

УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ



Марија Ристић

ОДРЕЂИВАЊЕ ТЕМАТСКИХ ОБЛАСТИ
НАУЧНИХ ЧАСОПИСА НА ОСНОВУ
ПУБЛИКОВАНИХ РАДОВА

мастер рад

Београд, 2026.

Ментор:

др Славко Гајин, ванредни професор
Универзитет у Београду, Математички факултет

Чланови комисије:

др Мирослав Марић, редовни професор
Универзитет у Београду, Математички факултет

Денис Аличић, асистент
Универзитет у Београду, Математички факултет

Датум одбране: _____

Захвалница

Захваљујем се својој породици, пре свега својој мајци и баки, које су ми биле подршка на сваком кораку на путу до дипломе!

Захваљујем се својим колегама и пријатељима који су били уз мене и онда када је било најтеже, али и онда када је било најлепше!

Захваљујем се и свом ментору, проф. др Славку, на стручном вођењу, стрпљењу и охрабрењу да кренем на пут у Италију, без којег овај мастер рад не би ни постојао!

Посебну захвалност дугујем и сваком члану свог тима на институту CNR, Алесији, Андреи, Клаудију, Ђамбатисти, Микелеу, Миријам, Паолу и Симонеу, који су веровали и били уз мене током читавог пројекта. Хвала им на свим саветима, сугестијама, подршци и дружењима, без којих Пиза никада не би била иста!

Наслов мастер рада: Одређивање тематских области научних часописа на основу публикованих радова

Резиме: Научно публикување готово у потпуности се преселило у електронски облик, па број часописа и количина метаподатака о њима непрестано расту. Податак о томе којој научној области часопис припада основа је за претрагу, вредновање и праћење научне продукције, али постојеће класификације обухватају углавном часописе индексиране у великим међународним базама, док они од националног и регионалног значаја остају без такве одреднице. Циљ овог рада јесте да се испита може ли се научна област часописа одредити аутоматски, искључиво на основу метаподатака публикација које је тај часопис објавио, дакле без ослањања на његов опис или на већ постојећу класификацију. Реализован је поступак у више етапа: наслови и сажеци публикација најпре се пречишћавају и обједињују у документе, потом се из сваког документа издвајају кључне фразе, које се групишу у тематске кластере, а сваки кластер се сажето описује носећим фразама и мапира на референтну таксономију научних области, и то на два начина, поређењем векторских репрезентација и упитом великом језичком моделу. Поступак је развијен и проверен над узорком од осамдесет часописа, насумично изабраних из свих шест области првог нивоа референтне таксономије. Примарна научна област часописа тачно је одређена у 79% случајева, односно у близу деведесет одсто ако се уместо прве узму прве три пронађене области, и то уз скромне рачунарске захтеве, малим језичким моделом од три милијарде параметара. Установљено је да главно ограничење није у самом поступку него у циљној таксономији, чији други ниво није довољно раздвојен и не обухвата све области које се у часописима заиста јављају.

Кључне речи: научни часописи, тематске области, таксономија научних области, метаподаци о публикацијама, издвајање кључних фраза, кластеровање, велики језички модели

Садржај

1	Увод	1
2	Преглед стања и постојећих решења	3
2.1	Таксономије научних области и проблем мапирања	3
2.2	Технологије	6
2.2.1	ETL поступак	7
2.2.2	Претпроцесирање	10
2.2.3	Машинско учење	13
2.2.4	Велики језички модели	16
3	Преглед предложеног решења	19
3.1	Радно окружење и коришћене библиотеке	21
3.2	Подаци	23
4	Формирање и опис тематских кластера	28
4.1	Приступ одозго-наниже	28
4.2	Приступ одоздо-навише	33
4.3	Одређивање носећих фраза и фраза подршке	40
5	Мапирање кластера на таксономију научних области	45
5.1	Приступ заснован на сличности векторских репрезентација	45
5.2	Приступ заснован на великом језичком моделу	48
6	Приказ и дискусија добијених резултата	52
6.1	Репрезентативан узорак часописа	52
6.2	Резултати приступа заснованог на сличности векторских репре- зентација	53
6.3	Резултати приступа заснованог на великом језичком моделу . . .	60

САДРЖАЈ

6.4	Анализа тематских промена часописа кроз време	67
7	Закључак	72
	Литература	75
	Листа слика	80
	Листа табела	81
	Списак скраћеница	82

Глава 1

Увод

Научно публикување се последњих година у потпуности преселило у електронски облик, чиме је нагло порасла и количина метаподатака о публикацијама којима треба управљати. Са порастом броја научних публикација расте и број научних часописа, како на светском тако и на националном нивоу. Широки спектар научних часописа даје појединцу слободу, али и потешкоћу да се одлучи где ће објавити следећи рад. Тачан податак о научној области часописа тај посао олакшава истраживачима, професорима и асистентима, зато што од области и категорије часописа зависи вредновање резултата, а тиме и испуњеност услова при избору у звање.

Исти податак потребан је и научним заједницама, попут радних група, универзитетских тимова и академских кругова, које прате промене у области научног публикувања и којима је припадност часописа научној области полазна тачка сваке такве анализе. За њих је значајно не само какво је стање него и како се оно мења, будући да померање тематског тежишта часописа говори и о померању саме научне области. Без поузданог податка о области такве анализе почивају на процени, па се њихови закључци тешко могу поновити. Информација о научној области часописа значајна је и базама које те часописе индексирају, како би омогућиле груписање и лакшу претрагу.

Са порастом броја часописа расту и различити извори у којима је могуће пронаћи податке о часопису. Како би се научни подаци уопште могли претраживати и анализирати на једном месту, развијају се графови научног знања (енгл. *scholarly knowledge graphs*), који публикације, истраживачке податке, ауторе, установе и пројекте повезују у једну мрежу. Најзначајнији међу њима јесте граф европске истраживачке инфраструктуре отвореног приступа *OpenAIRE*,

који настаје агрегирањем метаподатака из институционалних репозиторијума, издавача и националних инфраструктура [1], због чега обухвата и часописе релевантне само на националном или регионалном нивоу. Тај граф данас омогућава претрагу публикација, али не и претрагу самих часописа, а његово проширење подацима о часописима претпоставља уједначен начин одређивања њихове научне области.

Одређивање тематске области часописа отежано је с више страна. Постојеће таксономије научних области опште су природе, па њихове категорије могу одступати од стварне тематике часописа, а за велики број часописа класификација је непотпуна или уопште не постоји, будући да се сви часописи не индексирају у базама *Web of Science* (*WoS*) и *Scopus*. Уз то, описи часописа у метаподацима најчешће нису доступни, па остаје да је једини поуздан извор веб презентација самог часописа, што на нивоу неколико хиљада часописа, чије се адресе уз то мењају, није изводљив поступак.

Циљ овог рада јесте да се испита у којој мери се тематски профил научног часописа може одредити аутоматски, применом алгоритама машинског учења и обраде природног језика, уз поступке претпроцесирања текста, и то искључиво на основу метаподатака публикација објављених у том часопису. Такав поступак ослобађа људске ресурсе који се могу употребити за друге послове и олакшава прикупљање релевантних података у тренутку када истраживач одлучује где ће објавити рад. Уз аутоматизацију поступка смањује се могућност пристрасности и субјективности, а повећава се транспарентност добијене класификације.

Да би се област одредила прецизније, поступку је потребно дати шири контекст од самог назива часописа, а тај контекст чине управо публикације које је часопис објавио. Истраживачки рад, развој решења и иницијални резултати на којима овај рад почива настали су кроз сарадњу са Институтом за науку и технологије информација „А. Фаедо“ (*ISTI-CNR*) у Пизи, остварену у оквиру програма *Erasmus+ Traineeship*, чији су истраживачи аутори и одржаваоци кључних компоненти графа инфраструктуре *OpenAIRE*. Подаци над којима је рад изведен преузети су управо из тог графа, ради подршке при увођењу простора за претрагу часописа.

Глава 2

Преглед стања и постојећих решења

У овом поглављу најпре су приказане постојеће таксономије научних области и извори који часописима додељују област, заједно са ограничењима која из њих произлазе и са проблемом мапирања међу различитим поделама. Након тога дат је преглед досадашњих покушаја да се одређивање теме часописа аутоматизује, уз назнаку по чему се овај рад од њих разликује. У другом делу поглавља описане су технологије на којима почива предложено решење, од прикупљања и припреме података, преко претпроцесирања текста, до модела машинског учења и великих језичких модела.

2.1 Таксономије научних области и проблем мапирања

Фраскати таксономија је подела научних области која потиче из приручника *Frascati Manual*, који Организација за економску сарадњу и развој (*OECD*) издаје и допуњава од 1963. године [2]. Сврха приручника јесте да уједначи прикупљање статистичких података о истраживању и развоју, како би подаци из различитих земаља били међусобно упоредиви. Таксономија је хијерархијска и има два нивоа: први ниво чини шест широких области, а други четрдесет две уже научне области [2]. Области првог нивоа обележавају се једном цифром, од 1 до 6, а то су природне науке, техничко-технолошке науке, медицинске и здравствене науке, пољопривредне и ветеринарске науке, друштвене науке, као

и хуманистичке науке и уметност. Области другог нивоа обележавају се двома цифрама раздвојеним тачком, у распону од 1.1 до 6.5, при чему прва цифра означава област првог нивоа којој ужа област припада, тако да се надређена област чита непосредно из саме шифре. Разлози за избор ове таксономије јесу њена широка прихваћеност, језичка неутралност и то што је реч о затвореној листи од свега неколико десетина области. Области на другом нивоу нису довољно раздвојене и међу њима постоји знатно преклапање, што чини недостатак ове таксономије и отвара могућност за грешку између сродних области.

Класификацију *Science-Metrix* развила је истоимена компанија из Монтереала, специјализована за библиометријске анализе и вредновање науке, која је данас део групе *Elsevier* [3]. Класификација је хијерархијска и има три нивоа: пет домена, двадесет поља и сто седамдесет четири међусобно искључива потпоља [3]. Грађена је тако да поља буду што уједначенија по броју часописа, како ниједна област не би била толико широка да поређење учини безначајним. У овом раду је коришћено седмо јавно издање класификације, објављено 5. јануара 2022. године под ознаком верзије 2.00 и лиценцом *Creative Commons Attribution 3.0* [4]. То издање обухвата 34.655 различитих наслова часописа са тачно једном додељеном категоријом.

SCImago Journal & Country Rank (SJR) је јавно доступан портал који одржава истраживачка група *SCImago* из Шпаније, а који почива на подацима из библиографске базе *Scopus* издавача *Elsevier* [5]. Портал објављује показатеље утицаја часописа, а уз њих и предметну класификацију сваког часописа. Подела коју портал користи јесте *ASJC* (енгл. *All Science Journal Classification*), развијена уз базу *Scopus*, и она има хијерархијску структуру: часописи се разврставају у двадесет седам предметних области, које се даље деле на око три стотине тридесет ужих предметних категорија, а саме области груписане су у четири велике целине [6]. Уз сваку широку област постоји и категорија са ознаком *miscellaneous*, у коју се сврставају часописи општег усмерења унутар те области, што при поређењу треба имати у виду јер такве категорије носе мало информације о теми. Часопис по правилу припада већем броју категорија, уређених лексикографски, а не по значају, што чини једно ограничење ове класификације.

Web of Science (WoS) јесте библиографска и цитатна база компаније *Clarivate*, а уз њу се сваке године објављује и извештај *Journal Citation Reports (JCR)*, у ком се часописи разврставају по предметним категоријама [7]. Тих

категорија има двеста педесет четири, а часопис може припадати већем броју њих. Значајан аспект ове класификације јесте то што се категорија додељује по годишњем издању извештаја, па се тема часописа може пратити кроз време. Ограничење је то што нису сви часописи индексирани у тој бази.

ProRef (Провера референци) јесте информациони сервис који омогућава проверу категорије часописа у ком је објављена одређена библиографска референца, на основу обједињених података из више релевантних извора [8]. Сервис одржава Рачунарски центар Универзитета у Београду (*RCUB*). Намењен је универзитетским органима и телима, факултетима и институтима, као и свим установама којима је потребно вредновање резултата научноистраживачког рада.

Информациону основу сервиса *ProRef* чине три обједињена извора података: годишња издања *M*-категорија утврђених на основу извештаја *Journal Citation Reports*, категоризација домаћих часописа коју на предлог матичних научних одбора сваке године доноси министарство из ресора науке и годишња издања листе *SJR* [8]. Претрага је могућа по референци, по броју *DOI* (енгл. *digital object identifier*), по броју *ISSN* (енгл. *international standard serial number*) или по наслову часописа. Резултати се приказују по годинама, што омогућава да се прати како се категорија часописа мењала кроз време.

Граф научног знања *OpenAIRE* обједињује метаподатке о истраживачким производима као што су научне публикације, скупови података, истраживачки софтвер и осталим производима које повезује са организацијама, пројектима, финансијерима, изворима података и истраживачима [1]. Претрага се води над тим ентитетима, док се часописи у графу јављају само као извори података, описани административно и без тематске класификације. Простор за претрагу часописа, какав постоји за научне публикације, још увек не постоји. Како је замислио да се он у будућности изгради, потребан је уједначен начин за сврставање часописа у категорије. Различити извори из којих би се прикупљале информације о часописима индексирају различите часописе, при чему неки од њих уопште немају класификацију часописа, а неки имају класификацију која јасно не одређује примарну тему. Без обзира на то да ли подаци о часопису долазе из једног извора или из више њих, додељену класификацију потребно је потврдити, и то таксономијом коју инфраструктура *OpenAIRE* већ користи.

У сврху изучавања аутоматизације поступка за издвајање теме часописа из публикација, за потребе инфраструктуре *OpenAIRE*, у раду је коришћена управо таксономија којом та инфраструктура означава научне публикације у

свом графу [9]. Њу над публикацијама примењује систем *SciNoBo*, при чему су прва два нивоа таксономије преузета из Фраскати приручника, а трећи ниво чине категорије класификације *Science-Metrix* [9]. У раду су коришћена прва два нивоа, дакле онај део који потиче непосредно из приручника, али и он у односу на приручник има одступања.

Главна разлика је у томе што коришћена таксономија на другом нивоу има четрдесет области, а не четрдесет две [10]. Изостављене су две категорије које служе као скупљачи преосталих тема, „остале природне науке“ и „остале хуманистичке науке“, док су преостале четири категорије тог типа, у техничким, медицинским, пољопривредним и друштвеним наукама, задржане. Изостављање управо тих категорија за поступак је заправо повољно, јер класификатору не оставља лаку излазну категорију у коју би сврставао све што му није јасно. Друга разлика је у нумерацији области, које се пишу са водећом нулом, као 01 и 0101, што омогућава да се шифра првог нивоа добије узимањем прве две цифре. Додатак у односу на таксономију из приручника јесте то што уз сваку област стоји и опис у неколико реченица, као и списак кључних речи из те области.

2.2 Технологије

Развојем алата за прикупљање података сам процес прикупљања је значајно убрзан, али у њему и даље постоји људски фактор који га надгледа и контролише. Поред прикупљања, значајно време одузима и обрада великих количина података преузетих из различитих извора, будући да ти извори најчешће нису међусобно усклађени ни по формату ни по садржају. Како су људски ресурси ограничени, а количина расположивих података из године у годину све већа, у процесу дигитализације осмишљени су бројни алгоритми и модели који ову обраду убрзавају и делимично аутоматизују [11].

Данас значајну улогу у обради оваквих података имају модели машинског учења, а међу њима посебно велики језички модели (енгл. *large language model*, *LLM*). Иако могућности ових модела напредују из дана у дан, значајан део посла и даље се обавља пре њиховог покретања, у оквиру поступка *ETL* (енгл. *extract, transform, load*) и претпроцесирања података. Квалитет улазних података у великој мери одређује квалитет добијених резултата, па се овим корацима у пракси придаје пажња једнака избору самог модела [12]. Сви ови

кораци су детаљно изложени у наставку.

2.2.1 ETL поступак

ETL поступак представља устаљен начин организације обраде података, подељен у три фазе: екстракцију, трансформацију и учитавање [13]. Свака фаза има јасно одвојену улогу, што омогућава да се појединачни кораци развијају, испитују и понављају независно један од другог. У овом одељку описане су све три фазе, са тежиштем на трансформацији, у оквиру које се спроводе нормализација и уклањање дупликата.

Екстракција подразумева прикупљање сирових података из различитих извора. Извори могу бити базе података, датотеке различитих формата или веб сервиси, при чему се прикупљени подаци међусобно разликују и по формату записа и по шеми, односно по томе која поља уопште постоје и шта она значе. Управо та разноликост чини наредну фазу неопходном.

Трансформација подразумева уједначавање тако прикупљених података. Приликом трансформације примењују се различите технике нормализације, којима се сирови подаци преводе у уједначен облик погодан за даљу обраду. Циљ нормализације јесте да се исти садржај, записан на више различитих начина, сведе на јединствен запис, чиме се смањује број различитих облика истог податка и олакшава њихово касније поређење и груписање.

Нормализација текстуалних података може подразумевати свођење текста на мала слова, при чему се властите именице и скраћенице третирају подједнако као и остатак текста. Друга уобичајена техника јесте нормализација записа према стандарду Уникод (енгл. *Unicode*), којом се визуелно исти знаци своде на исти запис [14]. На тај начин се, на пример, средње црте које долазе из различитих извора или са различитих тастатура тумаче као исти знак.

Уклањање вишеструких размака и интерпункцијских знакова текст на први поглед чини мање читљивим, али смањује број различитих записа истог појма. Тиме се смањује и проређеност приказа над којим модел ради, односно уклања се шум који не носи информацију значајну за задатак. Из истог разлога се, у зависности од задатка, уклањају и знаци који не припадају писму изабраног језика.

Поред функција за нормализацију постојећег текста, у трансформацију података спада и уклањање дупликата, односно дедупликација. Дедупликација представља део чишћења података и подразумева проналажење група истих

или сличних записа који се могу заменити само једном инстанцом из групе [15]. Овим поступком се скуп података за обраду може значајно смањити, без битног губитка информација које тај скуп носи.

Проналажење истоветних записа своди се на непосредно поређење њихових вредности, док је проналажење сличних записа знатно сложеније. Приликом рада са текстуалним подацима за ту сврху се често користи Левенштајново растојање. Оно омогућава да се два записа прогласе сличним и онда када се не поклапају у потпуности, што је уобичајен случај код података прикупљених из више извора.

Левенштајново растојање јесте најмањи број уметања, брисања и замена знакова потребних да се једна ниска преведе у другу [16]. Уведено је у теорији кодирања, ради описивања кодова отпорних на грешке при преносу, а данас се широко примењује при упаривању и дедупликацији текстуалних записа. Свака од три дозвољене операције има једнаку цену, што ово растојање разликује од сродних мера које различитим операцијама додељују различите тежине.

Нека су дате ниске $a = a_1a_2 \dots a_m$ и $b = b_1b_2 \dots b_n$, а нека $lev(i, j)$ означава растојање између њихових префикса дужина i и j . Уколико је једна од тих дужина једнака нули, вредност је $lev(i, j) = \max(i, j)$, јер је тада преосталу ниску потребно у целости уметнути, односно обрисати. За $i > 0$ и $j > 0$ важи рекурентна релација

$$lev(i, j) = \min\{lev(i-1, j) + 1, lev(i, j-1) + 1, lev(i-1, j-1) + c(i, j)\} \quad (2.1)$$

где функција цене $c(i, j)$ узима вредност 0 уколико је $a_i = b_j$, а вредност 1 у супротном, док три члана унутар минимума одговарају редом брисању, уметању и замени једног знака. Тражено растојање између целих ниски јесте $d(a, b) = lev(m, n)$.

Левенштајново растојање је метрика, што значи да је ненегативно, да је једнако нули ако и само ако су ниске истоветне, да је симетрично и да задовољава неједнакост троугла. Захваљујући томе може се користити као мера сличности и на основу њега поставити праг за одлучивање о томе да ли два записа представљају исти податак. Растојање се израчунава динамичким програмирањем, попуњавањем матрице у којој поље (i, j) садржи вредност $lev(i, j)$, а сам поступак приказан је на слици 2.1 [17].

Када су подаци у потпуности обрађени, потребно их је сачувати у формату погодном за даље процесирање, чиме се скупи кораци обраде не морају понављати приликом сваке наредне анализе. Управо то представља трећу фазу,

```

УЛАЗ: ниске а и b, дужина m и n
ИЗЛАЗ: Левенштајново растојање d(a, b)

за i = 0 до m:
    D[i, 0] <- i
за j = 0 до n:
    D[0, j] <- j

за i = 1 до m:
    за j = 1 до n:
        ако је a[i] = b[j]:
            c <- 0
        иначе:
            c <- 1
        D[i, j] <- min( D[i-1, j]   + 1,      // брисање
                       D[i, j-1]   + 1,      // уметање
                       D[i-1, j-1] + c )      // замена

врати D[m, n]

```

Слика 2.1: Псеудокод за израчунавање Левенштајновог растојања

учитавање. Избор формата није формалност, јер од њега зависи и то које се информације о подацима чувају, и то у којим се све окружењима подаци касније могу читати.

Формат *CSV* (енгл. *comma-separated values*) и њему сродан формат *TSV* (енгл. *tab-separated values*) су преносиви и читљиви, али не чувају типове података и незгодни су за вредности које саме представљају низове. Формат *pickle* чува објекте програмског језика *Python* у изворном облику, али зависи од самог језика и од верзија коришћених библиотека. Колонски формати, попут формата *Parquet* [18], погодни су за велике скупове података јер омогућавају читање само оних колона које су заиста потребне.

Уколико се *ETL* поступак спроводи над количином података коју није могуће у целости учитати у радну меморију једног рачунара, за обраду се могу користити системи за дистрибуирану обраду. Један од најраспрострањенијих таквих система јесте алат *Apache Spark*. Употреба таквог система мења и начин на који се поједини кораци обраде записују, јер се они извршавају над деловима података распоређеним на више рачунара.

Apache Spark је систем отвореног кода намењен дистрибуираној обради великих количина података [19]. Настао је као одговор на ограничења ранијих система за пакетну обраду, пре свега на потребу да се међурезултати сваког ко-

рака уписују на диск. Основна замисао јесте да се подаци, кад год је то могуће, задрже у радној меморији рачунара у кластеру, чиме се вишеструко убрзавају поступци који над истим подацима извршавају више узастопних операција.

Скуп података се дели на партиције, делове који се обрађују независно један од другог и распоређују међу извршиоцима (енгл. *executor*), процесима који се извршавају на чворовима кластера. Оваква обрада назива се пакетном обрадом (енгл. *batch processing*), јер се унапред ограничен скуп података обрађује у целости, за разлику од обраде токова података, код које подаци пристижу непрекидно. Пошто је количина података која се обрађује у једном тренутку одређена величином партиције, а не укупном величином скупа, могуће је обрадити и скупове веће од радне меморије било ког појединачног рачунара, при чему се међурезултати који премаше расположиву меморију измештају на диск. Уз то, операције над подацима не покрећу обраду одмах, него се памте и извршавају тек позивом акције, попут пребројавања редова или уписа резултата у датотеку, што систему омогућава да пре покретања сагледа цео низ операција и да га оптимизује. У овом раду је алат *Apache Spark* коришћен за обраду прикупљених метаподатака, о чему је више речи у глави 3.

2.2.2 Претпроцесирање

Када су подаци успешно ускладиштени, потребно их је прилагодити конкретном моделу машинског учења који се користи. Сваки модел има своје захтеве у погледу тога какве податке очекује на улазу, због чега овај поступак зависи од избора модела. Из истог разлога претпроцесирање се не може у потпуности одвојити од самог моделовања, него се њихове одлуке међусобно условљавају.

За моделе који претежно раде над великом количином текстуалних података значајно је уклонити стоп-речи. Стоп-речи су врло честе речи које не носе тематску информацију, попут везника, предлога и заменица [20]. Уклањањем ових речи смањује се димензионалност приказа и уклања шум, а модел се у већој мери усредсређује на речи које заиста одређују садржај текста.

Велика количина текстуалних података подразумева и то да се једна иста реч јавља у различитим облицима. Најчешћи пример је појава исте речи у једнини и у множини. Иако такви облици семантички носе исту информацију, синтаксно се разликују, па их модел без додатне обраде третира као две различите речи.

Код енглеског језика примена једноставног правила да се са краја сваке речи уклони наставак не даје увек тачне резултате. Код речи какве су **news** или **analysis** примена оваквог правила може дати облик другачијег значења или облик који уопште не постоји у језику. Због тога се углавном користе поступци стемовања и лематизације. Стемовање је грубо свођење речи на корен, при чему добијени облик не мора бити постојећа реч [21], док лематизација представља свођење речи на речнички облик уз коришћење језичког модела и података о врсти речи.

Поједини модели машинског учења који раде над текстуалним подацима не примају текст непосредно, него његову векторску репрезентацију (енгл. *embedding*). Најједноставнији начин да се текст представи бројевима јесте представа заснована на учестаности речи (енгл. *bag of words*), где свака димензија одговара по једној речи речника, а вредност је учестаност те речи, евентуално пондерисана мером *TF-IDF* (енгл. *term frequency – inverse document frequency*) [22]. Такав приказ је редак и веома високе димензије, а његов основни недостатак је то што су сваке две различите речи међусобно подједнако удаљене, па синоними и сродни појмови немају ништа заједничко [20].

Векторске репрезентације решавају управо тај проблем, а почивају на дистрибутивној хипотези, према којој речи које се јављају у сличним контекстима имају и слична значења [23]. Први широко коришћени модели ове врсте, попут модела *word2vec*, уче густе векторе тако што предвиђају реч на основу њеног окружења, при чему свака реч добија тачно један вектор, независно од контекста у ком се појављује [24]. Долазак модела заснованих на архитектури трансформера, међу којима је најпознатији *BERT* (енгл. *Bidirectional Encoder Representations from Transformers*), донео је контекстуалне репрезентације, у којима иста реч добија различит вектор у зависности од реченице у којој се налази [25].

За задатке у којима се пореде целовите текстуалне јединице, а не појединачне речи, потребан је један вектор по таквој јединици, која се у даљем тексту назива документом. Просто усредњавање вектора појединачних речи даје слабе резултате, па су развијени модели који се посебно обучавају тако да косинусна сличност између два вектора одговара семантичкој сличности текстова које представљају, што је приступ уведен моделом *Sentence-BERT* [26]. На тај начин се поређење два текста своди на једноставну рачунску операцију над

њиховим векторима.

Пошто општи модели нису обучавани над научним текстом, за потребе обраде научних радова развијени су посебни модели. Модел *SciBERT* је претрениран над корпусом научних публикација, чиме боље препознаје стручну терминологију [27]. Модел *SPECTER* уводи битно другачију замисао: он користи мрежу цитата као извор надзора, полазећи од претпоставке да су радови који цитирају један другог тематски блиски, и уз њега је уведен и скуп *SciDocs*, који служи као мерило за поређење модела ове врсте [28].

Модел *SciNCL* надограђује овај приступ тако што узорке за обучавање бира из суседства у графу цитата, чиме прецизније одређује који су радови блиски, а који не, и на мерилу *SciDocs* постиже боље резултате од модела *SPECTER* [29]. Из тог разлога је у овом раду за рачунање векторских репрезентација изабран управо модел *SciNCL*. Његова је додатна предност то што су наслови и апстракти научних радова управо она врста улаза за коју је модел и обучаван.

Векторска репрезентација текста даје векторе великих димензија, најчешће реда неколико стотина. Иако модели машинског учења могу да раде и са овим великим бројем димензија, резултати у таквим просторима често постају непоуздани, што је познато као проклетство димензионалности. Проблем настаје управо при рачунању растојања између две тачке: са порастом броја димензија растојања се концентришу, односно однос између растојања до најдаљег и до најближег суседа тежи јединици, па мере засноване на растојању губе моћ разликовања [30].

Једна од техника за ублажавање овог проблема јесте метода *UMAP* (енгл. *Uniform Manifold Approximation and Projection*). Она полази од претпоставке да подаци, иако записани у простору велике димензије, заправо леже на много-структури знатно мање димензије. Метода најпре у изворном простору гради граф најближих суседа, који описује локалну структуру података, а затим тражи распоред тачака у простору мале димензије чији је граф што сличнији полазном [31]. Тиме се чува локална структура, а делом и глобални распоред група, па алгоритми који следе раде у простору у ком су растојања поново информативна.

Понашање методе *UMAP* одређује неколико параметара. Број суседа управља односом између локалне и глобалне структуре, најмање растојање одређује колико густо се тачке пакују у добијеном приказу, док се бројем излазних димензија и избором метрике задају циљни простор и начин мерења сличности.

Пошто је поступак стохастички, за поновљивост резултата неопходно је задати и семе генератора случајних бројева.

2.2.3 Машинско учење

Машинско учење је област вештачке интелигенције која се бави алгоритмима који решење задатка не добијају унапред записано у виду правила, него га изводе из података. Уместо да програмер опише поступак, он бира облик модела и меру успешности, а сам модел се обучава тако што се његови параметри подешавају на основу примера. Успешност се не процењује на подацима над којима је модел обучаван, него на подацима које током обучавања није видео, чиме се проверава да ли је модел заиста научио правилност или је само запамтио примере.

Алгоритми машинског учења се могу категоризовати у зависности од врсте обучавања и података које користе. Према оваквој подели разликују се четири основна типа машинског учења [32]. Избор типа не зависи од алгоритма него од тога какви су подаци на располагању, пре свега од тога да ли за улазне податке постоје познате излазне вредности.

- Надгледано учење — сваком улазном податку x придружена је позната излазна вредност y . Циљ је одредити хипотезу $h(x)$ која ће на основу нових улазних података предвиђати излазну вредност.
- Ненадгледано учење — познати су улазни подаци x , а излазне вредности y нису познате. Потребно је пронаћи правилност међу улазним подацима. Овакви алгоритми најчешће решавају проблеме кластеровања.
- Полунадгледано учење — комбинација надгледаног и ненадгледаног учења. За мањи део улазних података излазне вредности су познате, док су за већи део непознате.
- Учење са подршком — учење се врши на основу улазних података и сигнала подршке [33]. Сигнал подршке се шаље на крају обављања скупа акција и може бити позитиван или негативан, а његова вредност одређује да ли је скуп акција обављен исправно или погрешно. Своју примену најчешће проналази код сложених проблема.

Једна од важних области ненадгледаног учења је кластеровање. Кластеровање представља тражење група сличних података унутар целог скупа, при чему се групе не задају унапред него их алгоритам сам открива. Пошто тачан одговор није познат, квалитет добијене поделе се процењује посредно, преко мера унутрашње сличности група и њихове међусобне раздвојености.

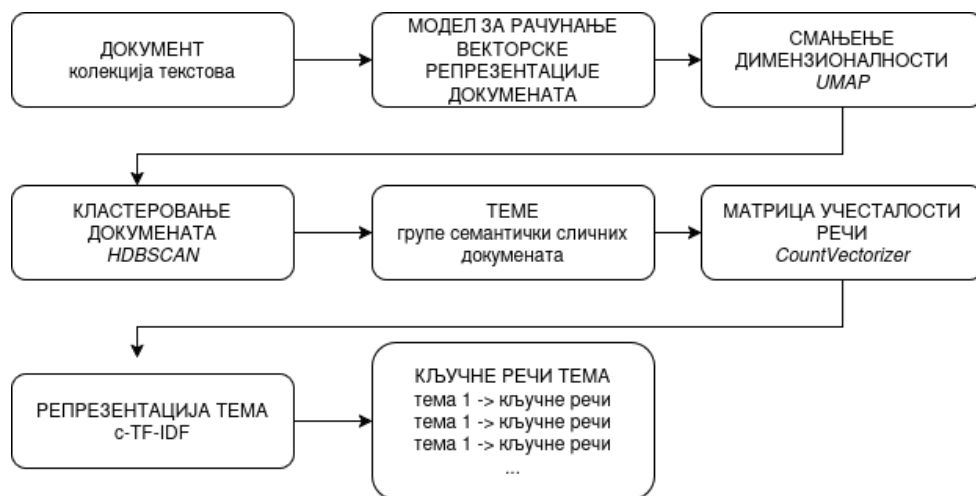
Један од алгоритама који се користи за кластеровање је *HDBSCAN* (енгл. *Hierarchical Density-Based Spatial Clustering of Applications with Noise*). Он припада породици алгоритама заснованих на густини, код којих је кластер област простора у којој су тачке густо распоређене, а раздвојен је областима мале густине [34]. За разлику од сродног алгоритма *DBSCAN* [35], који захтева да се унапред зада полупречник околине, *HDBSCAN* тај поступак понавља за читав опсег вредности и на основу њега гради хијерархију кластера.

Из добијене хијерархије се потом издваја једна равна подела, и то тако што се задржавају они кластери који опстају у најширем распону вредности параметра, односно они који су најстабилнији. Захваљујући томе, алгоритам не тражи унапред задат број кластера и подноси кластере различите густине и произвољног облика. Тачке које не припадају ниједном кластеру не додељују се насилно најближој групи него се означавају као шум, што је у обради текста корисно јер издвојене фразе често не припадају ниједној теми.

У категорију ненадгледаног учења спадају и модели који се баве тематским моделовањем. Замисао тематског моделовања јесте да постоји велики број докумената без унапред додељених ознака, а алгоритам покушава да открије које се теме природно појављују у збирци [36]. У овој групи модела најчешће се помињу *BERTopic* и *KeyBERT*.

Модел *BERTopic* није један алгоритам него оквир састављен од уланчаних корака [37]. Најпре се за сваки документ израчуна векторска репрезентација, потом се методом *UMAP* смањује број димензија, а затим се добијени вектори кластерују алгоритмом *HDBSCAN*. Сваки добијени кластер представља једну тему, а документи који су у њега сврстани сматрају се документима те теме. Дијаграм тока обраде података моделом *BERTopic* приказана је на слици 2.2.

Опис теме се добија у последњем кораку, применом статистичке мере *c-TF-IDF* (енгл. *class-based term frequency – inverse document frequency*) [37]. Сви документи једног кластера се посматрају као један велики документ, па се учестаности речи рачунају на нивоу кластера, а не појединачног документа, чиме се издвајају речи карактеристичне баш за ту тему. Овако добијен опис

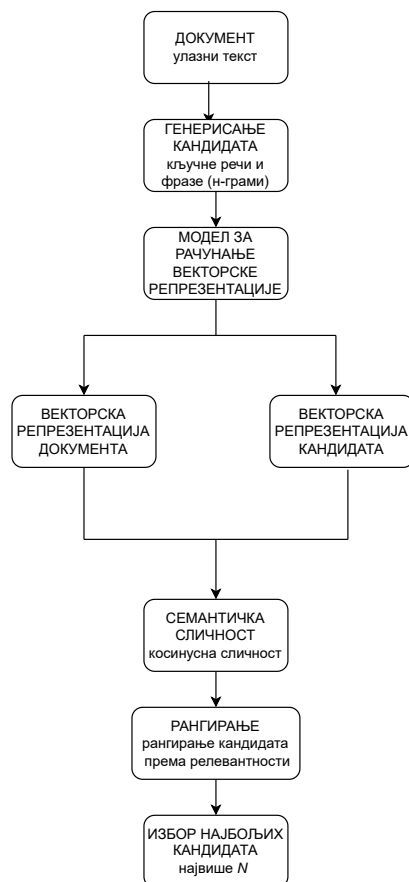


Слика 2.2: Дијаграм тока обраде података моделом *BERTopic*

се може додатно побољшати поступцима који смањују понављање сродних израза. Треба имати у виду да овакав приступ сваком документу додељује једну доминантну тему, што је ограничење када један документ припада већем броју тематских области.

Алат *KeyBERT* решава сродан задатак, али са супротне стране јер не тражи теме збирке него фразе које описују један појединачни документ [38]. Поступак је једноставан: најпре се израчуна векторска репрезентација целог документа, затим се из тог истог документа издвоје све фразе кандидати, обично сви n -грами задате дужине, и за сваку од њих се израчуна вектор. Фразе се потом рангирају по косинусној сличности са вектором документа, а као резултат се враћа задати број најсличнијих. Дијаграм тока обраде података алатом *KeyBERT* је на слици 2.3.

Оваквим поступком добијају се фразе које су заиста присутне у тексту, што описе чини разумљивим и лако проверљивим. Пошто фразе са највећом сличношћу често међусобно личе, на располагању су и поступци који бирају разноврсније фразе, међу којима је најпознатија максимална маргинална релевантност [39]. За разлику од модела *BERTopic*, алат *KeyBERT* не захтева обучавање нити читаву збирку докумената, него ради над сваким документом посебно, због чега се два приступа често називају приступима одозго-наниже и одоздо-навише.



Слика 2.3: Дијаграм тока обраде података алатом *KeyBERT*

2.2.4 Велики језички модели

Велики језички модели (енгл. *large language model*, *LLM*) јесу модели који садрже велики број параметара и који су обучавани над великом количином текста. Њихова основна намена јесте обрада природног језика: они представљају математички систем који је током обучавања научио обрасце у подацима и на основу њих генерише одговор који статистички и контекстуално одговара постављеном упиту [40].

Ови модели почивају на архитектури трансформера [41], а обучавају се у две основне фазе. У фази претренирања модел над огромним корпусом текста учи да предвиђа наредни токен, чиме посредно усваја и правила језика и знање садржано у том тексту. У другој фази модел се додатно обучава над примерима упутстава и жељених одговора, а потом и уз повратну информацију људи који рангирају понуђене одговоре, чиме се постиже да прати задато упутство уместо да само наставља започети текст [42]. Управо та друга фаза раздваја

моделе који настављају текст од оних који су употребљиви као саговорници и извршиоци задатака.

Велики језички модели се могу класификовати према више критеријума. Ниједна од подела није строга, јер се исти модел најчешће може сврстати по свакој од њих. У наставку су наведени критеријуми који су најзначајнији при избору модела за конкретан задатак.

- Према организацији која их развија, међу којима су *OpenAI*, *Anthropic*, *Google*, *Meta* и *Alibaba*. Модели различитих организација разликују се по архитектури, начину обучавања и начину коришћења.
- Према нивоу отворености. Затвореним моделима се најчешће приступа преко програмског интерфејса (енгл. *application programming interface*, *API*), док се код отворених модела разликује шта је све доступно: само научене тежине, или уз њих и архитектура, програмски код и подаци за обучавање. Значајни су и услови лиценце под којом је модел објављен, јер они одређују сме ли се модел користити у истраживању и у комерцијалне сврхе.
- Према намени. Постоје модели оптимизовани за писање програмског кода, модели усмерени на решавање сложених проблема кроз кораке закључивања, модели посебно обучавани да прате корисничка упутства, модели ограничени на одређени домен и општи модели намењени разговору, анализи и одговарању на питања. Ова подела није строга и један модел може имати више намена.
- Према величини, која се најчешће изражава бројем параметара. Већи број параметара не значи нужно и бољи модел.
- Према дужини контекста, односно броју токена које модел може да обради док генерише одговор. Токен је овде део текста добијен поделом улаза и не мора одговарати једној речи.

Локално покретање омогућено је чувањем тежина у формату половичне прецизности (`float16`), чиме се захтеви за меморијом преполовљују у односу на пуну прецизност, уз занемарљив губитак квалитета [43]. Тежине нису додатно квантизоване. Поред тога што не захтева слање података спољној услузи, локално покретање даје и потпуну контролу над верзијом модела, што је значајно

за поновљивост резултата. Цена квантизације јесте извесно одступање од одговора које би дао модел у пуној прецизности, па и то треба имати у виду при тумачењу добијених резултата.

При коришћењу великих језичких модела треба имати у виду и њихова ограничења. Подразумевано узорковање наредног токена чини одговоре недетерминистичким, па исти упит при узастопним покретањима може дати различите резултате. Модел може да произведе тврдњу која звучи уверљиво иако није тачна, што се назива халуцинацијом [44]. Излаз је осетљив и на саму формулацију упита, а када се тражи одговор у унапред задатом облику, ништа не гарантује да ће тај облик заиста бити испоштован, па је накнадна провера неопходна.

Глава 3

Преглед предложеног решења

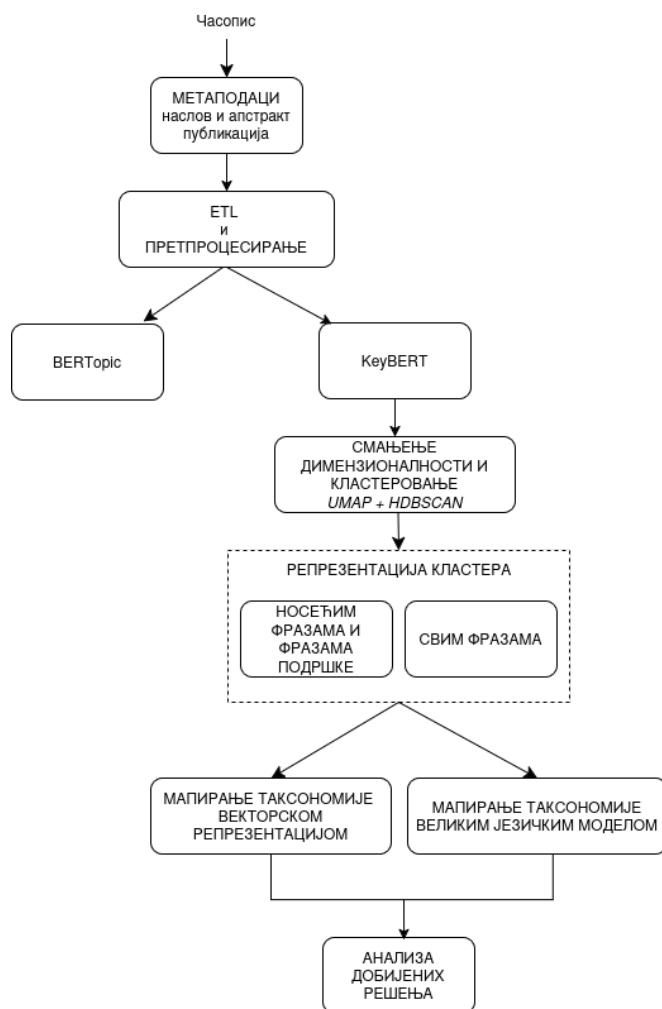
Подаци за овај рад прикупљени су кроз сарадњу са Институтом за науку и технологије информација „А. Фаедо“ (*ISTI-CNR*) у Пизи, остварену у оквиру програма *Erasmus+ Traineeship*. Истраживачи овог института аутори су и одржаваоци кључних компоненти графа *OpenAIRE*. Тај граф представља богату мрежу истраживачких података, која повезује публикације, скупове истраживачких података, ауторе, установе и пројекте, а настаје агрегирањем метаподатака из великог броја репозиторијума, издавача и националних инфраструктура [1].

Ради даљег обогаћивања графа потребан је и простор за чување података о научним часописима. Часопис се, поред наслова, бројева *ISSN* и *EISSN* (енгл. *electronic international standard serial number*), језика, лиценце, начина отвореног приступа и трошкова обраде рада (енгл. *article processing charge, APC*), описује и тиме о чему говори. Податак о теми часописа значајан је не само због лакше претраге него и самим истраживачима, у тренутку када треба да одлуче где ће објавити свој рад.

Најпоузданији начин да се тема часописа утврди јесте непосредан увид у његову веб презентацију, али такав поступак није изводљив на нивоу неколико хиљада часописа, јер за сваки од њих треба чувати веб адресу, која се уз то може мењати. Због тога је потребан аутоматски поступак који ће тему часописа извести из података који су већ доступни. Пошто су у графу доступни метаподаци публикација објављених у часописима, полазна претпоставка овог рада јесте да се тема часописа може одредити управо на основу њих, па је истраживачко питање формулисано на следећи начин: да ли је могуће аутоматски препознати тему часописа на основу метаподатака публикација објављених у

њему?

Ради лакшег сагледавања целокупног рада, у наставку овог поглавља најпре је укратко описан концепт примењеног решења, а затим се износе детаљи радног окружења у ком је решење реализовано, а потом се описују подаци над којима је рађено, заједно са поступком њихове припреме. Сам поступак издвајања тема детаљно је изложен у главама 4 и 5. Приказ међурезултата свуда је дат на једном истом часопису, како би се поступци могли непосредно упоредити. На слици 3.1 дат је шематски приказ целог система који олакшава праћење тока.



Слика 3.1: Дијаграм тока имплементiranог решења

Укратко, решење полази од метаподатака публикација једног часописа. Назлови и сажети се најпре чисте и обједињују у документе, а потом се моделима обученим над научним текстом преводе у векторске репрезентације. Груписа-

њем таквих репрезентација издвајају се тематски кластери и за сваки од њих одређује опис, при чему су упоређена два приступа — приступ одозго-наниже, заснован на моделу *BERTopic*, и приступ одоздо-навише, заснован на алату *KeyBERT*, из којег се даље издвајају носеће фразе и фразе подршке (глава 4). Тако добијени кластери се затим пресликавају на други ниво таксономије коју користи *OpenAIRE*, и то на два начина — рачунањем косинусне сличности између векторских репрезентација кластера и области, и упитом великом језичком моделу (глава 5). Резултат целог поступка јесу примарна и секундарне теме часописа, чија се тачност потом проверава поређењем са класификацијама *Science-Metrix* и *Scimago* (глава 6).

3.1 Радно окружење и коришћене библиотеке

Рад је изведен у два одвојена рачунарска окружења, а подела између њих није ствар избора него последица једне техничке чињенице: лични рачунар нема графичку картицу са подршком за платформу *CUDA* (енгл. *compute unified device architecture*). Пошто су сви кораци у којима се текст провлачи кроз трансформерску мрежу без такве картице неприхватљиво спори, они су пренети у клауд, док је све остало извршавано локално.

Локално окружење је виртуелно окружење језика *Python* 3.12.3 на рачунару са оперативним системом *Ubuntu* 24.04.3 *LTS* (енгл. *long-term support*), процесором *Intel Core i5-1135G7* са четири језгра и осам нити и 14 GiB радне меморије. Код је писан и извршаван у окружењу *JupyterLab* 4.5.7. У овом окружењу изведене су три фазе. Прва је прикупљање и претпроцесирање података, за коју је коришћен систем *Apache Spark* 4.1.1 за дистрибутивну обраду података које ради на виртуелној машини *Java* због чега у окружењу мора постојати и извршно окружење *OpenJDK* 21.0.12. Друга фаза је смањење броја димензија поступком *UMAP* 0.5.12 и кластеровање поступком *HDBSCAN* 0.8.43, уз помоћне библиотеке *scikit-learn* [45], *spaCy* 3.8.7 [46], *NLTK* [47] и *inflect* за нормализацију фраза и *Matplotlib* [48] за приказе. Трећа је мапирање кластера на таксономију *Frascati*, које се своди на једно множење матрица над већ израчунатим векторима, уз упис резултата у табеларну датотеку помоћу библиотеке *openpyxl* 3.1.5. У истом окружењу изведен је и експеримент са методом *BERTopic* 0.17.4.

Друго окружење је услуга под називом *JupyterHub at Google Cloud*, у оквиру виртуелне лабораторије *Blue-CloudLab* инфраструктуре *D4Science* [49]. При-

ступ је омогућен кроз сарадњу са Институтом за информационе науке и технологије (*ISTI*) у Пизи, који ту инфраструктуру и развија у оквиру организације *ISTI-CNR*. *JupyterHub* 5.3.0 при пријави покреће засебан контејнер са унапред припремљеном сликом, у ком је *Python* 3.10.12 и *JupyterLab* 4.4.4. Контејнер је привремен и брише се при заустављању сервера, док подаци остају у одвојеном простору за трајно складиштење. Расположиве су две графичке картице *NVIDIA Tesla V100-SXM2-16GB* од којих свака има 5.120 рачунских и 640 тензорских језгара и 16 GB меморије типа *HBM2* (енгл. *high bandwidth memory*), чија пропусност од око 900 GB/s и јесте оно што генерисање текста чини брзим, будући да се при сваком кораку из меморије мора прочитати цео модел. Програмски слој чине библиотека за рад са тензорима и неуронским мрежама *PyTorch* 2.3.0+cu121 [50], над платформом *CUDA* 12.1 за рачунање опште намене на графичкој картици, уз библиотеке *cuBLAS* 12.1.3.1 за операције линеарне алгебре и *cuDNN* 8.9.2.26 за слојеве неуронских мрежа, и библиотека *Accelerate* 1.14.0 за распоређивање модела по картицама.

У том окружењу изведена су два посла. Први је издвајање кључних фраза методом *KeyBERT* 0.8.5 над моделом *malteos/SciNCL*, учитаним преко библиотеке *sentence-transformers* 3.2.1. Истом свеском израчунате су и векторске репрезентације свих четрдесет области нивоа 2 таксономије, чиме су публикације и таксономија доведене у исти векторски простор. Други посао је издвајање тема часописа, помоћу великог језичког модела *Qwen2.5-3B-Instruct* [51], покретаног преко библиотеке *Transformers* 4.51.3 [52].

Убрзање које графичка картица доноси почива на три ствари. Прва је природа посла: трансформерска мрежа се своди на велика множења матрица, која процесор изводи на осам нити, а картица дели на хиљаде језгара која раде истовремено, уз меморију пропусности реда неколико стотина гигабајта у секунди. Друга је полупрецизни запис јер је модел уčitан са `torch_dtype=torch.float16`, чиме се заузеће меморије преполовљује и множења прелазе на тензорска језгра, наменски грађена за такав запис. Трећа је групна обрада: текстови се кроз мрежу провлаче у групама од 32 односно 64 одједном, као једна матрица, уместо један по један. Уз то, задавањем `device_map=„auto“` библиотека *Accelerate* [53] распоређује слојеве модела на обе картице; испис из свески показује заузеће од око 2,9 GiB по картици при издвајању тема и око 5,8 GiB при мапирању на таксономију.

Веза између два окружења остварена је искључиво преко датотека тако што

су резултати скуних корака, издвајање кључних фраза публикација, векторска репрезентација таксономије и сирови одговори језичког модела трајно сачувани. Захваљујући томе, свака наредна фаза те резултате само учитава, анализа се може поновити без поновног покретања модела, а неуспели упити поновити појединачно. Поновљивост је додатно обезбеђена задатим семеном 42 за стохастички поступак *UMAP* и усклађивањем верзија библиотека које се користе локално и у клауд окружењу. Једина преостала разлика између окружења, верзије програмског језика 3.12.3 наспрам 3.10.12, безопасна је управо зато што окружења размењују само податке.

3.2 Подаци

Подаци о научним публикацијама преузети из графа *OpenAIRE* испоручени су у четири компримоване датотеке у формату *JSON* (енгл. *JavaScript object notation*), укупне величине око 20 GB. У њима се налази 49.063.172 публикације и 156.099 различитих комбинација идентификатора *ISSN* и *EISSN*, односно толико различитих часописа. Обим података је знатно мањи од онога што граф *OpenAIRE* садржи, будући да су о публикацијама прикупљени само метаподаци релевантни за ово истраживање, и то у распону од 2020. до 2026. године. Ограничење на последњих седам година постављено је како би утврђена тема часописа одговарала садашњем стању публикација које он објављује.

Шема сваке од датотека приказана је на слици 3.2. Истоветна је за све четири датотеке, будући да су настале истим поступком извоза. Поља која она садржи одређују и оно што је у наставку уопште било могуће искористити.

```
root
|-- descriptions: array (nullable = true)
|   |-- element: string (containsNull = true)
|-- eissn: string (nullable = true)
|-- issn: string (nullable = true)
|-- journal_name: string (nullable = true)
|-- title: string (nullable = true)
|-- year: long (nullable = true)
```

Слика 3.2: Шема улазних датотека са метаподацима публикација

Шема приказује који су све метаподаци о научним публикацијама прикупљени: описи публикације, бројеви *ISSN* и *EISSN*, назив часописа, наслов пу-

бликације и година објављивања. Поље `descriptions` садржи сажетак, односно апстракт публикације, и представљено је низом, будући да метаподаци о једној истој публикацији могу бити прикупљени из различитих извора; током прикупљања обезбеђено је да у датотекама нема публикација чији је тај низ празан, а број његових елемената креће се од један до десет. Број *ISSN* представља идентификациони број часописа у штампаном издању, док број *EISSN* означава идентификациони број часописа у електронском издању. Поље `year` садржи годину објављивања публикације са насловом `title` у часопису чије је име у пољу `journal_name`. Пошто публикације пристижу из различитих извора, исти часопис се у подацима може јавити под различитим називима.

Свака публикација са својим метаподацима чини једну инстанцу података. Због велике количине података, коју није могуће одједном учитати у радну меморију, за брзу и оптимизовану претрагу, за анализу и за сам *ETL* поступак коришћен је алат *Apache Spark* [19]. Обрада је организована тако да се часописи обрађују један по један, чиме се у сваком тренутку у меморији држи само део укупног скупа.

Детаљан *ETL* поступак спроводи се над објектом типа `DataFrame` алата *Apache Spark*, уз задате бројеве *ISSN* и *EISSN* часописа над којим се изводи и задато одредиште за смештање резултата. Поступак обрађује један по један часопис, чиме се може покренути и поновити за сваки часопис независно.

Током фазе трансформације података коришћени су регуларни изрази, којима су из наслова и из сваког елемента низа описа уклоњене ознаке језика *HTML* (енгл. *hypertext markup language*) и *XML* (енгл. *extensible markup language*), ентитети тих језика, односно секвенце које почињу знаком `&`, вишеструки размаци, као и размаци са почетка и краја текста. Уочено је и да се у многим насловима, унутар заграде, налази назив или редни број конференције или серије, што за одређивање теме представља чист шум. Такве заграде уклоњене су изразом

$$(? i) \backslash ([^]) * (\text{conf} . ? \backslash \text{b} | \text{ser} . ? \backslash \text{b}) [^] * \backslash)$$

Овај израз обухвата заграду у којој се, било где унутар ње, јавља низ знакова `conf` или `ser` непосредно праћен границом речи, чиме се препознају скраћенице попут `conf.` и `ser.`, а не и пуни облици `conference` и `series`. Пуни облици се обрађују посебно: наслови у којима се јавља реч `conference` уклањају се у целисти, заједно са насловима који садрже изразе `peer review`, `preface`, `retraction`

или *corrigendum*. За такве публикације уочено је да углавном постоје и одговарајуће верзије без ових речи, па је ради смањења димензионалности одлучено да се оне даље не разматрају. Уклоњене су и инстанце са неинформативним насловом сачињеним од само једне речи која нема више од десет знакова, а сви преостали наслови су уједначени свођењем на мала слова.

Ради даљег смањења димензионалности спроведен је поступак уклањања дупликата. Уочено је да се исти наслови често јављају у записима који се разликују у свега неколико знакова, при чему је најчешћи узрок неуједначено коришћење знакова: у једном наслову се појам раздваја цртицом, а у другом размаком. Да би се такве неправилности препознале, као основна мера коришћено је Левенштајново растојање [16], уз неколико оптимизација без којих поступак не би могао да се спроведе у разумном рачунарском времену.

Прва оптимизација примењена је на саму функцију за рачунање растојања, увођењем раног заустављања. Поређење два текста је рачунски захтевно, па се функција прекида чим се утврди да је број неподударања већи од три знака. Граница од три знака одређена је емпиријски, тако да се не одбаци превише инстанци, а да поступак ипак буде осетно бржи.

Друга оптимизација спречава комбинаторну експлозију која би настала поређењем сваког наслова са сваким. Уместо тога, најпре се за сваки наслов израчуна префикс од четири знака, а потом се спајају само они наслови који деле исти префикс и исту годину објављивања. Иста година се захтева јер публикација са истим насловом објављена у две различите године не мора бити иста публикација. Додатно, не спајају се наслови чија се дужина разликује за више од пет знакова, јер би их функција са раним заустављањем ионако прогласила несличнима, нити се наслов спаја сам са собом, нити се исти пар образује два пута.

Након што су пронађени сви парови сличних наслова, потребно је од сваке групе задржати само један запис. Овде настаје проблем транзитивности: ако наслови A , B и C деле исти префикс и међусобно су довољно слични, добијају се парови $\{A, B\}$, $\{A, C\}$ и $\{B, C\}$. Ако би се из сваког пара издвојио по један представник, остао би скуп од више записа за које је већ утврђено да су међусобно слични.

Да би се то избегло, сличност је представљена неусмереним графом. Свакој инстанци се најпре додељује јединствен идентификатор, који постаје чвор графа, док грана између два чвора постоји ако је утврђено да су одговарајући

наслови довољно слични. Граф је неусмерен јер је релација сличности симетрична. Задатак се тиме своди на проналажење повезаних компоненти графа, при чему свака компонента представља једну групу дупликата из које треба задржати једног представника.

Повезане компоненте одређују се итеративно. Свакој инстанци се придружује вредност `rep_id`, која је на почетку једнака њеном идентификатору, а затим сваки чвор својим суседима шаље тренутно најмању познату вредност. Кроз неколико итерација најмањи идентификатор се прошири кроз целу повезану компоненту, тако да сви међусобно повезани дупликати добију исту вредност. Из сваке групе задржава се она инстанца која има највише елемената у низу описа, јер она носи највише информација.

Након уклањања дупликата, преостале инстанце захтевају и чишћење самих описа. Као и код наслова, из њих су уклоњене ознаке и ентитети језика *HTML*, вишеструки размаци и размаци са почетка и краја текста. Уочено је и да многи описи почињу речју **abstract**, па је и она избачена.

Осим појединачних делова реченица, уочено је да поједине реченице уопште нису информативне и да стварају додатан шум. Реченице које садрже изразе **peer review**, **issn**, **eissn**, **page**, **figure**, **http** или **https** углавном не доприносе садржају публикације него само понављају основне податке о њој или упућују на то где се о њој може сазнати више. Посебно су третиране реченице које садрже низ **pdf**, јер он може означавати и функцију густине вероватноће (енгл. *probability density function*), па је уведено додатно правило да реченица мора садржати и израз **available in** да би била уклоњена.

Тежиште је са уклањања делова реченица померено на уклањање целих реченица зато што поједини описи могу бити сачињени само од једне реченице која није информативна. Одбацивање целог описа само зато што има једну реченицу није било прихватљиво, јер су многи апстракти управо такви, па је дубљом анализом требало утврдити које реченице заиста нису информативне. Овај корак се показао корисним и при чишћењу дужих описа, а за све наведене провере коришћени су регуларни изрази.

Сваки члан низа описа потом је нормализован свођењем текста на мала слова, нормализацијом према стандарду Уникод, уклањањем сувишних знакова интерпункције и уклањањем стоп-речи. Из низа су уклоњени празни описи, као и инстанце чији је низ описа тиме постао празан. Да би се димензионалност додатно смањила, над сваким низом дужим од једног елемента спроведено је и

уклањање дупликата унутар самог низа.

Поступак подразумева да се задржавају најдужи описи, а одбацују они који су истоветни или садржани у неком дужем, као и они сачињени од само једне реченице. При препознавању делимично сличних описа коришћени су провера садржаности подниске и Левенштајново растојање са раним заустављањем, при чему се два описа сматрају истим ако им је сличност већа од 92%. За разлику од наслова, овде праг раног заустављања није фиксан број знакова него се рачуна из дужине самих описа, будући да су описи знатно дужи од наслова.

Када се фаза трансформације заврши, подаци се преводе у структуру `DataFrame` библиотеке *pandas* и чувају у датотеци формата *pickle*, како би при поновном учитавању били сачувани сви типови података. Тиме је припрема података завршена, а њихово поновно учитавање не захтева понављање скупог поступка чишћења. Резултат сваког часописа чува се одвојено, у посебној датотеци, чиме се обрада може прекинути и наставити без губитка већ обављеног посла.

Учитавање тако припремљених података претходило је примени модела машинског учења. Након учитавања подаци су проширени тако што је сваки члан низа описа постао засебна инстанца са истим насловом, будући да модели који следе не примају листу текстова као улаз. Свака инстанца је потом проширена још једном колоном, добијеном надовезивањем наслова и описа, и она је оно што у овом раду представља документ.

Глава 4

Формирање и опис тематских кластера

Пошто је циљ овог истраживања проналажење тема скривених међу мета-подацима публикација једног часописа, полазна тачка била је употреба модела за тематско моделовање. Испробана су два приступа: модел *BERTopic*, који представља приступ одозго-наниже, и алат *KeyBERT*, који представља приступ одоздо-навише. Ради приказа међурезултата у наставку је коришћен часопис из области рачунарства, као области сродне овом раду, под називом „*Semantic Web*“, за који је након чишћења остало 333 инстанци.

4.1 Приступ одозго-наниже

Модел *BERTopic* је, како је раније описано, оквир сачињен од више етапа које заједно дају тематску репрезентацију докумената [37]. Најпре је потребно изабрати модел за векторску репрезентацију докумената, а у овом случају употребљен је модел *SPECTER*, који добро моделује текст научног типа [28]. Добијене векторске репрезентације су нормализоване, како би се уклонио утицај апсолутне величине вектора и тежиште пренело на његов правац, односно на семантичку сличност.

Нормализацијом се постиже и то да еуклидско растојање и косинусна сличност дају исти поредак, јер за јединичне векторе u и v важи

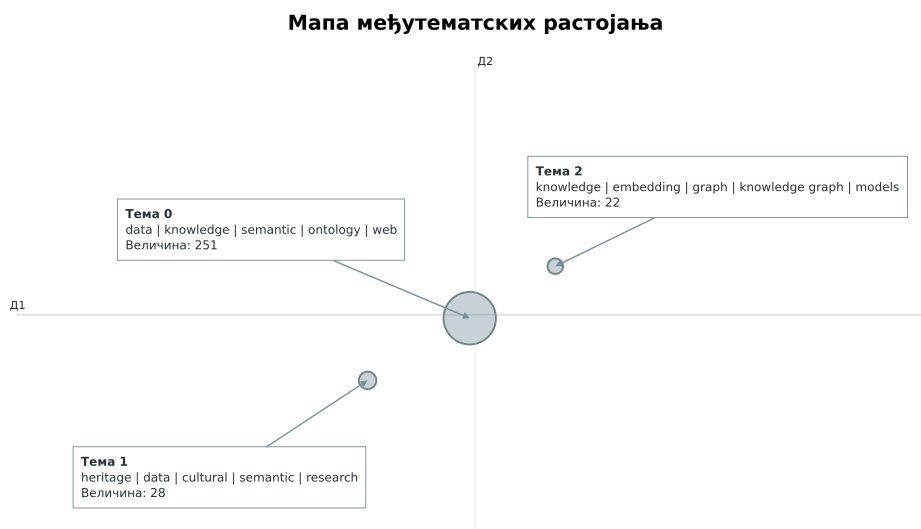
$$\|u - v\|^2 = 2(1 - \cos(u, v)) \quad (4.1)$$

одакле следи да је еуклидско растојање монотono опадајућа функција косинусне

сличности. Кластеровање засновано на еуклидском растојању тиме заправо групише векторе по семантичкој сличности.

Једна од етапа оквира користи и алгоритам за смањење димензионалности *UMAP* [31]. Пошто је реч о стохастичком алгоритму, а како би резултати били исти при сваком покретању, семе генератора случајних бројева постављено је на вредност 42. Последња етапа користи матрицу *c-TF-IDF* за издвајање речи које представљају тему сваког кластера, при чему подразумевано издваја само једночлане изразе. Како теме најчешће нису једна реч него синтагма, моделу је прослеђена измењена верзија текстуалног векторизатора, која издваја *n*-граме дужине од једне до три речи, односно униграме, биграме и триграме.

Резултат примене модела *BERTopic* на часопис „*Semantic Web*“ приказан је на слици 4.1. Модел је пронашао три теме, при чему највећој припада 75,38% докумената, односно 251 документ, док преосталим двама темама припада између 6,6% и 8,4% докумената. Преосталих око десет одсто докумената проглашено је шумом и не може се доделити ниједној од издвојених тема.



Слика 4.1: Теме које је модел *BERTopic* издвојио за часопис „*Semantic Web*“

Иако овакав резултат на први поглед изгледа добро, будући да је модел пронашао сличност у преко седамдесет одсто докумената, дубља анализа доводи до занимљивих запажања. На слици 4.2 дате су кључне речи које представљају сваку тему, заједно са њиховим релативним вредностима мере *c-TF-IDF*, које се користе за рангирање термина унутар теме. Два запажања посебно се издвајају и оба говоре о ономе што овај приступ не успева да обухвати.



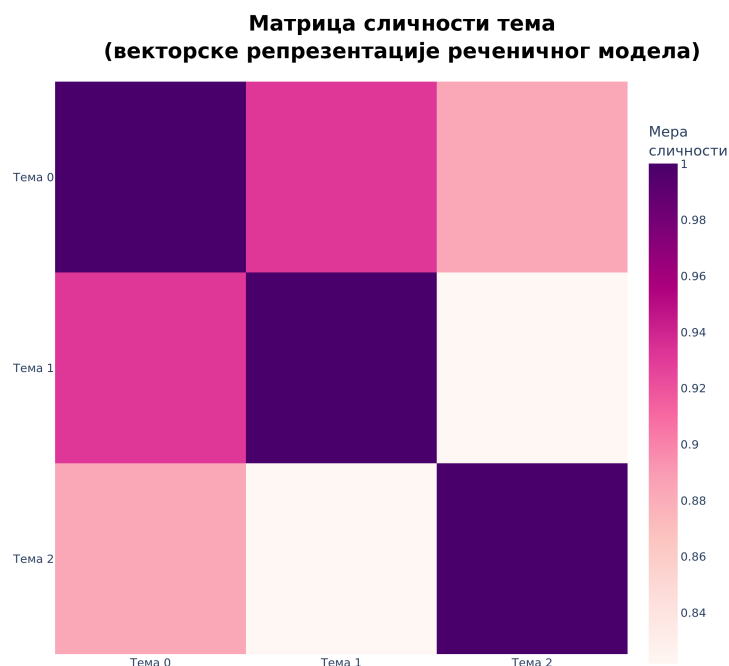
Слика 4.2: Кључне речи по темама и њихове вредности мере $c\text{-}TF\text{-}IDF$

У оквиру теме означене бројем два уочава се да се једнина и множина исте речи не третирају једнако, о чему сведоче парови **graph** и **graphs**, односно **embedding** и **embeddings**, што знатно смањује број појављивања истог појма при рачунању мере $c\text{-}TF\text{-}IDF$. Иако то на први поглед делује као пропуст у претпроцесирању, проблем је у томе што не постоји поуздан начин да се ове речи уједначе. Примена напреднијих поступака, уместо пуког одбацивања наставка са краја речи, доводи до непотпуних резултата: стемовање даје корен речи, али се притом губи њено семантичко значење, што може погоршати векторску репрезентацију.

Испробана је и лематизација, која такође није дала добар резултат на свим подацима. Један такав пример је реч **data**, која лематизацијом добија облик **datum**, чиме израз **linked datum** губи смисао. То је само један пример, а на нивоу више од сто педесет хиљада часописа из различитих области није могуће предвидети све изузетке који могу настати. Услед недостатка поузданог поступка уједначавања, прихваћено је да се такви облици посматрају као засебне речи.

Друго запажање односи се на сличност између самих тема, приказану на сликама 4.3 и 4.4. На првој се уочава да су све теме врло кохерентне и да је мера сличности између сваке две теме виша од осамдесет одсто. Та мера израчуната је на основу косинусне сличности векторских репрезентација тема, односно векторских репрезентација докумената који темама припадају.

На другој слици сличност између различитих тема мања је од шездесет одсто. Разлика потиче отуда што мера приказана на другој слици није заснована на семантичкој сличности, него на косинусној сличности вектора изведених из



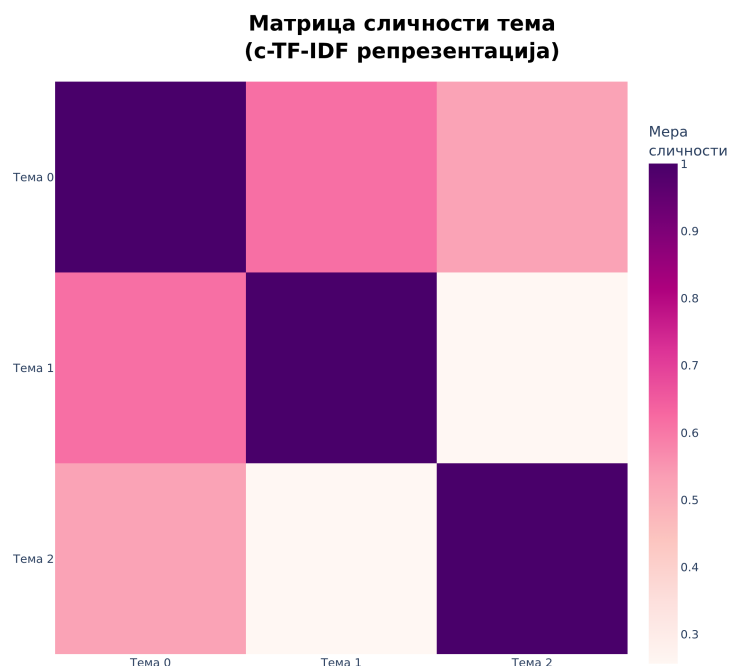
Слика 4.3: Топлотна мапа сличности тема на основу векторских репрезентација

матрице *c-TF-IDF*. Ова разлика показује да теме имају прилично блиску семантику, али да се у њима користе различити карактеристични термини.

Иста разлика уочава се и на слици 4.2, која приказује репрезентативне речи по темама. Из приказаног се види да се тема означена бројем нула може разумети као семантички веб и онтологије, док се тема означена бројем два може разумети као графови знања и њихове векторске репрезентације, што су врло блиске области. Уз то, обе теме деле реч **knowledge**, и то заједно са кластером који је означен као шум.

Овај проблем настаје зато што алгоритам *HDBSCAN* не кластерује појединачне речи него векторску репрезентацију целог документа. Документ може садржати реч **knowledge**, али и много других речи, услед чега његов укупни вектор није довољно близу ниједном постојећем кластеру. Пример једног таквог документа, означеног као шум, у ком се јавља реч **knowledge** и који говори о унапређењу таксономије графа употребом великих језичких модела, дат је у наставку.

„large language models creation enrichment evaluation taxonomic graphs.



Слика 4.4: Топлотна мапа сличности тема на основу матрице $c\text{-TF-IDF}$

taxonomies play crucial role organizing knowledge various natural language processing tasks. recent advancements large language models llms opened new avenues automating taxonomy-related tasks greater accuracy. paper explore potential contemporary llms learning evaluating predicting taxonomic relations multiple lexical semantic tasks. propose novel method creation taxonomy-based instruction datasets. use dataset based wordnet build taxollama unified model fine-tuned designed handle wide range taxonomy-related tasks taxonomy construction enrichment hypernym discovery lexical entailment. experimental results demonstrate model achieves state-of-the-art performance 11 16 tasks ranked second 4 tasks. explore llm ability constructed taxonomies graph refinement present comprehensive ablation study thorough error analysis supported manual automated techniques.“

Пошто је примена модела *BERTopic* кренула од рачунања векторске репрезентације целог документа, коју је потом покушала да кластерује, овај приступ решавању проблема издвајања тема назван је приступом одозго-наниже. Полази се, дакле, од целине и тражи се њена подела на теме. У томе је и његово основно ограничење, јер документ који припада већем броју области нема је-

динствен положај у простору векторских репрезентација.

4.2 Приступ одоздо-навише

Пошто је главна мана модела *BERTopic* то што сличност тражи између целих докумената, друга замисао заснива се на томе да се најпре издвоје речи које добро описују појединачни документ, па да се тек оне групишу. У овом приступу коришћен је алат *KeyBERT*, који рачуна векторску репрезентацију целог документа и пореди је са векторским репрезентацијама свих фраза од једне до три речи које се у том документу јављају [38]. Из сваког документа издваја се десет најрепрезентативнијих фраза, а архитектура самог алата приказана је на слици 2.3 у одељку 2.2.3.

Број фраза које се издвајају параметар је самог алата, а изабрана је вредност десет јер простор за касније кластеровање расте сразмерно броју публикација у часопису. За векторске репрезентације докумената и фраза коришћен је модел *SciNCL*, као поузданији избор за текст научне тематике [29]. Као најрепрезентативније издвајају се фразе са највећом вредношћу мере

$$score(k) = \cos(e(k), e(d)) \quad (4.2)$$

где је k фраза кандидат, d документ из ког је издвојена, а $e(\cdot)$ означава векторску репрезентацију. Уз сваку издвојену фразу алат враћа и њену вредност ове мере, али се она не сме тумачити као мера квалитета фразе. Тако фраза *knowledge graph applying* може имати већу вредност од фразе *knowledge graph* иако је мање употребљива, будући да дужа фраза носи више контекста, па њена репрезентација може бити ближа репрезентацији целог документа.

Проблем који се јавио код модела *BERTopic*, постојање једнине и множине истог појма, јавља се и овде. Разлика је у томе што су, због природе алата *KeyBERT*, овде издвојене готово све значајне кључне фразе, што омогућава да се поступак уједначавања унапреди и да се могућност погрешних облика знатно смањи. Другим речима, сам скуп издвојених фраза носи податак о томе који су облици у корпусу заиста уобичајени.

Уместо да се сваки облик своди на једину без провере, најпре се формира скуп свих појединачних речи које се јављају у издвојеним фразама. Свака фраза се потом раставља на речи, а за сваку реч се испитује њен облик у једини: ако тај облик постоји у поменутом скупу, реч се замењује њиме, а ако не

постоји, задржава се изворни облик. Тако се, на пример, реч *ontologies* своди на *ontology* само ако се облик *ontology* већ јавља међу издвојеним речима; у супротном то значи да су све издвојене фразе већ у множини, па сама промена облика не би донела уједначавање него ново неслагање. Проблем претварања речи *data* у *datum* и даље је могућ, али са знатно мањом вероватноћом, јер би облик *datum* морао да буде издвојен као кључна фраза неког документа. Примери таквог уједначавања дати су у табели 4.1.

Табела 4.1: Примери уједначавања облика издвојених фраза

пре нормализације	после нормализације
<code>knowledge graphs models</code>	<code>knowledge graph model</code>
<code>links entities</code>	<code>link entity</code>
<code>links rdf data</code>	<code>link rdf data</code>
<code>ontological concepts model</code>	<code>ontological concept model</code>

Када фразе прођу кроз уједначавање, за њих се рачунају нормализоване векторске репрезентације, како би, по угледу на оквир *BERTopic*, могле да се кластерују. Опонашајући тај оквир, најпре је спроведено смањење димензионалности методом *UMAP*, чије је семе генератора случајних бројева такође постављено на вредност 42, а потом и кластеровање алгоритмом *HDBSCAN* [34]. Разлика у односу на претходни приступ није у самом низу корака него у томе шта се кластерује: овде су то фразе, а не цели документи.

Након кластеровања свакој инстанци придружују се још две информације: којим све кластерима она припада и које су речи из кластера довеле до тог придруживања. Пошто је број кластера велик, будући да се кластерује десет пута више фраза него што има публикација, ове две информације помажу у избору најинформативнијих кластера. За сваки кластер може се пребројати у колико се докумената јављају речи из њега, па се кластери могу сортирати по том критеријуму.

Нека је D скуп свих докумената часописа, а $D(c)$ скуп докумената у којима се јавља бар једна фраза кластера c . Кластери се уреди опадајуће по величини скупа $D(c)$, чиме се добија низ $c_1, c_2, \dots$, а кумулативна покривеност после k кластера дефинише се као

$$Cov(k) = \frac{|\bigcup_{i \leq k} D(c_i)|}{|D|} \quad (4.3)$$

Као репрезентативни задржавају се први k^* кластера, где је k^* најмањи број за који важи $Cov(k^*) \geq 0,70$. Кластери се, дакле, додају један по један све док нови кластер доприноси унији већ покривених докумената, а поступак се зауставља када буде покривено седамдесет одсто различитих публикација.

Праг је постављен зато што за одређивање тема које се претежно јављају у часопису нису потребне теме заступљене у врло малом броју публикација. Такве теме говоре о појединачном раду, а не о профилу часописа. На примеру часописа „*Semantic Web*“, кластер са следећим кључним речима појавио се у само једном документу.

```
['combinational creativity iii', 'computational creativity artificial',
'creativity artificial', 'creativity artificial creativity',
'iii transformational creativity', 'regarding types creativity',
'related computational creativity', 'technologies creative ai',
'transformational creativity relevant', 'web computational creativity']
```

Очигледно је да овај кластер говори о рачунарској креативности и креативности вештачке интелигенције и да није у непосредној вези са тематиком часописа, па се, да не би стварао додатан шум, одбацује. Слично важи и за остале кластере који се јављају у свега једној или две публикације. Овим се поље даље анализе знатно сужава и у наставку се посматра само репрезентативан број кластера.

На слици 4.5 приказано је десет највећих репрезентативних кластера, заједно са највише десет најфреквентнијих фраза у сваком од њих, док је у загради дат број докумената које кластер покрива. Приказ је ограничен на десет кластера зато што их је за овај часопис укупно тридесет један, па би сви заједно заузели неколико страна, а облик расподеле види се већ и на овом делу. Код прва четири кластера најфреквентнија фраза задаје саму тему. Већ и сам број покривених докумената показује да међу кластерима постоји јасна разлика у значају, будући да прва четири кластера покривају вишеструко више докумената од оних на крају низа.

Даљим разматрањем уочава се да се поједине теме преклапају. То није случајност него последица избора параметара алгоритма *HDBSCAN*. Параметар `min_cluster_size` постављен је на вредност 5 и одређује најмањи број фраза које чине кластер, чиме се спречава да се свака мања група прогласи темом. Параметар `min_samples` постављен је на вредност 2 и одређује колико суседа тачка мора имати да би се сматрала унутрашњом тачком области веће густине;

ГЛАВА 4. ФОРМИРАЊЕ И ОПИС ТЕМАТСКИХ КЛАСТЕРА



Слика 4.5: Десет највећих репрезентативних кластера часописа „Semantic Web“ са најфреквентнијим фразама и бројем покривених докумената

ниже вредности дају мање строг критеријум и мање тачака означених као шум. Параметар *metric* постављен је на еуклидско растојање, што је оправдано тиме да су векторске репрезентације претходно нормализоване, у складу са једначином (4.1).

Преклапање се може и измерити. Од докумената који улазе у бар један репрезентативни кластер, нешто више од половине припада двама или више њих.

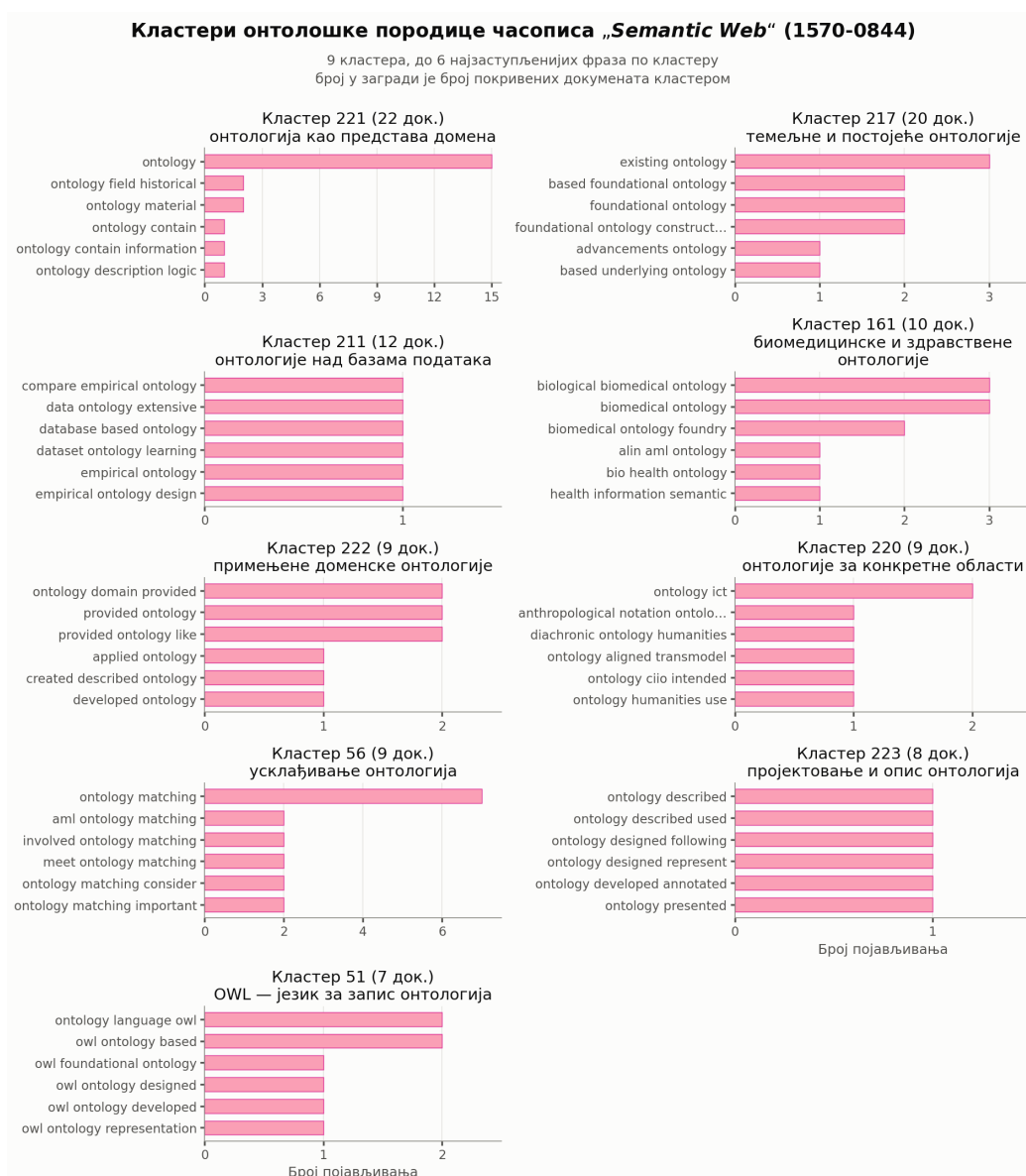
То значи да преклапање није изузетак него редовно својство овог приступа, што и одговара стварном стању у научним публикацијама, будући да један рад најчешће говори о више ствари одједном.

Избором ових вредности није се тежило великим кластерима него што прецизнијим кластерима на нивоу уже области. Најизразитији пример су кластери окренути онтологијама, приказани на слици 4.6. Анализом кључних речи уочава се да је кластер означен бројем 222 усмерен ка прављењу и коришћењу онтологије као представе неког домена, кластер означен бројем 217 ка коришћењу, повезивању и тумачењу постојећих и темељних онтологија, док је кластер означен бројем 220 усмерен ка онтологијама направљеним за конкретне области. Тумачење свих девет кластера ове породице дато је у табели 4.2.

Табела 4.2: Тумачење кластера онтолошке породице часописа „*Semantic Web*“

ознака кластера	тема кластера
221	онтологија као представа домена — сам појам онтологије и оно што она садржи
217	темељне и постојеће онтологије — њихово коришћење, комбиновање и повезивање
211	онтологије над базама података — приступ подацима посредством онтологије
161	биомедицинске и здравствене онтологије — онтологије у наукама о животу
222	примењене доменске онтологије — моделовање неког домена помоћу онтологије
220	онтологије за конкретне области — информационе технологије, материјали, хуманистика
56	усклађивање онтологија — довођење двеју или више онтологија у склад
223	пројектовање и опис онтологија — систематски развој и представљање онтологије
51	<i>OWL</i> — језик за запис онтологија и његова изражајна моћ

Кластери у табели уређени су опадајуће по броју докумената које покривају, па се већ из тог редоследа види да су уопштенија схватања онтологије заступљенија од оних везаних за једну област. Ако се кластери упореде међусобно, уочава се хијерархија која се креће од општијих ка специфичнијим

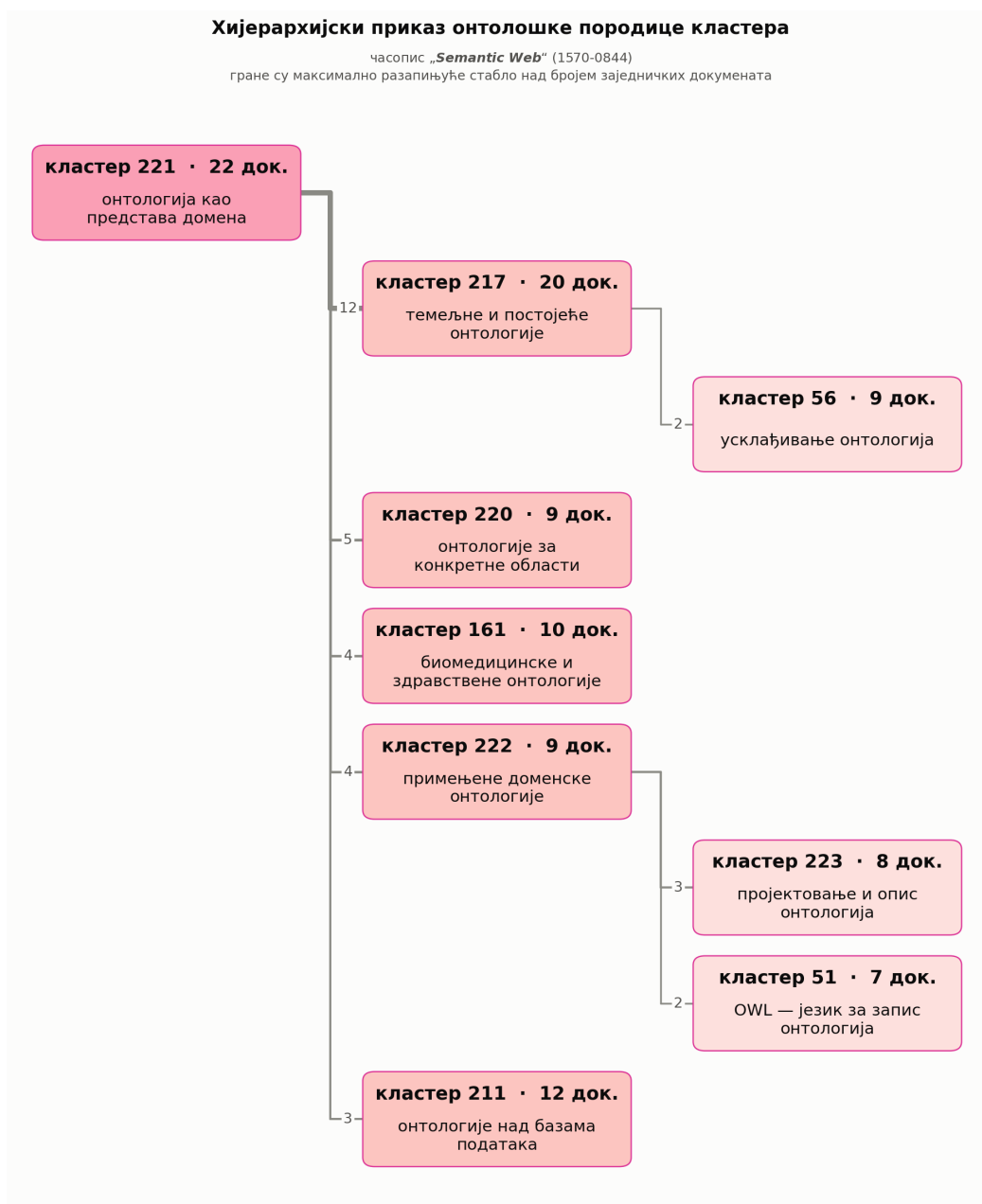


Слика 4.6: Кластери онтолошке породице часописа „Semantic Web“ са најфреквентнијим фразама

темама и која је приказана на слици 4.7. Ту хијерархију алгоритам не гради изричито, него она произлази из тога што се фразе опште природе јављају у више кластера.

Управо зато што настаје из преклапања, хијерархија се може и измерити, па гране на слици 4.7 нису претпостављене него добијене као максимално развијено стабло над бројем докумената које два кластера деле; број уз грану показује колико је тих заједничких докумената. Најјача веза стоји између кла-

стера који онтологију посматра уопштено и кластера који се бави темељним и постојећим онтологијама, што и одговара односу ширег и ужег појма. Овакве хијерархије кластера могу се уочити и у другим областима, што говори да поступак не издваја само равни списак тема него и однос међу њима.



Слика 4.7: Хијерархија кластера онтолошке породице часописа „Semantic Web“

Пошто је приступ заснован на алату *KeyBERT* кренуо од докумената појединачних публикација, односно од надовезаних наслова и сажетака, а по-

том груписао саме фразе издвојене из њих, овај приступ назван је приступом одоздо-навише. За разлику од претходног, овде један документ може припадати већем броју кластера, што одговара стварном стању у научним публикацијама. Цена тога је знатно већи број кластера, због чега је и био потребан поступак избора репрезентативних.

4.3 Одређивање носећих фраза и фраза подршке

Због уочене хијерархије међу кластерима, у наставку је било могуће применити и други степен кластеровања ради добијања ширих тема, али би се тиме изгубила специфичност сваког кластера. Уместо тога, приступило се проналажењу фраза које најбоље описују сваки кластер, односно сваку тему. Такве фразе у даљем тексту називамо носећим фразама.

Ако се за носећу фразу изабере најфреквентнија фраза кластера, за првих неколико кластера види се да она добро описује целу групу. Проблем настаје код кластера у којима су све фразе подједнако заступљене, какав је кластер означен бројем 185. Постоје и кластери у којима најфреквентнија фраза није добар представник, какав је кластер означен бројем 38, где би уместо фразе **reasoning approach** **iot** бољи избор биле фразе **urban iot** или **semantic iot**, јер боље представљају заједничко значење већег броја фраза, док у остатку кластера нема довољно јаког сигнала да је **reasoning** главна тема.

Други приступ заснива се на већ израчунатим векторским репрезентацијама фраза у сваком кластеру. Најпре се одреди центроид кластера, а потом фраза која му је по косинусној сличности најближа. Овим приступом кластер означен бројем 30 за најближу фразу добија **iot data**, што добро описује све фразе у њему, док кластер означен бројем 148 уместо фразе **biological biomedical ontology** добија општији опис **health ontology**.

Пошто је центроид у извесној мери побољшао квалитет избора, коначни поступак заснива се на комбинацији учестаности и дужине фразе, уз центроид као допунски критеријум. Полази се од запажања да фраза која добро описује кластер најчешће није фраза са највећом учестаношћу него она која се као целина понавља у више различитих фраза тог кластера. Такву фразу зато не треба тражити међу издвојеним фразама него међу њиховим деловима.

Нека је $P(c)$ скуп фраза кластера c , а U скуп свих издвојених фраза. За сваку фразу из кластера образују се све њене узастопне подфразе које и саме припадају скупу U , чиме се добија скуп кандидата. За фразу од n речи број таквих подфраза износи $n(n+1)/2$, дакле једна за униграм, три за биграма и шест за триграма. Подршка кандидата k дефинише се као број фраза кластера у којима се тај кандидат јавља, а његова оцена као

$$score(k) = supp(k) \cdot |k|^2 \quad (4.4)$$

где је $|k|$ број речи у кандидату. Квадрирање дужине уводи се да би се предност дала дужим и информативнијим изразима: фраза од две речи која се јавља у три фразе кластера претпоставља се фрази од једне речи која се јавља у истом броју фраза. Кандидати се потом уређују опадајуће по оцени, а при истој оцени по дужини и по подршци.

Ако највећу оцену дели највише пет кандидата, сви се задржавају као носеће фразе. Ако их је више, рачуна се центроид њихових векторских репрезентација и задржава се само онај кандидат који је центроиду најближи по косинусној сличности. Тиме се спаја оно што су претходна два приступа радила одвојено: учестаност и дужина одређују шта уопште улази у избор, а центроид разрешава случајеве у којима сама оцена није довољна.

Носеће фразе, међутим, описују само средиште теме. Да би опис кластера био потпунији, уз њих се бирају и фразе подршке, и то тако да свака доноси нешто ново у односу на оно што је већ описано. Полазни облик тог поступка био је похлепан: фразе кластера разматрају се редом опадајуће учестаности и прихвата се она која уводи бар једну реч још непокривену носећим фразама ни раније изабраним фразама подршке, све док се не сакупи осам фраза. Провера над стварним подацима показала је, међутим, да такав поступак, ограничен на фразе које се у кластеру јављају бар два пута, за део кластера уопште не даје опис.

Разлог је природа фраза које алат *KeyBERT* издваја. Оне су дугачке и врло специфичне, па се у оквиру једног часописа ретко понављају; у појединим кластерима свака фраза јавља се тачно једном. За такве кластере услов о најмање два појављивања одбацује све кандидате, због чега је опис остајао сведен на саму носећу фразу, што је премало за поступак мапирања који следи.

Друго запажање тиче се самих фраза. Алат *KeyBERT* гради n -граме над припремљеним текстом не водећи рачуна о границама реченица, па се међу

издвојеним фразама налазе и спојеви краја једне и почетка наредне реченице. Фразе `sparql interactive map` и `rdf propose` примери су таквих спојева: ни у једној публикацији не јављају се као непрекидан текст, а као опис теме наводе на погрешан закључак. Фраза коју корпус нигде не изговара не може бити ни његов опис.

Коначни поступак, приказан на слици 4.8, зато уводи две измене. Прва је да се фраза прихвата само ако за њу у публикацијама кластера постоји реченица у којој се она заиста јавља; поступак проналажења те реченице описан је у наставку. Друга је да се уместо једног праве три пролаза кроз кластер, уређена по строгости. У првом се разматрају само фразе које се јављају бар два пута. Ако после њега изабраних фраза има мање од пет, други пролаз понавља исти похлепни избор над фразама које се јављају једном. Ако их је и тада мање од пет, трећи пролаз допуњава списак најучесталијим преосталим фразама, овога пута без услова о новој речи, јер је у том тренутку важније да опис уопште постоји него да свака фраза буде нова.

Основа речи добија се стемовањем, чиме се избегава да се фразе које се разликују само по облику речи рачунају као различити садржај. Носеће фразе унапред се уклањају из скупа кандидата, а при истој учестаности кандидати се уређују лексикографски, како би избор био детерминистички и поновљив. Горња граница t од осам фраза држи опис кластера довољно кратким да буде употребљив као улаз у наредни корак, док доња граница m од пет фраза обезбеђује да ниједан репрезентативни кластер не остане описан само носећом фразом.

Провера да за фразу постоји реченица не служи само као услов при избору. Пронађена реченица чува се и уз саму фразу, па опис кластера, поред фраза, садржи и делове текста из којих су оне потекле. Фразе су, наиме, нормализоване и лишене интерпункције, па саме по себи не откривају у ком су контексту употребљене; реченица тај контекст враћа и чини опис разумљивим и читаоцу и великом језичком моделу који се у глави 5 користи за мапирање.

Реченица се тражи међу публикацијама које припадају датом кластеру. За сваку такву публикацију узимају се њене кључне фразе додељене том кластеру, а потом се редом испробавају три облика тражене фразе, од најпрецизнијег ка најширем: најпре кључна фраза те публикације која се нормализује тачно у тражену фразу, затим њен изворни, ненормализовани облик, и на крају сама нормализована фраза. Прва реченица у којој се неки од тих облика нађе —

```

УЛАЗ: кластер  $s$ , носеће фразе  $K$ , горња граница  $t$ , доња граница  $m$ 
ИЗЛАЗ: списак фраза подршке  $S$ 

покривено  $\leftarrow$  скуп основа речи свих фраза из  $K$ 
 $S \leftarrow$  празан списак

кандидати(услов):
    врати фразе кластера  $s$  које нису у  $K$  ни у  $S$  и
    испуњавају услов, уређене опадајуће по учестаности,
    а при истој учестаности лексикографски

похлепни_избор(списак):
    за сваку фразу  $r$  из списак:
        ако је  $|S| = t$ :
            изађи
        нове  $\leftarrow$  скуп основа речи фразе  $r \setminus$  покривено
        ако је нове празан скуп:
            настави са следећом фразом
        ако за  $r$  не постоји реченица у кластеру  $s$ :
            настави са следећом фразом
        додај  $r$  у  $S$ 
        покривено  $\leftarrow$  покривено  $\cup$  нове

похлепни_избор(кандидати(учестаност  $\geq 2$ ))
ако је  $|S| < m$ :
    похлепни_избор(кандидати(учестаност  $= 1$ ))
ако је  $|S| < m$ :
    за сваку фразу  $r$  из кандидати(без услова):
        ако је  $|S| = m$ :
            прекини
        ако за  $r$  не постоји реченица у кластеру  $s$ :
            настави са следећом фразом
        додај  $r$  у  $S$ 
        покривено  $\leftarrow$  покривено  $\cup$  скуп основа речи фразе  $r$ 

врати  $S$ 

```

Слика 4.8: Псеудокод за избор фраза подршке кластера

најпре у опису, а ако га тамо нема, у наслову — постаје реченица те фразе, и то у изворном облику, са интерпункцијом.

Средњи корак потребан је зато што се носећа фраза бира из скупа свих фраза корпуса, па често није самостална кључна фраза ниједне публикације него се јавља само унутар дужих фраза, и то у другом граматичком облику. Тако се носећа фраза *ontology* у једној публикацији јавља само као део фразе *ontologies based data*. Пошто нормализација описана у одељку 4.2 ради реч

по реч и чува њихов број и редослед, положаји речи у нормализованој и у изворној фрази поклапају се, па се изворни облик добија узимањем речи са истих положаја, у наведеном примеру *ontologies*.

Ако дословно поређење не успе, оно се понавља над текстом из ког су уклоњени сви знаци који нису слова ни цифре. Алат *KeyBERT* враћа фразе без интерпункције, док у тексту стоје цртице и тачке, па се фраза `rdf star dataset` тек овако проналази у изразу `rdf-star datasets`. Уклањање интерпункције служи, дакле, само за проналажење, док у опис кластера увек улази изворна реченица.

Опис сваког репрезентативног кластера тако чине број публикација које он покрива, његов удео у укупном броју публикација часописа, носеће фразе и фразе подршке, а уз сваку фразу и реченица у којој се она јавља. Уз носеће фразе чува се једна реченица када кластер има бар пет фраза подршке, а пет реченица када их има мање, како би оскудан опис био надокнађен већим бројем примера употребе носеће фразе. Пошто се исти поступак користи и као услов при избору фраза и као извор реченица, не може се десити да фраза прође избор а да у опису остане без реченице. Овако описани кластери чине улаз у главу 5, у којој се пресликавају на таксономију научних области.

Глава 5

Мапирање кластера на таксономију научних области

У даљем раду изабрана је таксономија коју користи *OpenAIRE*. Као референтна таксономија узет је други ниво изабране таксономије, који даје финију поделу области, а из којег се једноставно, сажимањем, изводи и први ниво. У наставку су приказана два приступа, приступ заснован на већ израчунатим векторским репрезентацијама текста и приступ заснован на коришћењу великих језичких модела.

5.1 Приступ заснован на сличности векторских репрезентација

Почетна идеја била је да се израчунају нормализоване векторске репрезентације изабране таксономије и да се на њих потом мапирају већ добијени репрезентативни кластери. Векторска репрезентација таксономије рачуната је истим моделом *SciNCL* који је коришћен и за формирање тематских кластера [29], чиме се обезбеђује да се и кластери и области налазе у истом простору. Репрезентација је рачуната на основу другог нивоа таксономије, где за сваку област постоје назив, опис и списак кључних речи.

Репрезентативни кластери, у ознаци K^* , односно првих k^* кластера изабраних поступком описаним у одељку 4.2, долазе у два облика. Први облик је скуп свих фраза које се налазе у једном кластеру, а други је представљање кластера носећим фразама и фразама подршке. За сваки од та два облика посебно су

израчунате нормализоване векторске репрезентације кластера, такође моделом *SciNCL*: кластеру c придружује се вектор $e(c)$, добијен применом модела на текстуални опис кластера у датом облику. Пошто су обе репрезентације у истом векторском простору, два облика могу се непосредно упоредити.

За сваки кластер из скупа K^* рачуна се косинусна сличност са сваком облашћу другог нивоа таксономије. Најпре је постављен услов да сличност мора бити већа од 0,90 да би се за неки кластер рекло да припада изабраној области. Тај праг показао се претростогим, јер многи кластери, па и читави часописи, нису успели да остваре тако високо поклапање ни са једном облашћу.

Због тога су покренуте две варијанте овог поступка, које ублажавају првобитно постављени услов. Прва варијанта задржава праг, али га спушта на емпиријски одређену вредност, док друга варијанта праг у потпуности напушта и за сваки кластер задржава три најближе области. Обе варијанте покренуте су над истим скупом кластера, тако да се њихови резултати могу непосредно упоредити.

Праг у првој варијанти одређен је у два корака. Нека је T_2 скуп области другог нивоа таксономије, $e(\cdot)$ векторска репрезентација, а K_j^* скуп репрезентативних кластера часописа j . За сваки $c \in K_j^*$ најпре се одреди сличност са његовој најближој области

$$s(c) = \max_{a \in T_2} \cos(e(c), e(a)) \quad (5.1)$$

а потом се, као једна вредност по часопису, узима медијана тих сличности, $m_j = \text{med}_{c \in K_j^*} s(c)$. Она представља вредност за коју већина кластера тог часописа може бити придружена бар једној области. У другом кораку узима се медијана тих вредности по свим часописима из посматраног скупа J :

$$\tau = \text{med}_{j \in J} m_j \approx 0,82 \quad (5.2)$$

Друга варијанта не користи праг већ за сваки кластер из скупа K^* издвајају се три области таксономије које имају највећу косинусну сличност са тим кластером. Поред саме области, чува се и вредност косинусне сличности између те области и испитаног кластера, јер она улази у наредни корак. Тиме сваки кластер увек добија тачно три придружене области, што ће се касније показати и као извор једне врсте грешке.

У обе варијанте алгоритма на крају се пролази кроз све кластере из скупа K^* и води се мапа која свакој области таксономије придружује њен допринос.

Допринос области a рачуна се кумулативно, множењем косинусне сличности те области са кластером и покривености часописа тим кластером:

$$w(a) = \sum_{c \in K^*} \cos(e(c), e(a)) \cdot cov(c) \quad (5.3)$$

Тежина, односно покривеност часописа $cov(c)$, уведена је да би већи кластери више доприносили резултату, будући да је њихова тема шире распрострањена кроз цео часопис. Добијени доприноси се нормализују и уређују опадајуће, од области са највећим ка области са најмањим доприносом. Допринос области A првог нивоа таксономије добија се сабирањем нормализованих доприноса свих области другог нивоа које јој припадају:

$$W(A) = \frac{\sum_{a \in A} w(a)}{\sum_{b \in T_2} w(b)} \quad (5.4)$$

Главни проблем оба приступа јесте недовољно специјализована таксономија. Она је преопшта, не садржи кључне речи из појединих ужих области и вероватно не покрива све области које се у пракси јављају. На примеру часописа „*Virtual Reality*“, иако носећа фраза дословно гласи **virtual reality**, мера косинусне сличности сврстава ову тему најближе уметности.

На примеру часописа из области науке о материјалима резултати сврставају часопис готово подједнако у области физике, хемије и нано-технологије, а тек у малом проценту у област инжењерства материјала. Иако би такав резултат донекле био тачан, поставља се питање да ли наука о материјалима треба да буде засебна област таксономије и како уопште закључити да добијени резултат можда указује на нову тему. Ово је ограничење које се не може решити избором модела него само допуном саме таксономије.

На примеру часописа „*Semantic Web*“, осим што ниједна област не постиже сличност већу од 0,90 ни са једним кластером, при прагу од 0,82 добија се да је само један кластер довољно близак области математике. То не искључује могућност да је та област добро одређена за тај кластер, али она свакако није добра област за сам часопис. Друга варијанта, са избором три најближе теме по кластеру, јасно истиче да овај часопис припада области рачунарске и информационе науке.

Проблем је у томе што се као друга по значају област за часопис „*Semantic Web*“ појављује хемија. Дубљом анализом кластера уочава се да се та област увек јавља као друга или трећа област кластера, никада као прва, при чему

је разлика у односу на прву изабрану област често мања од 0,02. То је непосредна последица тога што поступак сваком кластеру мора доделити најмање три области, па и онда када за то нема стварног основа.

Кластери у којима је ово понашање препознато садрже фразе попут *ontology*, *knowledge representation*, *semantic web* и *knowledge graph*. Са друге стране, опис области хемије садржи речи *structure*, *representation* и *model*, које су заједничке и хемији и формалном представљању знања. Модел *SciNCL* обучаван је над научним текстом уопште, па му изрази *ontology* и *molecular structure* делују сродно, будући да оба говоре о структури и о односима међу ентитетима.

Да би се тема часописа прецизно одредила поређењем векторских репрезентација области и издвојених фраза часописа, пре свега се мора унапредити сама таксономија дефинисаних области. Овај приступ ипак има две значајне предности: детерминистичан је, па при истом улазу увек даје исти излаз, и лако се поново примењује. Промена таксономије захтевала би само да се једном изнова израчунају векторске репрезентације нових области, док би сам поступак остао непромењен.

5.2 Приступ заснован на великом језичком моделу

Полазна замисао била је да се изабраном великом језичком моделу проследи други ниво изабране таксономије и репрезентативни кластери, на основу којих би модел одлучио којим областима часопис припада. Упит који се шаље моделу може да обухвати највише 32.768 токена [51], а број токена потребан за кластере представљене свим фразама и сам по себи може прећи ту границу, па заједно са описом таксономије премашује оно што модел може да обради; то је показано и у резултатима, у одељку 6.3. Због тог ограничења поступак је разбијен на два дела, односно на два позива великог језичког модела.

Први корак подразумевао је да се моделу остави слобода да сам именује теме које пронађе у репрезентативним кластерима и да за сваку тему да оцену у којој мери она одговара кластерима. Оцена се заснива на покривености кластера, која се овде мери бројем публикација које припадају том кластеру и не своди се на сто одсто. У упутству су постављени и захтеви да модел издвоји једну

или више што ширих научних дисциплина, да их уреди по оцени и да одговор врати као ваљан *JSON* запис.

Од модела се тражи и да наведе разлог на основу ког је извео тему, као и главне теме целог часописа, при чему се разлог заснива на уоченим фразама из кластера. Током испитивања уочено је да поступак генерисања разлога може да упадне у бесконачно понављање исте фразе, све док се не исцрпи највећи дозвољени број токена одговора. Узрок је похлепно декодирање [54]: модел у сваком кораку бира најизвеснији наредни токен, па када једном уђе у низ који сам себе потврђује, у самом поступку генерисања нема ничега што би га из тог низа извело.

Овај проблем покушан је да се реши на два начина. Оба полазе од претпоставке да се понављање може спречити самим обликом упита, без измене поступка генерисања. Испоставило се да та претпоставка не важи.

- Постављањем ограничења у самом упиту да се фразе из кластера не понављају у објашњењу разлога. Ово није донело побољшање, јер је ограничење исказано у упиту, а не у поступку генерисања, па га модел поштује само у мери у којој је обучен да прати упутства; код модела са малим бројем параметара, какав је *Qwen2.5* са три милијарде параметара, та мера је ниска.
- Уклањањем поља за разлог. Проблем је у томе што то поље моделу даје потпуну слободу, па он не зна када да стане. Тиме проблем није решен него само померен, будући да и поље у ком модел издваја главне теме часописа даје исту слободу и доводи до истог понављања.

Сагледавањем упита уочено је да и поље за навођење научне области може бити похлепно. У упиту не постоји затворена листа научних области из које модел бира одговор, него му је остављено да га сам изведе. Због тог самосталног закључивања овај позив модела називамо кораком индукције.

У кораку индукције часопис може добити неваљан одговор настао пресецањем. Пресецање значи да је модел прекинуо генерисање пре него што је завршио набрајање уочених тема, као последица малог броја дозвољених токена у одговору. Пошто уз сваки додатни кластер модел добија нов повод да наведе још једну ужу област, дужина одговора расте са бројем кластера, па је пресецање вероватније управо код часописа са више кластера.

Пошто није пронађен јединствен начин да се проблеми похлепног декодирања отклоне, одлучено је да се корак индукције изводи у два наврата. Ако за неки часопис модел не врати ваљан одговор, покреће се поново, али са другачијим захтевима. Ако је узрок пресецање, довољно је повећати број токена одговора и поновити исти упит; ако је узрок бесконачно понављање, повећање броја токена не помаже, јер ће модел поново вртети исте фразе док их не исцрпи, па се шаље сведени упит у ком се не траже поља која су се показала као извор понављања.

И при сведеном упиту модел може наићи на иста два проблема. Понављање се тиме не разрешава, јер нема начина да се ограничи број научних области, док пресецање даје ваљан али непотпун одговор. Повољна је околност што је код пресецања уочено да оцена научне области монотono опада и да је при крају знатно нижа него код прве наведене области, одакле следи да модел није успео да обухвати финију поделу која се ређе јавља, али да је главне области ипак препознао.

Коначан одговор за такав часопис добија се накнадним сређивањем *JSON* записа. Последња област на којој је модел стао задржава се са истом оценом као претходна, али само ако је њен назив до краја исписан. Овим поступцима опоравка знатно већи број часописа добија употребљив резултат, уместо да буде одбачен као неваљан.

Корак индукције покреће се засебно за два облика представљања кластера из скупа K^* : једном за представљање носећим фразама и фразама подршке, а други пут за представљање свим фразама. Пошто и улаз мора да стане у дозвољени број токена, за сваки часопис се проверава да ли је то ограничење испуњено. Када се ограничење премаши, модел одсеца део улаза, а пошто су кластери уређени по броју публикација у којима се јављају, одсеца се управо реп са кластерима најмање покривености.

Такво одсецање уноси грешку у резултат, али теме које су широко распрострањене остају део упита. Управо зато је значајно што представљање носећим фразама и фразама подршке довољно добро апроксимира представљање свим фразама, јер се њиме у исти број токена смешта већи број кластера. Тиме се и мањи део садржаја губи одсецањем.

Два облика представљања подразумевају и два различита упита. Захтеви су у оба случаја исти, да модел изведе једну или више што ширих научних дисциплина, да их уреди од најзаступљеније до најмање заступљене и да свакој

додели оцену која одговара покривености кластера, а разликују се само улазни подаци. Када се корак индукције заврши, сви резултати се чувају у табели и представљају улаз за наредни корак.

Други корак састоји се од прослеђивања резултата претходног позива у нови позив модела. Уз њих се кроз упит прослеђују и категорије другог нивоа изабране таксономије, а од модела се тражи да сваку научну област пронађену у кораку индукције придружи категоријама које јој највише одговарају. Улаз чине пронађене области, оцена додељена свакој од њих и главне теме часописа, док се разлог одбацује, јер би само повећао број улазних токена и поновио фразе из којих је област већ изведена.

Пошто су научне области из корака индукције потпуно произвољне, не може се очекивати да се пресликају једна на једну на области таксономије, па се моделу допушта да једну пронађену област додели већем броју категорија. Од модела се тражи да у резултату наведе шифре области другог нивоа, из којих се потом сажимањем добија и први ниво. Како овај корак пресликава индуковане теме у затворену листу од 40 категорија таксономије, називамо га кораком мапирања.

Одговор корака мапирања такође мора бити ваљан *JSON* запис, са шифрама пронађених категорија и, уз сваку, списком области из корака индукције које ту шифру објашњавају. У списак се уписује само редни број пронађене области, чиме се смањује могућност грешке модела. Корак мапирања покренут је само над оним часописима за које је корак индукције, укључујући и све поступке опоравка, дао ваљан одговор.

Глава 6

Приказ и дискусија добијених резултата

У овом поглављу најпре је описан подкуп података над којим је испитивање вршено, заједно са референтним изворима према којима се тачност приступа може мерити. Након тога приказани су резултати оба испитана приступа, уз дискусију сваког од испробаних корака. У последњем делу разматра се анализа тематских промена часописа кроз време.

6.1 Репрезентативан узорак часописа

Скуп часописа над којим је анализа покренута описан је паровима бројева *ISSN* и *EISSN*, назив часописа, категорије којима часопис припада према класификацији *Scimago*, као и категорију часописа према класификацији *Science-Metrix*. Свака од те две класификације има другачију улогу при провери резултата.

Класификација *Science-Metrix* даје само једну репрезентативну категорију по часопису и као таква је добар избор за референцу при одређивању примарне теме. *Scimago*, са друге стране, даје више категорија и поткатегорија по часопису, уређених лексикографски. Она зато може послужити за потврду примарне теме и за потврду секундарних тема, али не и за одређивање тога која је тема доминантнија, будући да поредак није по значају.

Репрезентативни узорак чини 80 часописа, при чему је јединица узорка један часопис. Часописи су бирани насумично из сваке од шест области првог нивоа таксономије коју користи *OpenAIRE*, међу онима који у подацима имају

довољну количину публикација. Довољном количином публикација сматра се часопис који има између 500 и 3.000 инстанци, при чему је горње ограничење постављено због временске сложености поступка, али и због ограничења великог језичког модела, односно броја токена које он може да обради. Такво ограничење је оправдано и тиме што се за часописе са већим бројем публикација може посматрати ужи распон година.

Узорак је подељен на шест делова, по један за сваку област првог нивоа таксономије, и по њима су делови и добили називе. У сваком делу налазе се часописи у чијем се опису појављује област из назива тог дела. Расподела часописа по областима приказана је у табели 6.1.

Табела 6.1: Расподела репрезентативног узорка часописа по областима првог нивоа таксономије

назив области у таксономији	назив области на српском	број часописа
<i>Natural sciences</i>	природне науке	10
<i>Engineering and technology</i>	инжењерство и технологија	16
<i>Medical and health sciences</i>	медицинске и здравствене науке	15
<i>Agricultural sciences</i>	пољопривредне науке	14
<i>Humanities</i>	хуманистичке науке	11
<i>Social sciences</i>	друштвене науке	14
укупно		80

Због насумичног избора часописа, 78 од 80 часописа има класификацију у бази *Scimago*. База *Science-Metrix* обухвата 34.655 часописа, међу којима се налази 74 часописа из репрезентативног узорка. У пресеку, 73 часописа имају обе класификације, док само један часопис нема ниједну.

6.2 Резултати приступа заснованог на сличности векторских репрезентација

Поступак је за сваки часопис из репрезентативног узорка дао осам резултата, приказаних у табели 6.2, где је уз сваки наведено на који је начин кластер представљен и којим је поступком резултат добијен. Сваки резултат садржи уређен списак области са припадајућим уделима, добијен поступком описаним једначинама (5.3) и (5.4).

Табела 6.2: Резултати приступа заснованог на сличности векторских репрезентација

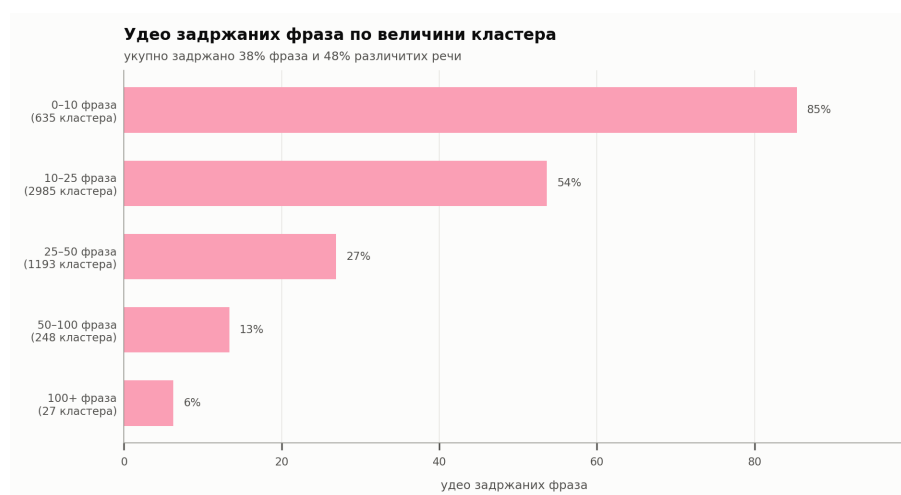
представљање кластера	ниво таксономије	поступак одређивања области
носеће фразе и фразе подршке	други	три области другог нивоа са највећом косинусном сличношћу са кластером
носеће фразе и фразе подршке	други	области другог нивоа чија је косинусна сличност са кластером већа од прага 0,82
носеће фразе и фразе подршке	први	сажимање области добијених избором три најсличније области, по једначини (5.4)
носеће фразе и фразе подршке	први	сажимање области добијених применом прага 0,82, по једначини (5.4)
све фразе	други	три области другог нивоа са највећом косинусном сличношћу са кластером
све фразе	други	области другог нивоа чија је косинусна сличност са кластером већа од прага 0,82
све фразе	први	сажимање области добијених избором три најсличније области, по једначини (5.4)
све фразе	први	сажимање области добијених применом прага 0,82, по једначини (5.4)

Приликом излагања поступка у одељку 5.1 објашњене су и главне мане овог приступа, а на примеру више часописа размотрена је и тачност резултата. Осим тачности, овај приступ омогућава и упоредно посматрање два облика представљања кластера из скупа K^* . Поређењем резултата поступка покренутог само над носећим фразама и фразама подршке, које ћемо у наставку краће називати и сажетим описом кластера, са резултатима поступка покренутог над свим фразама може се измерити колико добро сажети опис апроксимира цео кластер. Оба поступка могу подједнако грешити, али је идеја да се провери да ли и када греше, праве грешке у истом смеру.

Број фраза у скупу носећих фраза и фраза подршке готово је сталан и његова медијана износи 9, док се број свих фраза кластера креће од неколико

десетина до неколико стотина. Због тога је најпре испитано који удео целог скупа фраза описује сажети скуп. Просечан број фраза по кластеру у сажетом скупу износи 8,4, а у скупу свих фраза 22,0, што даје однос од 38% и означава количину текста коју сажети опис задржава. Код великих кластера одбацује се знатан део фраза, док се код мањих кластера, којих је знатно више, задржава преко половине текста. Удео задржаних фраза по величини кластера приказан је на слици 6.1.

Пошто модел *SciNCL* има ограничење на број токена које може да обради, значајно је размотрити и дужину текста. Ако се све фразе једног кластера поређају у низ, тај низ у сажетом скупу има у просеку 197,8 знакова, а у скупу свих фраза 527,8 знакова. Однос ове две вредности износи 38%, готово исти као однос броја фраза, и показује да је количина текста коју модел треба да обради у сажетом опису знатно мања.

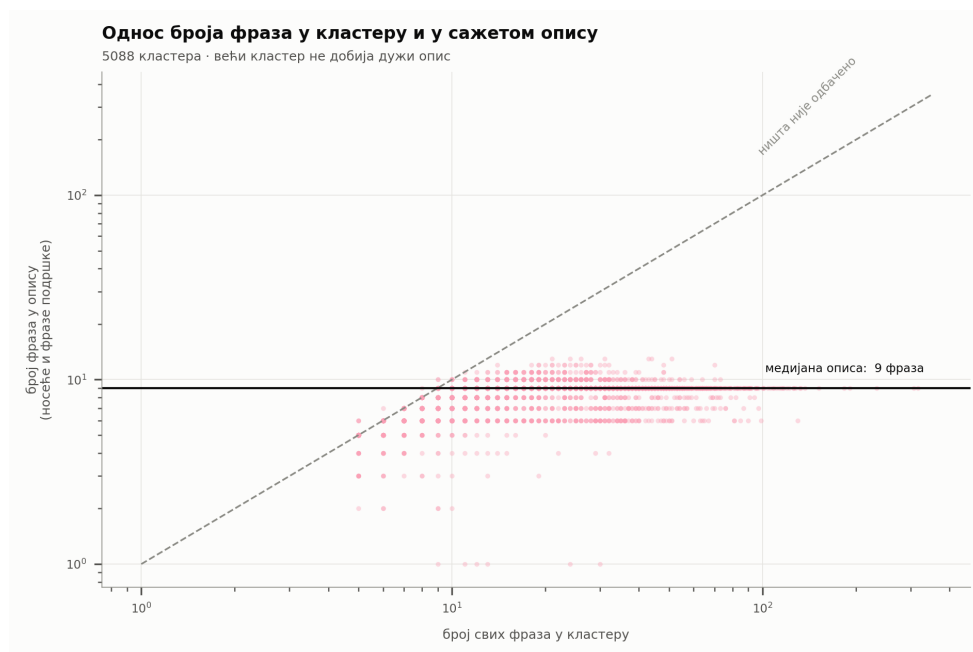


Слика 6.1: Удео задржаних фраза у зависности од величине кластера

Графикон на слици 6.2 приказује сваки пронађени кластер на скупу репрезентативног узорка часописа као једну тачку, чије су координате укупан број фраза у кластеру и број фраза у сажетом скупу. Тачке које леже на дијагонали значе да сажимање није одбацило ниједну фразу. Ако се из сваке тачке повуче нормала на дијагоналу, растојање између тачке и дијагонале представља број фраза одбачених у том кластеру.

Поједине тачке налазе се и изнад дијагонале, што значи да сажети скуп има већи број фраза него скуп свих фраза кластера. То је могуће због начина на који се одређује носећа фраза, будући да она не мора бити једна од фраза кластера него може бити и њена подфраза. Приметно је и то да описано растојање расте

са порастом броја фраза у кластеру, док је код малих кластера, којих је знатно више, оно осетно мање.



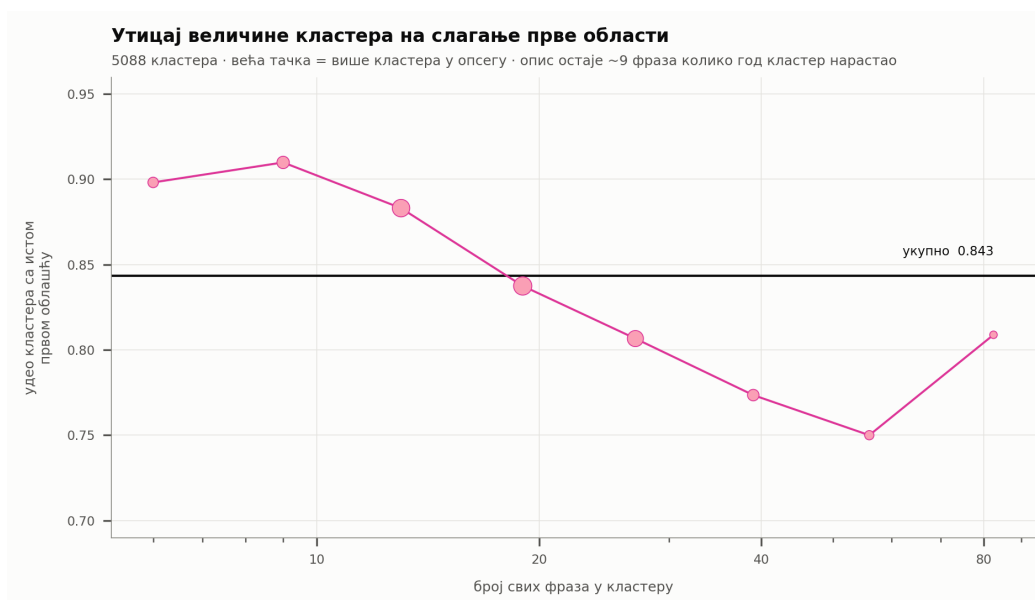
Слика 6.2: Однос укупног броја фраза и броја фраза у сажетом опису кластера

Осим количине, при сажимању текста значајан је и квалитет задржаног садржаја. У сажетом скупу у просеку се јавља 12,2 различитих речи, док их у скупу свих фраза има у просеку 25,6 по кластеру. Однос од 47,7% показује колико је јединствених речи сажети опис обухватио у односу на цео скуп фраза. Ова мера је значајна јер говори да се, када је број јединствених речи у кластеру мали, фразе граде понављањем истих речи, па сам број фраза прецењује количину информација којом кластер располаже, те се њиховим одсецањем квалитет не умањује.

Квалитет се потврђује и тиме што 84,3% кластера дели прву, односно косинусно најсличнију пронађену област другог нивоа таксономије, и на целом и на сажетом опису. Ова мера показује да сажети кластери имају високо поклапање са целим кластерима упркос значајном одсецању. На првом нивоу таксономије чак 91,8% кластера дели исту прву област, што значи да се, и када се разликују, најчешће разликују за област унутар исте групе првог нивоа, а не преласком у другу науку.

На слици 6.3 приказан је удео кластера код којих се прва придружена област поклапа између сажетог и целог описа кластера, за пет група кластера обра-

зованих према броју фраза које садрже. Са слике се уочава да поклапање монотono опада са порастом броја фраза у кластеру, све до четврте групе, што одговара томе да удео задржаног текста по величини кластера монотono опада. Последња група обухвата кластере са више од 80 фраза, где се уочава благи пораст, али он се не може узети у обзир јер је број кластера у тој групи мали. Рачунање интервала поверења Вилсоновим поступком¹ показује да би поновљено мерење на другом скупу часописа, које би дало исти број од 64 велика кластера, дало поклапање на првој теми између 0,66 и 0,86. На основу тога може се претпоставити да би поклапање врло вероватно наставило да опада.



Слика 6.3: Поклапање прве придружене области у зависности од величине кластера

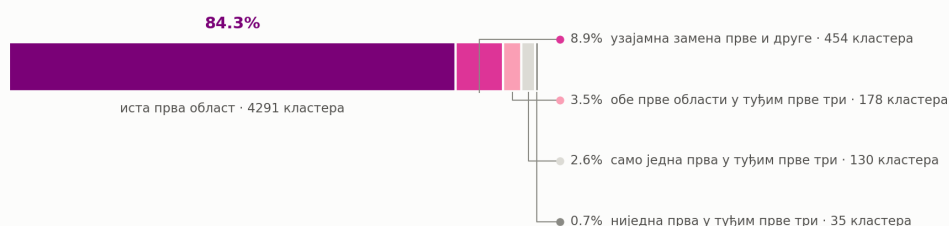
Слика 6.4 приказује како се односе прве изабране области таксономије по кластеру за два облика представљања кластера. Осим високог поклапања у првој области, које износи 84,3%, преостала грешка од 15,7% може се поделити на четири врсте, приказане у наставку. Врсте су поређане по озбиљности, од најблаже до најтеже, будући да не носе исту последицу по коначно одређену област часописа.

¹Вилсонов интервал изабран је зато што на малим узорцима даје поузданије границе од уобичајеног нормалног приближења, које при уделима блиским јединици може дати и границе изван интервала $[0, 1]$. Треба имати у виду да поступак претпоставља међусобно независна мерења, што овде није у потпуности испуњено, будући да сви кластери потичу из истог скупа часописа.

- У 8,9% случајева прва и друга област замене места у два облика представљања кластера.
- У 3,5% случајева прва област једног облика налази се међу прве три области другог облика, и обрнуто.
- У 2,56% случајева прва област једног облика налази се међу прве три области другог облика, али обрнуто не важи.
- У 0,69% случајева прва област једног облика не налази се ни у једној од три области другог облика.

Однос неслагања прве области у две репрезентације кластера

свих 5088 кластера · пет искључивих скупова, све мерене у ОБА смера
84.3% има исту прву област, 96.8% се поклапа бар кроз прве три · обостраног неслагања има 0.7%
сиво је стварно неслагање, розе су степеници блажег слагања



Слика 6.4: Однос првих изабраних области у два облика представљања кластера

Однос величине кластера и врсте грешке приказан је у табели 6.3, где су вредности изражене у процентима укупног броја кластера. Кластери су подељени у пет група према броју фраза које садрже. Уз сваку групу наведен је и број кластера у њој, јер од њега зависи колико се добијеним уделима може веровати.

У табели се види тенденција пораста сваке врсте грешке са порастом величине кластера, при чему за групу са више од 80 фраза и даље важи да се мери на малом узорку кластера. Ако се свакој врсти грешке придружи тежина, тако да замена има најмању, а изостанак сваког поклапања највећу, уочава се да са порастом величине кластера расте и тежина грешке. Код малих кластера долази до просте замене сродних области, што може бити последица тога да изабрана таксономија на другом нивоу није довољно раздвојена и да међу областима има доста преклапања.

ГЛАВА 6. ПРИКАЗ И ДИСКУСИЈА ДОБИЈЕНИХ РЕЗУЛТАТА

Табела 6.3: Врсте грешке у зависности од величине кластера, у процентима

број фраза у кла- стеру	број кластера	замена	обе међу прве три	само једна
0–10	635	7,6	1,3	0,8
10–20	2.241	8,2	2,9	1,8
20–40	1.721	9,8	4,2	3,4
40–80	427	11,9	6,8	4,7
преко 80	64	6,2	4,7	9,4

Код великих кластера однос замене и изостанка сваког поклапања износи око 50%, што значи да поступак ту знатно чешће греши. На слици 6.5 могу се видети и области које су најчешће изазвале замену прве и друге области у два облика представљања кластера. Управо те области указују на места на којима таксономија није довољно раздвојена.



Слика 6.5: Области које најчешће изазивају замену прве и друге области у два облика представљања кластера

Из ове анализе може се закључити да сажети опис добро апроксимира цео кластер, али и да би замена сталног броја фраза у сажетом опису бројем сразмерним величини кластера могла додатно унапредити резултате. Грешка је неминовна, јер се било којим поступком сажимања мора изгубити део информације. Ипак, при одсецању од око 62% фраза, сажети кластери и даље добро апроксимирају изворне.

6.3 Резултати приступа заснованог на великом језичком моделу

Након покренутог корака индукције добијени су резултати приказани у табели 6.4. Сваки ред одговара једном исходу, односно поступку који је био потребан да би се за часопис добио употребљив одговор. Последњи ред даје укупан број часописа за које је добијен ваљан резултат.

Табела 6.4: Исходи корака индукције по поступку опоравка

исход	поступак	носеће фразе и фразе подршке	све фразе
исправно из првог покушаја	први пролаз дао ваљан <i>JSON</i> запис, ништа се не мења	65	47
исти упит са већим ограничењем	пресецање, исти упит уз повећање са 800 на 2.500 токена	0	1
сведени резервни упит	понављање, упит без поља за разлог, главне теме и потпорне кластере, само област и оцена	8	13
опорављено из неваљаног записа	без новог позива модела, пресечени одговор се затвара, а недовршена последња област се одбацује	5	9
није прошло	ниједан од два упита, ни поступак опоравка	2	10
укупно ваљаних		78 од 80	70 од 80

Од девет часописа опорављених из неваљаног записа при представљању свим фразама, у осам случајева реч је о пресецању, а у једном о неисправном облику записа. Представљање кластера носећим фразама и фразама подршке чини улаз у модел мањим, па већи број часописа прође већ при првом покретању. Из приложених резултата види се и то да повећање броја излазних токена није донело значајну промену, што значи да би решење било смањење улаза, а не проширење излаза.

Највећи добитак доноси сведени резервни упит, који уклања поља што најчешће изазивају понављање. Разлика између часописа означених као опорављени

и оних који нису прошли уочава се у понашању оцене придружене пронађеним областима. Код опорављених оцена монотono опада, па се одсецањем уклања само реп области мањег значаја, док се код часописа код којих оцена не опада не може сматрати да је корак индукције пронашао све главне области.

Главни упит, заједно са сведеним, даје око пет области по часопису. Код часописа чији је одговор опорављен из неваљаног записа тај број је знатно већи, између 33 и 40, али то нису нове теме часописа. Реч је о последици похлепног декодирања: модел је упао у понављање и наставио да ниже готово истоветне области све док одговор није пресечен ограничењем броја излазних токена. Управо зато се при сређивању записа задржава само почетни део списка, у ком оцена монотono опада, док се поновљени реп одбацује.

Након корака индукције у корак мапирања улази 78 часописа са кластерима представљеним носећим фразама и фразама подршке и 70 часописа са кластерима представљеним свим фразама. Резултат корака мапирања приказан је у табели 6.5.

Табела 6.5: Исходи корака мапирања

представљање кластера	исправно	преписивање	понављање	нема улаза
носеће фразе и фразе подршке	74	4	0	2
све фразе	66	3	1	10

У овом кораку нема проблема пресецања одговора, јер су улазни подаци знатно мањег обима, али се јављају понављање и преписивање. Преписивање се препознаје када модел, уместо да изабере шифре таксономије које одговарају теми часописа, готово дословно препише целу таксономију. Тиме не поштује ограничење да не сме навести више шифара него што је области пронађено у часопису, па се добија синтаксно исправан, али садржајно безвредан одговор.

Исправке у кораку мапирања, односно поновно покретање модела са другачијим упитом над часописима који нису прошли из првог покушаја, нису рађене. Разлог је дубља анализа одговора, из које се види да модел уме да врати шифру суседну тачној области таксономије. Такви су примери пресликавања области *Education* у шифру 0504 *Sociology* и области *Economics* у шифру 0501 *Psychology*.

Области за које модел није знао где да их сврста претежно су додељиване шифрама 0401 и 0104. Тако је часопис „*Journal of Conservative Dentistry*“, који је чист стоматолошки пример, препознат као пољопривредни часопис у оба облика представљања кластера, док област ветерине није додељена ниједном часопису иако у узорку постоји часопис из те области. У неким случајевима иста област, попут области *Digital Humanities*, сврстана је под више различитих суседних шифара.

Пошто резултати корака мапирања нису добро описали часописе, даља анализа изведена је над резултатима корака индукције. Разлог је и то што класификација *Science-Metrix* дели области финије од другог нивоа изабране таксономије, будући да има 174 поткатегорије наспрам 40 категорија тог нивоа. Уз то, *Scimago* има категорије и поткатегорије које нису увек на истом нивоу општости, па се не могу непосредно упоређивати са једним нивоом друге таксономије.

У оквиру табеле 6.6 приказано је за колико се часописа, од оних са ваљаним резултатом корака индукције, тај резултат уопште може проверити помоћу двеју референтних класификација. Први ред даје вредности за цео репрезентативни узорак, а наредна два за часописе који имају излаз корака индукције. Разлика између првог и осталих редова показује колико се часописа губи већ у самом кораку индукције.

Табела 6.6: Број часописа са расположивом референтном класификацијом

скуп часописа	укупно	са <i>Science-Metrix</i>	са <i>Scimago</i>	са обе
репрезентативни узорак	80	74	78	73
носеће фразе и фразе по- дршке	78	72	76	71
све фразе	70	64	69	64

Пошто излаз корака индукције, поред области, даје и оцену, поставља се питање да ли је та оцена добро мерило за одређивање примарне теме часописа. При одговору главни ослонац је класификација *Science-Metrix*, јер у њој часопис има само једну додељену категорију. Тешкоћа је у томе што су категорије произвољан текст, па дословно поређење двају низова даје само доњу границу тачности резултата.

Дословним поређењем примарне области поклапање постоји код 11 часописа од 72 при представљању носећим фразама и фразама подршке, односно код 6 часописа при представљању свим фразама. Осим те синтаксне провере, урађена је и семантичка. Пошто класификација *Science-Metrix* бира једну од 174 поткатегорије, пронађеној првој области додељује се њој најближа поткатегорија по косинусној сличности, која се потом пореди са категоријом додељеном часопису.

Овим се не мери исправност поступка него његово слагање са класификацијом *Science-Metrix*. Ако поступак да тачнији опис теме него сама референца, резултат се ипак бележи као промашај. Пример је часопис „*Medical Informatics*“, чији је најближи вектор међу поткатегоријама управо поткатегорија истог назива, док референца за тај часопис наводи „*Nuclear Medicine & Medical Imaging*“, па се резултат рачуна као промашај.

У обзир улази и то да је мера такмичарска, односно да узима највеће поклапање. Тако је за часопис „*Agronomy for Sustainable Development*“ пронађеној области „*Agriculture*“ ближа поткатегорија „*Agricultural Economics & Policy*“ него поткатегорија која заиста јесте област часописа, „*Agronomy & Agriculture*“, и то за свега 0,007, па се и то рачуна као промашај. И семантичка провера тако даје доњу границу поклапања примарне теме, али са нешто бољим резултатом: 33 од 72 при представљању носећим фразама и фразама подршке и 25 од 64 при представљању свим фразама.

У оквиру табеле 6.7 приказан је број часописа чија се примарна категорија није поклопила са категоријом из класификације *Science-Metrix*, као и однос колико од њих своју примарну категорију ипак има међу темама класификације *Scimago*. И сама референтна класификација има грешке, што се види на примеру часописа „*Journal of Electronic Science and Technology*“, чија је категорија означена као „*Optics*“, иако се из описа часописа може закључити да су примарна тема електронске науке и технологије. Управо зато је важно утврдити да ли поступак заиста греши или референтна категорија није добра.

Часописи за које примарна тема није поклопљена са класификацијом *Science-Metrix*, њих 45 у унији оба облика представљања, проверени су ручно. За сваки су забележени назив часописа, број *ISSN*, категорија из класификације *Science-Metrix*, тема добијена сваким од два облика представљања кластера, потврда преузета из описа часописа и веб адреса извора те потврде. Тамо где је поклапања за неки облик представљања већ раније било, часопис за тај

Табела 6.7: Промашаји примарне теме и њихов однос према темама класификације *Scimago*

исход	носеће фразе и фразе подршке	све фразе
промашаја укупно	38	39
прва област јесте међу <i>Scimago</i> темама	19	17
прва област није ни у једној <i>Scimago</i> теми	19	22

облик није поново провераван.

Поклапања су добијена ручном претрагом часописа и његовог описа, а колоне са поклапањем носе ознаку да ли је примарна тема потпуно тачна, преопширна, делимично тачна или је у питању грешка. Преопширна тема обухвата стварну област часописа, али је шира од ње, док делимично тачна обухвата само део те области. Резултати те провере сажети су у табели 6.8.

Табела 6.8: Ручна оцена промашених примарних тема

оцена	носеће фразе и фразе подршке	све фразе
потпуно тачно	23	17
преопширно	7	10
делимично тачно	2	6
грешка	6	6
укупно тачних	57	42

Из приложених резултата види се да је тачност знатно порасла, што значи да аутоматска провера поклапања прве теме са класификацијом *Science-Metrix* није довољно прецизна или да сама референца греша при избору теме часописа. Уочава се и то да у оба случаја поступак греша код шест часописа, при чему само три часописа имају грешку у оба облика представљања, док код преосталих грешка настаје само по једном од њих. Ако се тачним сматрају само часописи чија је примарна тема потпуно тачна, укупној тачности доприносе две групе: часописи чија је тема потврђена већ семантичком провером, њих 34 односно 25, и часописи којима је ручна провера дала оцену потпуно тачно, њих 23 односно 17. Тако се добија укупна тачност од 57 од 72, односно 79% при

представљању носећим фразама и фразама подршке, и 42 од 64, односно 66% при представљању свим фразама.

Уочено је да, ако се уместо прве пронађене теме узме скуп од прве две или три, тачност расте на близу 90% при представљању носећим фразама и фразама подршке, односно на око 88% при представљању свим фразама, и то без иједне измене у моделу или у упиту. У првом облику представљања за 7 часописа утврђено је да им се примарна област налази на другом или трећем месту, а у другом чак код 14 часописа. То значи да поступак примарну област најчешће и пронађе, али је не рангира увек на прво место.

Код великог броја часописа између прве и друге теме практично нема разлике у оцени, што значи да их поступак заправо није ни рангирао. Ту меру би требало оптимизовати или је рачунати без учешћа великог језичког модела. Ни највећа забележена разлика није велика и износи 0,30 при представљању носећим фразама и фразама подршке, односно 0,22 при представљању свим фразама.

Резултати по квантилима фрагментације кластера приказани су у табелама 6.9 и 6.10. Квантили су одређени просечним бројем радова по кластеру, тако да први квантил обухвата најфрагментисаније, а четврти најкрупније кластере. У сваком квантилу налази се приближно једнак број часописа, што омогућава да се удели међу квантилима непосредно упоређују.²

Табела 6.9: Резултати по квантилима фрагментације, носеће фразе и фразе подршке

квантил	опсег	часописа	мера	потпуно тачно	преопширно	делимично	грешка
Q1	2,7–5,8	18	7	5	3	0	3
Q2	5,9–8,8	18	8	7	1	0	2
Q3	9,0–11,5	18	8	6	1	2	1
Q4	11,6–33,2	18	11	5	2	0	0
укупно		72	34	23	7	2	6

Иако се због резултата при представљању свим фразама не може извести закључак да грешка монотонно опада са смањењем фрагментације, што може бити и последица малог узорка, уочава се да у четвртом квантилу, где је број радова

²Ознака *мера* означава аутоматско поклапање семантичком мером, док ознака *потпуно тачно* означава случајеве у којима је поклапање ручно потврђено иако их мера није спојила.

Табела 6.10: Резултати по квантилима фрагментације, све фразе

квантил	опсег	часописа	мера	потпуно тачно	преопширно	делимично	грешка
Q1	2,7–5,8	16	4	5	4	2	1
Q2	5,9–8,8	16	6	4	2	0	4
Q3	9,0–11,5	16	8	4	1	2	1
Q4	11,6–33,2	16	7	4	3	2	0
укупно		64	25	17	10	6	6

по кластеру велик, нема ниједне грешке ни у једном ни у другом облику представљања. Види се и то да представљање свим фразама даје више преопширних и делимично тачних одговора, а мање аутоматски ухваћених поклапања, при истом броју правих грешака. Из тога следи да оно не греша чешће, него чешће одговара прешироко и слабије рангира, што иде у прилог тврдњи да носеће фразе и фразе подршке добро апроксимирају кластере.

Иако се преширока тема може сматрати прилогом добре класификације, делимично поклопљена прва област ипак се рачуна као грешка. На примеру часописа „*Veterinary Sciences*“, чија је тема описана као „*Veterinary Cardiology*“, оба облика представљања дају делимичну тачност. Представљање носећим фразама и фразама подршке чак и погађа више, јер обухвата део „*Cardiology*“, али се област ветерине појављује тек међу каснијим темама, што за примарну тему није довољно.

Слично је и на примеру часописа „*Political Science*“, где поступак препознаје да је област часописа политика, али нигде не обухвата и то да тема говори и о психологији. Код делимичног погађања примарне теме узрок би морао бити додатно испитан. Такви случајеви се аутоматском провером не хватају, па су утврђени тек ручним прегледом.

Поред примарних тема, корак индукције издваја и секундарне. Оне се могу упоређивати само са класификацијом *Scimago*, при чему то поређење не мери тачност него сличност, будући да би ручна провера захтевала знатно више времена. Мера сличности рачуна се тако што се свакој пронађеној области додели најближа од 117 категорија класификације *Scimago*, а потом се проверава да ли се та категорија налази у класификацији датог часописа.

У подацима је утврђено да је просечан удео пронађених тема које се налазе и у класификацији *Scimago* између 37 и 40%. Тај удео је готово исти у оба

облика представљања и не расте код часописа код којих је прва област добро одређена, нити опада код оних код којих није, што значи да квалитет прве области не говори о квалитету осталих тема. Скоро сваки часопис има бар једну погођену област из класификације *Scimago*, па се може рећи да поступак показује слагање са том класификацијом.

Представљање свим фразама производи знатно већи број области, али сразмерно томе не погађа више, посматрано према класификацији *Scimago*. То додатно потврђује да већи број издвојених области сам по себи не значи и бољи опис часописа. Тиме се и овај део анализе слаже са закључком изведеним при поређењу примарних тема.

6.4 Анализа тематских промена часописа кроз време

Прикупљени метаподаци о публикацијама у научним часописима из графа научног знања *OpenAIRE* укрштени су са подацима из базе *ProRef*, како би се пронашао часопис о ком постоји довољан број публикација, а за који је током претходних седам година уочена промена области објављивања. База *ProRef* садржи податке о теми часописа по годинама, претежно прикупљене из базе *Web of Science* (*WoS*). Поред података из те базе, за домаће часописе који се у њој не индексирају, министарство Републике Србије надлежно за ресор науке на годишњем нивоу објављује тему сваког часописа за ту годину.

Укрштањем двеју база уочено је да за домаћи часопис „*Genetika (Beograd)*“ постоји мала промена теме у претходних седам година. Тај часопис у скупу прикупљених података из графа *OpenAIRE* има укупно 687 публикација из периода од 2020. до 2025. године. Након покренутог поступка уклањања дупликата број публикација смањено се на 468, а њихова расподела по годинама приказана је у табели 6.11.

Из табеле се уочава да за 2024. и за 2025. годину постоји знатно мањи број публикација него за остале године, што може нарушити поузданост резултата. Управо те две године носе највећи ризик од погрешног закључка, будући да се тема часописа изводи из мањег узорка. То треба имати у виду при тумачењу свих даљих резултата.

Након фазе уклањања дупликата, за сваку годину посебно израчунате су векторске репрезентације докумената и покренут је поступак извођења репре-

Табела 6.11: Број публикација часописа „*Genetika (Beograd)*“, са бројем ISSN 0534-0012, по годинама

година	број публикација
2020	99
2021	118
2022	109
2023	78
2024	41
2025	31
2026	0

зентативних кластера из скупа K^* у оба облика представљања, носећим фразама и фразама подршке и свим фразама. Потом је покренут само корак индукције. Корак мапирања није извршен јер је за откривање малих промена у часопису значајнија финија подела области, коју управо корак индукције и даје, а и зато што корак мапирања није дао завидне резултате на репрезентативном узорку часописа.

Корак индукције открио је да се проблем понављања јавио у 2021. години, што је отклоњено покретањем сведеног упита над том годином. Корак је покренут засебно за оба облика представљања кластера, а прве пронађене области по годинама, за оба облика, дате су у табели 6.12.

Табела 6.12: Прве пронађене области по годинама, за оба облика представљања кластера

година	носеће фразе и фразе подршке	све фразе
2020	<i>Biology</i>	<i>Plant Breeding and Genetics</i>
2021	<i>Biology</i>	<i>Biology</i>
2022	<i>Agricultural Sciences</i>	<i>Plant Breeding and Genetics</i>
2023	<i>Agriculture</i>	<i>Agriculture</i>
2024	<i>Environmental Science</i>	<i>Plant Sciences</i>
2025	<i>Agriculture and Genetics</i>	<i>Agriculture and Plant Sciences</i>

Да би се промена теме измерила, а не само описала, потребна је мера која пореди два скупа области. Жакаров индекс јесте мера сличности двају скупова и рачуна се као количник броја елемената њиховог пресека и броја елемената

њихове уније:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (6.1)$$

где $|A \cap B|$ означава број елемената који се налазе и у скупу A и у скупу B , а $|A \cup B|$ укупан број различитих елемената оба скупа заједно. Резултат је увек број из интервала $[0, 1]$, при чему јединица означава да су скупови истоветни, а нула да немају ниједан заједнички елемент. Пошто је овде од значаја колико се тема променила, а не колико је остала иста, у наставку се користи Жакарово растојање:

$$d(A, B) = 1 - J(A, B) \quad (6.2)$$

Оно узима вредност нула када су скупови истоветни, а јединицу када немају ниједан заједнички елемент. Све вредности приказане у наставку јесу управо то растојање, па веће вредности означавају већу промену. Тиме се промена теме исказује једним бројем, што омогућава да се године међусобно упоређују.

Да би се мера уопште могла применити, називе области је најпре требало свести на исти облик. Пошто је број области мали, то је урађено простим ручним пресликавањем области на канонски облик. Резултат је дат у табели 6.13, у којој се поред скупа области за сваки облик представљања налази и растојање између текуће и претходне године.

Табела 6.13: Канонски скупови области по годинама и Жакарово растојање (d) у односу на претходну годину

година	носеће фразе и фразе подршке	d	све фразе	d	грешка методе
2020	<i>Biology, Agriculture</i>	—	<i>Genetics, Agriculture</i>	—	0,67
2021	<i>Biology, Genetics, Agriculture</i>	0,33	<i>Biology, Genetics, Agriculture</i>	0,33	0,00
2022	<i>Biology, Genetics, Agriculture</i>	0,00	<i>Plant Pathology, Genetics, Agriculture</i>	0,50	0,50
2023	<i>Genetics, Agriculture</i>	0,33	<i>Bioinformatics, Genetics, Agriculture</i>	0,50	0,33
2024	<i>Genetics, Agriculture, Environmental Science</i>	0,33	<i>Genetics, Agriculture</i>	0,33	0,33
2025	<i>Genetics, Agriculture, Environmental Science</i>	0,00	<i>Genetics, Agriculture</i>	0,00	0,33

Колона грешка методе означава растојање између скупова добијених двама облицима представљања за исту годину. Та вредност не потиче само од грешке корака индукције него и од ручног поступка свођења на канонски облик, који зависи од редоследа обраде. Због тога је треба тумачити као горњу границу неслагања двају облика, а не као тачну меру грешке.

Посматрањем резултата уочава се да је тема часописа добро одређена, али да се не могу утврдити значајни скокови у промени теме. Поређењем са резултатима из базе *ProRef* види се да ни тамо тема не осцилује много: од 2020. до 2022. године наводе се агрономија, генетика и наслеђивање, за 2023. годину нема података, док су за 2024. и 2025. годину наведене биологија, биотехнологија и пољопривреда.

Занимљиво је запажање да се управо у публикацијама из 2023. године јавља област биоинформатике, која би према класификацији из базе *ProRef* требало да буде задржана и у 2024. и у 2025. години. Ово одступање потенцијално се може оправдати тиме што часопис за те две године нема довољно података, односно има скоро три пута мање публикација него у претходним годинама. Та тврдња се, међутим, не може потврдити, јер ниједан други часопис код ког је уочена и најмања промена теме није доступан у скупу података из графа *OpenAIRE* у већем обиму публикација од часописа „*Genetika (Beograd)*“.

Занимљиво је и то што представљање носећим фразама и фразама подршке није успело да пронађе биоинформатику као област часописа за 2023. годину. И то се може оправдати тиме што тај облик врши велика одсецања при сажимању кластера, па је, при овако малој количини података, могуће да је одсецањем управо та област превиђена.

Представљање носећим фразама и фразама подршке проналази, међутим, једну област коју представљање свим фразама не види, а то је животна средина. То може бити шум у подацима, али ако се боље погледа образложење које је модел дао, уочава се да се заправо говори о генетици и о начину на који животна средина утиче на генетику и на принос усева. Уз то, тема животне средине нигде се не јавља међу главним темама а да није везана за претходно поменуте две области, а у 2025. години њена оцена је осетно нижа.

Из приложеног се закључује да овај поступак треба покренути на још неколико часописа са већим корпусом публикација по години. Тек би се тако утврдило да ли представљање свим фразама заиста може да ухвати финију гранулацију у којој би се разлика скупова препознала и да ли представљање

ГЛАВА 6. ПРИКАЗ И ДИСКУСИЈА ДОБИЈЕНИХ РЕЗУЛТАТА

носећим фразама и фразама подршке одсеца управо ту финију поделу. Овај часопис је, наиме, био једини у прикупљеним подацима код ког је уочена промена теме, а истовремено има довољан број публикација по години.

Глава 7

Закључак

У овом раду реализован је поступак у више етапа који научну област часописа одређује аутоматски, на основу метаподатака публикација објављених у њему. Предложени поступак развијан је и проверен над подацима прикупљеним из графа инфраструктуре *OpenAIRE*, који су припремљени алатом *Apache Spark*, уз чишћење наслова и описа и уклањање дупликата Левенштајновим растојањем, а сам поступак изведен је над репрезентативним узорком од осамдесет часописа из свих шест области првог нивоа таксономије коју користи *OpenAIRE*. Приступ одозго-наниже, моделом *BERTopic*, напуштен је зато што једна публикација по правилу припада већем броју области, па је уведен приступ одоздо-навише алатом *KeyBERT*, којим се издваја највише десет кључних фраза по публикацији, а саме фразе се потом кластерују методама *UMAP* и *HDBSCAN*.

Предложено је да се задржавају само кластери који покривају седамдесет одсто публикација часописа, а да се сваки кластер представи у два облика, свим својим фразама и сажето, носећим фразама и фразама подршке. Оба облика се потом мапирају на изабрану таксономију на два начина, косинусном сличношћу векторских репрезентација и великим језичким моделом у два корака. Сажети опис задржава око тридесет осам одсто текста, а са целим кластером дели прву придружену област у 84,3% случајева. Примарна тема часописа тачно је одређена у 79% случајева, односно у близу деведесет одсто ако се уместо прве узму прве три пронађене теме, док је анализа промена кроз време изведена на часопису „*Genetika (Beograd)*“, уз Жакарово растојање као меру.

Основна предност поступка јесте то што не тражи ништа осим метаподатака самих публикација: ни класификацију часописа, ни његов опис, ни податак да

је индексан у међународној бази. Приступ одоздо-навие при том допушта да публикација припада већем броју тема, што одговара стварном стању, а сажети опис кластера смањује улаз на око трећину уз мали губитак, чиме се у ограничење броја токена уклапа већи број кластера. Оба приступа мапирању су детерминистичка — приступ са векторским репрезентацијама по својој природи, а приступ са великим језичким моделом зато што је покренут похлепним декодирањем, па поновљен позив над истим улазом даје исти одговор.

Значајно је и то што су резултати добијени малим језичким моделом од три милијарде параметара и уз одсецање улаза код часописа са највише кластера, дакле уз скромне рачунарске захтеве. Поступак уз то теме уређује по доприносу, док их класификација *Scimago* наводи лексикографски, па се из ње и не види која је тема доминантна. Ручна провера показала је да је у 23 од 38 случајева које је аутоматско поређење прогласило промашајем пронађена тема ипак била потпуно тачна, а међу њима су и часописи код којих сама референтна класификација греша или је грубља од добијеног резултата.

Посебна предност тиче се понављања поступка кроз време. Ако се уз сваку публикацију сачува и највише десет фраза које су је описале, а које припадају репрезентативним кластерима, реч је о врло малом додатном обиму података, којим се одређивање теме за сваку наредну годину знатно убрзава, јер не захтева ни поновно рачунање векторских репрезентација ни поновно издвајање кључних фраза. Исти запис отвара и могућност да се класификују саме публикације, а не само часописи.

Установљено је да главно ограничење није у поступку него у циљној таксономији: њен други ниво није довољно раздвојен, поједине области у њему не постоје, а пошто је реч о затвореној листи од четрдесет категорија, област које у таксономији нема не може ни бити препозната. Похлепно декодирање, које обезбеђује поновљивост, уједно је узрок понављања фраза и пресецања одговора, корак мапирања дао је осетно слабије резултате од корака индукције, а оцена коју модел придружује темама често не раздваја прву од друге. Ограничен дomet има и сама евалуација: узорак од осамдесет часописа премали је да би разлике од неколико процената међу варијантама биле значајне, а поређење се изводи према класификацијама које и саме греше, па мери слагање, а не тачност у строгом смислу. Анализа тематских промена кроз време изведена је на једном часопису, и то на оном који у последње две године има знатно мањи број публикација.

Најнепосреднији правац унапређења води кроз саму таксономију — изоштравање описа и допуну недостајућих области — уз то да број фраза у сажетом опису буде сразмеран величини кластера, а рангирање тема рачунато из покривености кластера, без учешћа језичког модела. Већи инструкциони модел и ограничено декодирање уклонили би понављање и пресецање одговора. Евалуацију би требало проширити на већи број часописа и измерити колико се саме референтне класификације међусобно не слажу.

Оно што је данас мана поступка може постати употребљив податак. Кластери који се ни са једном облашћу не поклапају довољно добро могу се тумачити двојачко: као сигнал да у таксономији недостаје област, чиме поступак постаје механизам за предлагање њене допуне, или као сигнал изражене мултидисциплинарности часописа, што ниједна постојећа класификација не даје. Предност имплементираног решења јесте и то што за сваку област може издвојити носеће фразе, које се могу употребити за употпуњавање саме таксономије. Могуће је даље испитати и да ли се груписањем свих носећих фраза једне области може извести наредни ниво таксономије или открити област коју треба засебно издвојити на истом нивоу. Проширење анализе кроз време на већи број часописа отворило би праћење померања фокуса часописа и уочавање научних праваца у настајању.

Практична примена решења непосредно следи из проблема од ког је пошло. Оно попуњава простор за часописе у графу инфраструктуре *OpenAIRE*, а може послужити и дигиталним библиотекама, библиографским системима и националним репозиторијумима, којима недостаје уједначен начин сврставања часописа у научне области. Агрегаторима олакшава мониторинг и евалуацију, а истраживачу који бира где ће објавити рад даје податак који је до сада морао да прикупља ручно.

Литература

- [1] Manghi P., Atzori C., Bardi A., Baglioni M., Schirrwagen J. и др., „OpenAIRE Research Graph Dataset“, Zenodo, 2022. doi:10.5281/zenodo.3516917
- [2] OECD, *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*, OECD Publishing, Paris, 2015. doi:10.1787/9789264239012-en
- [3] Archambault É., Beauchesne O. H., Caruso J., „Towards a Multilingual, Comprehensive and Open Scientific Journal Ontology“, *Зборник радова 13. међународне конференције ISSI*, Durban, стр. 66–77, 2011.
- [4] *Science-Metrix Journal Classification*, верзија 2.00, Science-Metrix/Elsevier, Montreal, 2022, лиценца Creative Commons Attribution 3.0, доступно на: <https://science-metrix.com/classification/> (приступљено 30.08.2026.)
- [5] SCImago, *SJR — SCImago Journal & Country Rank*, доступно на: <https://www.scimagojr.com> (приступљено 30.08.2026.)
- [6] Elsevier, *Scopus Content Coverage Guide*, Elsevier B.V., 2023, доступно на: <https://www.elsevier.com/products/scopus/content> (приступљено 30.08.2026.)
- [7] Clarivate, *Journal Citation Reports*, Clarivate Analytics, доступно на: <https://clarivate.com/academia-government/scientific-and-academic-research/research-funding-analytics/journal-citation-reports/> (приступљено 30.08.2026.)
- [8] Рачунарски центар Универзитета у Београду, *ProRef — Провера референци*, доступно на: <https://proref.rcub.bg.ac.rs/> (приступљено 30.08.2026.)

- [9] Kotitsas S., Pappas D., Manola N., Papageorgiou H., „SciNoBo: a novel system classifying scholarly communication in a dynamically constructed hierarchical Field-of-Science taxonomy“, *Frontiers in Research Metrics and Analytics*, 8, чланак 1149834, 2023. doi:10.3389/frma.2023.1149834
- [10] *SciNoBo Field-of-Science taxonomy*, доступно на: <https://github.com/iNoBo/scinobo-fos-classification> (приступљено 30.08.2026.)
- [11] Hey T., Tansley S., Tolle K. (ур.), *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, Redmond, 2009.
- [12] García S., Ramírez-Gallego S., Luengo J., Benítez J. M., Herrera F., „Big data preprocessing: methods and prospects“, *Big Data Analytics*, 1(9), 2016. doi:10.1186/s41044-016-0014-0
- [13] Vassiliadis P., „A Survey of Extract-Transform-Load Technology“, *International Journal of Data Warehousing and Mining*, 5(3), стр. 1–27, 2009.
- [14] Whistler K., *Unicode Standard Annex #15: Unicode Normalization Forms*, The Unicode Consortium, 2023, доступно на: <https://www.unicode.org/reports/tr15/> (приступљено 30.08.2026.)
- [15] Elmagarmid A. K., Ipeirotis P. G., Verykios V. S., „Duplicate Record Detection: A Survey“, *IEEE Transactions on Knowledge and Data Engineering*, 19(1), стр. 1–16, 2007.
- [16] Levenshtein V. I., „Binary codes capable of correcting deletions, insertions and reversals“, *Soviet Physics Doklady*, 10(8), стр. 707–710, 1966.
- [17] Wagner R. A., Fischer M. J., „The String-to-String Correction Problem“, *Journal of the ACM*, 21(1), стр. 168–173, 1974.
- [18] Apache Software Foundation, *Apache Parquet*, доступно на: <https://parquet.apache.org> (приступљено 30.08.2026.)
- [19] Zaharia M., Xin R. S., Wendell P., Das T., Armbrust M. и др., „Apache Spark: A Unified Engine for Big Data Processing“, *Communications of the ACM*, 59(11), стр. 56–65, 2016.
- [20] Manning C. D., Raghavan P., Schütze H., *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, 2008.

- [21] Porter M. F., „An algorithm for suffix stripping“, *Program*, 14(3), стр. 130–137, 1980.
- [22] Spärck Jones K., „A statistical interpretation of term specificity and its application in retrieval“, *Journal of Documentation*, 28(1), стр. 11–21, 1972.
- [23] Harris Z. S., „Distributional Structure“, *Word*, 10(2–3), стр. 146–162, 1954.
- [24] Mikolov T., Chen K., Corrado G., Dean J., „Efficient Estimation of Word Representations in Vector Space“, *Зборник радова ICLR 2013 Workshop*, arXiv:1301.3781, 2013.
- [25] Devlin J., Chang M.-W., Lee K., Toutanova K., „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“, *Зборник радова NAACL-HLT 2019*, стр. 4171–4186, 2019.
- [26] Reimers N., Gurevych I., „Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks“, *Зборник радова EMNLP-IJCNLP 2019*, стр. 3982–3992, 2019.
- [27] Beltagy I., Lo K., Cohan A., „SciBERT: A Pretrained Language Model for Scientific Text“, *Зборник радова EMNLP-IJCNLP 2019*, стр. 3615–3620, 2019.
- [28] Cohan A., Feldman S., Beltagy I., Downey D., Weld D. S., „SPECTER: Document-level Representation Learning using Citation-informed Transformers“, *Зборник радова ACL 2020*, стр. 2270–2282, 2020.
- [29] Ostendorff M., Rethmeier N., Augenstein I., Gipp B., Rehm G., „Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings“, *Зборник радова EMNLP 2022*, стр. 11670–11688, 2022.
- [30] Beyer K., Goldstein J., Ramakrishnan R., Shaft U., „When Is Nearest Neighbor Meaningful?“, *Зборник радова ICDT 1999*, LNCS 1540, Springer, стр. 217–235, 1999.
- [31] McInnes L., Healy J., Melville J., „UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction“, arXiv:1802.03426, 2018.
- [32] Bishop C. M., *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

- [33] Sutton R. S., Barto A. G., *Reinforcement Learning: An Introduction*, 2. издање, MIT Press, Cambridge, 2018.
- [34] Campello R. J. G. B., Moulavi D., Sander J., „Density-Based Clustering Based on Hierarchical Density Estimates“, *Зборник радова PAKDD 2013*, LNCS 7819, Springer, стр. 160–172, 2013.
- [35] Ester M., Kriegel H.-P., Sander J., Xu X., „A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise“, *Зборник радова KDD-96*, AAAI Press, стр. 226–231, 1996.
- [36] Blei D. M., „Probabilistic Topic Models“, *Communications of the ACM*, 55(4), стр. 77–84, 2012.
- [37] Grootendorst M., „BERTopic: Neural topic modeling with a class-based TF-IDF procedure“, arXiv:2203.05794, 2022.
- [38] Grootendorst M., *KeyBERT: Minimal keyword extraction with BERT*, Zenodo, 2020. doi:10.5281/zenodo.4461265
- [39] Carbonell J., Goldstein J., „The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries“, *Зборник радова SIGIR '98*, ACM, стр. 335–336, 1998.
- [40] Zhao W. X., Zhou K., Li J., Tang T., Wang X. и др., „A Survey of Large Language Models“, arXiv:2303.18223, 2023.
- [41] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L. и др., „Attention Is All You Need“, *Зборник радова NIPS 2017*, стр. 5998–6008, 2017.
- [42] Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C. L. и др., „Training language models to follow instructions with human feedback“, *Зборник радова NeurIPS 2022*, стр. 27730–27744, 2022.
- [43] Micikevicius P., Narang S., Alben J., Diamos G., Elsen E. и др., „Mixed Precision Training“, *Зборник радова ICLR 2018*, arXiv:1710.03740, 2018.
- [44] Ji Z., Lee N., Frieske R., Yu T., Su D. и др., „Survey of Hallucination in Natural Language Generation“, *ACM Computing Surveys*, 55(12), чланак 248, 2023. doi:10.1145/3571730

- [45] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B. и др., „Scikit-learn: Machine Learning in Python“, *Journal of Machine Learning Research*, 12, стр. 2825–2830, 2011.
- [46] Honnibal M., Montani I., Van Landeghem S., Boyd A., *spaCy: Industrial-strength Natural Language Processing in Python*, Zenodo, 2020. doi:10.5281/zenodo.1212303
- [47] Bird S., Klein E., Loper E., *Natural Language Processing with Python*, O’Reilly Media, Sebastopol, 2009.
- [48] Hunter J. D., „Matplotlib: A 2D Graphics Environment“, *Computing in Science & Engineering*, 9(3), стр. 90–95, 2007.
- [49] Assante M., Candela L., Castelli D., Cirillo R., Coro G. и др., „Enacting open science by D4Science“, *Future Generation Computer Systems*, 101, стр. 555–563, 2019.
- [50] Paszke A., Gross S., Massa F., Lerer A., Bradbury J. и др., „PyTorch: An Imperative Style, High-Performance Deep Learning Library“, *Зборник радова NeurIPS 2019*, стр. 8024–8035, 2019.
- [51] Qwen Team, „Qwen2.5 Technical Report“, arXiv:2412.15115, 2024.
- [52] Wolf T., Debut L., Sanh V., Chaumond J., Delangue C. и др., „Transformers: State-of-the-Art Natural Language Processing“, *Зборник радова EMNLP 2020: System Demonstrations*, стр. 38–45, 2020.
- [53] Gugger S., Debut L., Wolf T., Schmid P., Mueller Z. и др., *Accelerate: Training and inference at scale made simple, efficient and adaptable*, 2022, доступно на: <https://github.com/huggingface/accelerate> (приступљено 30.08.2026.)
- [54] Holtzman A., Buys J., Du L., Forbes M., Choi Y., „The Curious Case of Neural Text Degeneration“, *Зборник радова ICLR 2020*, arXiv:1904.09751, 2020.

Листа слика

2.1	Псеудокод за израчунавање Левенштајновог растојања	9
2.2	Дијаграм тока обраде података моделом <i>BERTopic</i>	15
2.3	Дијаграм тока обраде података алатом <i>KeyBERT</i>	16
3.1	Дијаграм тока имплементираних решења	20
3.2	Шема улазних датотека са метаподацима публикација	23
4.1	Теме које је модел <i>BERTopic</i> издвојио за часопис „ <i>Semantic Web</i> “ . .	29
4.2	Кључне речи по темама и њихове вредности мере <i>c-TF-IDF</i>	30
4.3	Топлотна мапа сличности тема на основу векторских репрезентација	31
4.4	Топлотна мапа сличности тема на основу матрице <i>c-TF-IDF</i>	32
4.5	Десет највећих репрезентативних кластера часописа „ <i>Semantic Web</i> “ са најфреквентнијим фразама и бројем покривених докумената . . .	36
4.6	Кластери онтолошке породице часописа „ <i>Semantic Web</i> “ са најфре- квентнијим фразама	38
4.7	Хијерархија кластера онтолошке породице часописа „ <i>Semantic Web</i> “ .	39
4.8	Псеудокод за избор фраза подршке кластера	43
6.1	Удео задржаних фраза у зависности од величине кластера	55
6.2	Однос укупног броја фраза и броја фраза у сажетом опису кластера	56
6.3	Поклапање прве придружене области у зависности од величине кла- стера	57
6.4	Однос првих изабраних области у два облика представљања кластера	58
6.5	Области које најчешће изазивају замену прве и друге области у два облика представљања кластера	59

Листа табела

4.1	Примери уједначавања облика издвојених фраза	34
4.2	Тумачење кластера онтолошке породице часописа „ <i>Semantic Web</i> “ . .	37
6.1	Расподела репрезентативног узорка часописа по областима првог нивоа таксономије	53
6.2	Резултати приступа заснованог на сличности векторских репрезентација	54
6.3	Врсте грешке у зависности од величине кластера, у процентима . . .	59
6.4	Исходи корака индукције по поступку опоравка	60
6.5	Исходи корака мапирања	61
6.6	Број часописа са расположивом референтном класификацијом	62
6.7	Промашаји примарне теме и њихов однос према темама класификације <i>Scimago</i>	64
6.8	Ручна оцена промашених примарних тема	64
6.9	Резултати по квантилима фрагментације, носеће фразе и фразе подршке	65
6.10	Резултати по квантилима фрагментације, све фразе	66
6.11	Број публикација часописа „ <i>Genetika (Beograd)</i> “, са бројем <i>ISSN</i> 0534-0012, по годинама	68
6.12	Прве пронађене области по годинама, за оба облика представљања кластера	68
6.13	Канонски скупови области по годинама и Жакарово растојање (d) у односу на претходну годину	69

Списак скраћеница

Скраћенице су наведене абецедним редом. Свака се при првом појављивању у тексту објашњава и у самом тексту.

<i>APC</i>	<i>Article Processing Charge</i>
<i>API</i>	<i>Application Programming Interface</i>
<i>ASJC</i>	<i>All Science Journal Classification</i>
<i>BERT</i>	<i>Bidirectional Encoder Representations from Transformers</i>
<i>CNR</i>	<i>Consiglio Nazionale delle Ricerche</i>
<i>CSV</i>	<i>Comma-Separated Values</i>
<i>c-TF-IDF</i>	<i>class-based Term Frequency — Inverse Document Frequency</i>
<i>CUDA</i>	<i>Compute Unified Device Architecture</i>
<i>DBSCAN</i>	<i>Density-Based Spatial Clustering of Applications with Noise</i>
<i>DOI</i>	<i>Digital Object Identifier</i>
<i>EISSN</i>	<i>Electronic International Standard Serial Number</i>
<i>ETL</i>	<i>Extract, Transform, Load</i>
<i>HBM2</i>	<i>High Bandwidth Memory 2</i>
<i>HDBSCAN</i>	<i>Hierarchical Density-Based Spatial Clustering of Applications with Noise</i>
<i>HTML</i>	<i>HyperText Markup Language</i>
<i>ISSN</i>	<i>International Standard Serial Number</i>
<i>ISTI</i>	<i>Istituto di Scienza e Tecnologie dell'Informazione</i>
<i>JCR</i>	<i>Journal Citation Reports</i>
<i>JSON</i>	<i>JavaScript Object Notation</i>
<i>LLM</i>	<i>Large Language Model</i>
<i>LTS</i>	<i>Long-Term Support</i>
<i>OECD</i>	<i>Organisation for Economic Co-operation and Development</i>

СПИСАК СКРАЋЕНИЦА

<i>OWL</i>	<i>Web Ontology Language</i>
<i>RCUB</i>	Рачунарски центар Универзитета у Београду
<i>SJR</i>	<i>SCImago Journal Rank</i>
<i>TF-IDF</i>	<i>Term Frequency — Inverse Document Frequency</i>
<i>TSV</i>	<i>Tab-Separated Values</i>
<i>UMAP</i>	<i>Uniform Manifold Approximation and Projection</i>
<i>WoS</i>	<i>Web of Science</i>
<i>XML</i>	<i>Extensible Markup Language</i>
