Класификација докумената из медицинског домена екстраховањем таксономских концепата из MeSH онтологије

УДК 004.82:025.43MESH

САЖЕТАК: Рад је настао као одговор на задатак класификације медицинских докумената, постављен током летње школе *Keyword Search in Big Linked Data*, одржане у оквиру COST акције Keystone 2017. године на Технолошком универзитету у Бечу. У њему се приказују резултати специфичног приступа класификацији заснованог на креирању минималних сурогата текста. Као основа класификације узета је MeSH онтологија, заснована на тезаурусу *Medical Subject Headings*. У текстовима, претходно класификованим помоћу таксономије ове онтологије, најпре се проналазе појмови од важности, а потом се замењују таксономским референцама. Тако екстраховане референце користе за класификацију унутар MeSH таксономије помоћу простог алгоритма, а резултати се евалуирају у односу на ручно класификоване документе.

КЉУЧНЕ РЕЧИ: класификација докумената, MeSH, онтологије, екстракција информација

РАД ПРИМЉЕН: 21. април 2019.
РАД ПРИХВАЋЕН: 30. август 2019.

Михаило Шкорић
mihailo.skoric@rgf.bg.ac.rs
Универзитет у Београду
Београд, Србија

Мауро Драгони
dragoni@fbk.eu
Фондација „Бруно Кеслер“
Тренто, Италија

1. Увод

1.1 О задатку

Овај рад описује решење проблема који је задао предавач на једнодневном хакатону летње школе Keystone 3rd training school:

Keyword Search in Big Linked Data,[1] одржаној у Бечу 21–25. августа 2017. у оквиру *COST* акције IC1302 - *Keyword Search in Big Linked Data*. Проблем се састоји у класификовању 10.000 докумената из дигиталне колекције Националне медицинске библиотеке Сједињених Америчких Држава (слика 1), које је обезбедио предавач, док је на учесницима било да те документе класификују у класе предефинисане у MeSH (*Medical Subject Headings*) онтологији (Dragoni, 2017).

> *... Goserelin in the adjuvant treatment of breast cancer An update of the Zoladex Early Breast Cancer Research Association (ZEBRA) trial was presented by Professor R Blamey (Nottingham City Hospital, UK). Goserelin was found to be better ... Results were presented by the Austrian Breast and Colorectal Cancer Study Group comparing ...*

Слика 1. Одломак једног од докумената који треба класификовати

Класификација у овом раду направљена је, сходно правилима, на основу таксономије медицинских појмова из верзије MeSH онтологије из 2016. године.[2] Онтологија се може претраживати на вебу[3], где се могу наћи предефинисани упити међу којима су они за класе и предикате, или где се новим SPARQL упитима могу добити други жељени подаци.

Документација, RDF тројке и преузимање примера су такође доступни је на вебу,[4] док за прегледање предиката постоји и шира опција[5] која нуди табеларни приказ предиката, њихов опис и XML етикету која је посебно важна ако се користи локална копија MeSH онтологије у виду XML серијализације. Онтологија се састоји од 56.309 медицинских концепата, описаних и систематски сврстаних у хијерархијско дрво (слика 2).

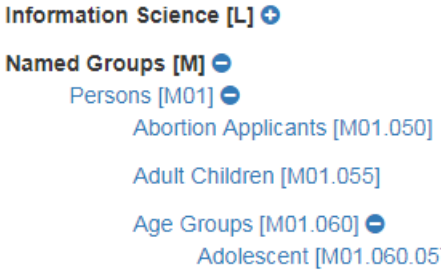
[1] Keyword Search in Big Linked Data (на вебу)

[2] Класификација докумената из медицинског домена заснована на онтологијама је предмет истраживања бројних тимова, при чему се користе различити приступи, али као онтологија се најчешће користи управо MeSH.

[3] Приступна тачка за претраживање је на вебу

[4] Документација, RDF тројке и преузимање примера на вебу

[5] Прегледање предиката на вебу



Слика 2. Приказ исечка хијерархијског дрвета.

Концепти из онтологије су хијерархијски поређани, а сваки од њих има додељен и идентификатор који се састоји од блокова цифара раздвојених тачкама, а који описују надређене концепте у опадајућем редоследу, од највишег до најнижег у хијерархији. Класе за класификацију се, конкретно у овом задатку, односе на други ниво хијерархије дрвета – има их укупно 1.718 – а препознају се тако што се њихови идентификатори састоје два блока, на пример [M01.055] *Adult children* (слика 2), где први блок *M01* означава да му је надређен чвор [M01] *Persons*, а други блок *055* је јединствен међу сабратним чворовима.

1.2 О класификацији текста

Проблем класификације докумената, у најопштијем смислу, јавља се у две варијанте: класификација у непознатом, и класификација у познатом, ограниченом домену класа. За обе варијанте проблем се своди на одређивање сличности два документа, при чему је уобичајено коришћење такозваног Дајсовог[6] индекса, односно коефицијента сличности.

$$sim(D_i, D_j) = \frac{2|S_i \bigcap S_j|}{|S_i| + |S_J|} \tag{1}$$

Он нам каже да ако је S_i скуп термина додељених документу D_i, а S_j скуп термина додељених документу D_j, онда се овај индекс може дефинисати као двоструки број заједничких термина према укупном броју термина у оба документа (ако је S скуп, онда је $|S|$ број елемената

[6] Lee Raymond Dice – амерички биолог (1887–1977)

скупа). Ако документа немају заједничких термина у том случају је $sim(D_i, D_j) = 0$ и одражава минималну сличност двају докумената, а уколико два документа имају додељене потпуно исте скупове термина онда је $sim(D_i, D_j) = 1$ и одражава максималну сличност. Када је домен класа унапред познат и ограничен, што је наш случај, проблем се своди на проналажење одговарајуће класе са највећом sim(документ, класа докумената) вредношћу, за сваки документ који је предмет класификације.[7]

Проблем који се јавља приликом израчунавања коефицијента сличности између текстова је висока цена, која се мора платити било процесорском моћи или временом извршења. Управо због тога први корак у класификацији најчешће представља креирање сурогата текста, где се он преводи у векторски простор који чине вектори речи са фреквенцијама појављивања. Речи у индексу се некада добијају стемовањем, некада лематизацијом, некада заменом синонима, или хипонима хиперонимима, како би се сурогати текста додатно смањили и како би се убрзало време извршења израчунавања. За квалитетну класификацију у овом поступку, неопходно је да сурогати буду исправно креирани и да верно представљају текст.

У радовима (Trieschnigg et al., 2009) и (Elberrichi et al., 2012) тестирано је више метода класификације медицинских докумената заснованих на MeSH онтологији или тезаурусу. Направљени класификатори користили су:

- Само MeSH тезаурус (*'Thesaurus-oriented' classifiers*);
- Тренинг скуп за изградњу експлицитних модела за сваки MeSH концепт (*'Concept-oriented' classifiers*);
- Ручно креиране анотације докумената слично као код обичних класификатора текста, да се одреди одговарајући концепт (*'K-Nearest Neighbor' classifier*);
- Хибридне и ручно профињене системе који комбинује више приступа (*'Hybrid' classifiers*).

У оба рада је закључено да *K-Nearest Neighbor* класификатор (KNN) даје најбоље резултате, али да је, упркос својим предностима, много спорији од класификатора заснованих на тезаурусу, као и да се са растом скупа докумената за тестирање његове перформансе додатно смањују, што није било пожељно при решавању добијеног задатка.

[7] Проналажење информација (на вебу)

У овом раду експериментише се једноставним приступом класификацији да би се оценио значај благовременог управљања великом количином података, као и употребна вредност семантике похрањене у MeSH онтологији. Циљ је био направити класификатор који би био брз и једноставан, како би се решио проблем велике количине текста коју треба класификовати. Примењује се драстична сумаризација докумената и самих класа – класе (концепти другог нивоа онтологије) сведене су на један једини термин – њихов назив. Са друге стране, документи су сведени само на појављивања термина (назива концепата из MeSH онтологије) који се уз једноставно мапирање (похрањено у њиховим идентификаторима у самој онтологији) поистовећују са термином који означава неку класу, њима надређени објекат. Ово умногоме олакшава и убрзава процес израчунавања сличности, јер свака класа сада има само један термин. На овај начин документ ће помоћу Дајсовог индекса увек бити класификован у класу чији (једини) термин се највише пута у њему јавља, чиме се израчунавање избегава и своди на проналажење најфреквентнијег термина у сурогату текста.

2. Поставка експеримента

Циљ експеримента је провера могућности и успешности класификације медицинских докумената на основу таксономије из MeSH онтологије и као и на основу система заснованог на правилима – појављивању термина везаних за концепте из MeSH онтологије у документима који треба да се класификују. Ток експеримента се може поделити на пет сукцесивних етапа.

1. **Екстракција таксономије из MeSH онтологије коришћењем SPARQL упита**. Ова етапа је неопходна како би се прикупила листа идентификатора и утврдила таксономска позиција термина употребљених у документима, као и позиција њихових чворова у односу на позиције чворова могућих класа.
2. **Конверзија докумената у векторе идентификатора концепата који се у њима јављају**.Ова етапа омогућава да документима доделимо атрибуте који су непосредно и нераскидиво повезани са класама у које те документе треба сврстати.
3. **Уклањање шумова**. Ова етапа треба да омогући и обезбеди што боље резултате при класификацији докумената.

4. **Класификација докумената у класе на основу њихових вектора идентификатора и једноставних скупова правила**.
5. **Евалуација успешности класификација докумената за сваки од коришћених скупова правила**. Ова етапа нам омогућава да сагледамо и упоредимо различита правила класификације, као и да установимо да ли се неки скупови правила могу сматрати успешним, који су то скупови и колико су успешни.

У наредним поглављима, етапе експеримента биће описане детаљније, како би се стекао бољи увид у коришћене методе и добијене резултате.

2.1 Екстракција таксономије концепата из MeSH онтологије

Екстракција матрице назива концепата и њихових идентификатора позиција у класификационом дрвету обавља се помоћу SPARQL упита. Како у овој онтологији постоје тројке који се састоје од концепта, предиката и објекта тог предиката, а који одражава позицију у таксономији, овај део задатка своди се на екстракцију субјекта и објекта за сваку од ових тројки.

Најпре је потребно одредити име предиката у онтологији који одражава идентификатор позиције у таксономији облика **[A-Z][0-9][0-9]([.][0-9][0-9][0-9])***.[8] Коришћен је једноставан SPARQL упит, који за било који концепт из онтологије излистава све предикате и објекте свих тројки из MeSH 2016 онтологије чији је тај концепт део (слика 3). Упит смо илустровали концептом (*mesh2016:D049916*).

```
PREFIX mesh2016: <http://id.nlm.nih.gov/mesh/2016/>
SELECT DISTINCT ?predikat ?objekat
FROM <http://id.nlm.nih.gov/mesh/2016>
WHERE  mesh2016:D049916 ?predikat ?objekat
ORDER BY ?class
```

Слика 3. SPARQL упит за добављање таксономијие свих концепата из MeSH 2016 онтологије.

[8] Овај регуларни израз описује конструкцију која се састоји од обавезног дела (велико слово, цифра, цифра) и опционог дела (тачка, цифра, цифра, цифра), који може да се понавља.

На основу упита добијен је излаз из кога се, поред осталог, закључује да је тражени предикат *meshv:treeNumber*, јер садржи блокове цифара који описују хијерархију (слика 4). Овај податак се користи у наредном упиту који има циљ да излиста све називе концепата (термине) и њихове *meshv:treeNumber* вредности.

rdf:type; meshv:TopicalDescriptor
rdfs:label; Polyplacophora
meshv:identifier; D049916
meshv:dateEstablished; 2006-01-06
meshv:historyNote; 2006
meshv:publicMeSHNote; 2006
meshv:treeNumber; mesh2016:B01.050.500.644.600

Слика 4. Неки од Излаза SPARQL упита, међу којима су и жељени предикат и објекат

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX meshv: <http://id.nlm.nih.gov/mesh/vocab#>
PREFIX mesh2016: <http://id.nlm.nih.gov/mesh/2016/>
SELECT DISTINCT ?naziv ?treeNumber
FROM <http://id.nlm.nih.gov/mesh/2016>
WHERE ?koncept rdfs:label ?naziv .
?koncept meshv:treeNumber ?treeNumber
ORDER BY DESC(STRLEN(?naziv))
```

Слика 5. SPARQL upit за добављање листе назива и позиција свих концепата из онтологије MeSH 2016.

Други SPARQL упит прибавља називе концепата на основу предиката *rdfs:label*, а затим тражи и *mesh:treeNumber* тог истог концепта. Термини су на излазу поређани по величини назива од најдужег до најкраћег, како би се тим редом претраживали у документима да би се избегло да дужа поклапања не буду препозната због краћих (слика 5).

Резултат овог упита је CSV датотека чији сваки ред садржи назив (*rdfs:label*) и таксономски идентификатор (*treeNumber*) сваког од концепата (слика 6). Ова датотека биће коришћена у наредном кораку, у којем се у документима проналазе називи концепата, тј. термини, након чега се замењују идентификаторима чворова у таксономском дрвету. Треба истаћи да се могу наћи и једночлани термини као и вишечлани (нпр. *Bacteria* и *Gram-Negative Bacteria*), али имајући у виду редослед примене замена (по дужини ниске, од најдуже до најкраће) неће доћи до погрешне замене и препознавања само дела термина.

```
Ganglia;A08.340
Neurons;A08.675
Malleus;A09.246.397.247.524
Cochlea;A09.246.631.246
Eyelids;A09.371.337
Choroid;A09.371.894.223
Tissues;A10
Chorion;A10.615.284.473
Muscles;A10.690
```

Слика 6. Примери линија CSV документа који садржи називе и идентификаторе чворова концепата.

2.2 Конверзија докумената у векторе концепата

Ова етапа састоји се од два корака. Најпре се у свим документима проналазе и замењују претходно излистани термини, а затим се уклања преостали текст како би се документи трансформисали у листу вектора идентификатора. Никаква нормализација ни документа ни термина није рађена, што за енглески језик, који нема богат флективни систем може да буде прихватљиво, али би за флекитвно богат језик, као што је српски, била неопходна претходна лематизација или неки други вид нормализације оба ресурса.

Проналажење концепата из онтологије у тексту. Како је оно што желимо да постигнемо да у документима нешто (термине

који означавају концепте) пронађемо и потом нечим (одговарајућим таксономским идентификаторима) заменимо, при чему те две ствари имамо излистане заједно у претходно генерисаној датотеци, било је могуће листу из датотеке директно трансформисати у C# функцију која би то радила.

Ово је постигнуто заменом карактера **;** у листи са ниском каракатера **", "** затим конкатенацијом (или дописивањем) ниске карактера **doc = doc.Replace("** на почетак сваког новог реда и ниске карактера **");** на крај сваког реда, где се doc односи на варијаблу која представља комплетан садржај документа (слика 7).[9]

```
doc = doc.Replace("Ganglia", "A08.340");
doc = doc.Replace("Neurons", "A08.675");
doc = doc.Replace("Malleus", "A09.246.397.247.524");
doc = doc.Replace("Cochlea", "A09.246.631.246");
doc = doc.Replace("Eyelids", "A09.371.337");
doc = doc.Replace("Choroid", "A09.371.894.223");
doc = doc.Replace("Tissues", "A10");
doc = doc.Replace("Chorion", "A10.615.284.473");
doc = doc.Replace("Muscles", "A10.690");
```

Слика 7. Одломак C# скрипте за проналажење и замену која се заснива на претходно наведеним концептима и идентификаторима (слика 6)

За замену у свим документима које треба класификовати припремљен је други C# код који учитава један по један документ за класификацију и на њима примењује скрипту генерисану у претходном кораку како би концепти били пронађени према именима и замењени идентификатором чвора из онтологије. Ова етапа је најдужа и временски најзахтевнија јер се за наш експеримент подразумева примењивање 56.309 replace израза над 10.000 докумената што даје укупно 563.090.000 трансформација. Даље истраживање може ићи у правцу коришћења коначних аутомата и трансдуктора за решење овог

[9] Накнадним размишљањем закључили смо да би се сличним приступом могао генерисати и другачији вид замене заснован на петљама или регуларним изразима, што би убрзало замене и умањило ефекте вишеструког парсирања истог документа.

проблема, које је комплексније за имплементацију, али брже врши обраде овог типа.

Трансформација докумената у векторе идентификатора (креирање сурогата) Након што су документи успешно обележени идентификаторима концепата који се у њима појављују, било је неопходно документе очистити од осталог, неупареног текста. Како се ово не би радило појединачно за сваки документ, они су спојени у један, величине ∼230MB, са новим редовима као границом између докумената. Информације од важности – називи докумената (**[0-9]+[.]txt**), идентификатори у њима (**[A-Z][0-9][0-9]([.][0-9][0-9][0-9])***) и ознаке за нови ред (**[\r\n]+**) - проналазе се помоћу регуларног израза (**[A-Z][0-9][0-9]([.][0-9][0-9][0-9])***)|(**[0-9]+[.]txt**)|(**[\r\n]+**), док се све остало уклања. Овим се величина датотеке смањила преко 450 пута (нова величина: ∼0.5MB).

По завршетку трансформације добија се датотека у којој сваки нови ред представља нови документ: ред започиње називом документа (без екстензије) иза кога следи тачка зарез, а затим сви идентификатори концепата који су у њему пронађени одвојени зарезима (слика 8).

```
2875592;M01.060.116
2875593;D13.444.308,D13.444.308
2875594;C04.557.465.625.650.510,D13.444.735,D13.444.735
2875595;A01.236,A01.236;A01.236
2875596;D13.444.735
2875598;
2875599;D02.455.612
```

Слика 8. Одломак датотеке која садржи називе докумената и све идентификаторе у њима.

Илустроваћемо на једном примеру трансформацију једног од полазних докумената по корацима. У делу документа са Слике 1 препознати су неки термини (слика 9), који су потом замењени идентификаторима (слика 10). Уочава се да је у примеру *Colorectal* дошло до препознавања дела речи и тако је ниска **Color** погрешно

замењена са **G01.590.540.199**[10] јер се термини *colorectal cancer* и *colorectal* не налазе као појмови у коришћеној верзији онтологије. Слика 11 приказује коначан сурогат тескта са Слике 1.

> ... **Goserelin** in the adjuvant treatment of breast cancer An update of the Zoladex Early **Breast** Cancer **Research Association** (ZEBRA) trial was presented by Professor R Blamey (Nottingham City Hospital, UK). **Goserelin** was found to be better ... Results were presented by the **Austrian Breast** and **Color**ectal Cancer Study Group comparing ...

Слика 9. Одломак оригиналног документа са оболеженим терминима који се налазе у MeSH онтологији.

> ... **D06.472.699.327.740.320.340** in the adjuvant treatment of breast cancer An update of the Zoladex Early **A01.236** Cancer **H01.770.644 F02.463.425.069** (ZEBRA) trial was presented by Professor R Blamey (Nottingham City Hospital, UK). **D06.472.699.327.740.320.340** was found to be better... Results were presented by the **Z01.542.088 A01.236** and **G01.590.540.199**ectal Cancer Study Group comparing ...

Слика 10. Одломак документа у коме су термини из MeSH онтологије замењени својим таксономским идентификаторима.

> ...D06.472.699.327.740.320.340;A01.236;H01.770.644;F02.463.425.069;D06.472.699.327.740.320.340;A01.236;G01.590.540.199;...

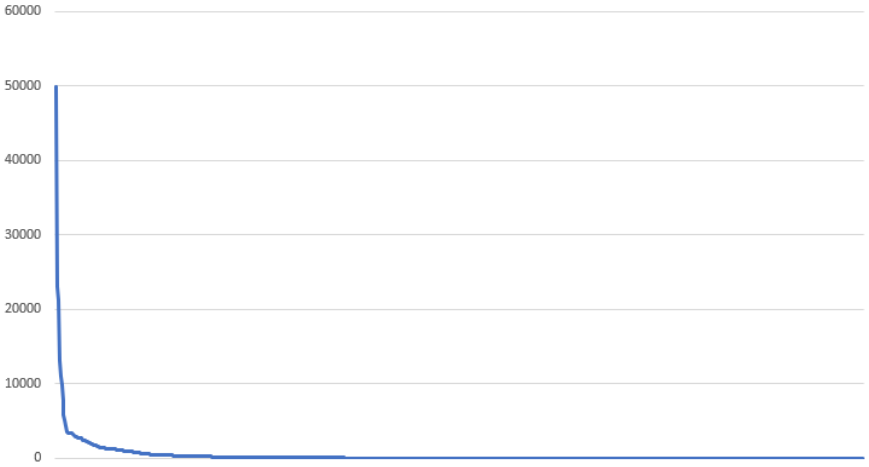
Слика 11. Коначан сурогат одломка оригиналног документа.

[10] Оваква грешка могла би се избећи претходном токенизацијом текста.

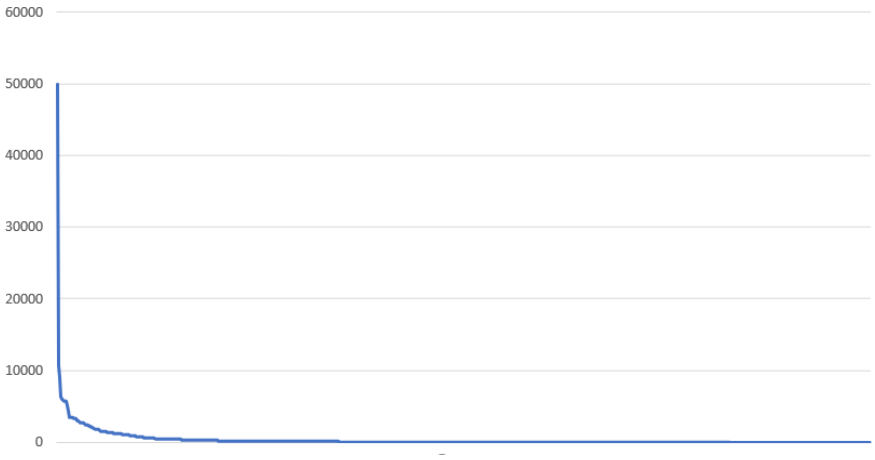
2.3 Уклањање стоп речи и шумова

Пре преласка на класификацију било је потребно је детектовати могуће шумове у виду погрешних обележја или појмова који се јављају у превеликом број докумената те нису дискриминаторни, као и концепата који се често јављају, а тешко да могу успешно класификовати документ. За сваку класу избројан је укупан број понављања, што је пружило увид у неравномерну расподелу (Слика 12).

Први идентификатори који су додати у листу стоп речи и одстрањени јесу они који се односе на географске локације побројане у онтологији под класом Z01 — географске локације као и хомоними попут термина *back*, који се појављује у документима 11.440 пута, очигледно не увек да би означио део људског тела (леђа). На Слици 13 се, међутим, види да је неравномерна расподела фреквенција класа остала и након овог корака.



Слика 12. Расподела фреквенција пре уклањања шумова.



Слика 13. Фреквенција класа у документима после уклањања стоп речи.

2.4 Класификација докумената према идентификаторима

Над документима који су предмет класификације примењена су два поступка обраде, па су тако настала два тест скупа. Контролна метода је била замена идентификатора њиховом класом, општијом хијерархијском ознаком. У експерименталној методи је узета у обзир и дужина идентификатора, па су они замењивани одређеним бројем понављања надређене класе у зависности од њихове дужине, како би се тестирала претпоставка да је коришћење уже специјализованих термина од веће важности за одређивање класе докумената. Дакле, уместо да идентификатори буду скраћени на прва два блока цифара, у обзир је узета њихова дужина, тј. дубина сваког од концепата у дрвету. Идентификатори су замењени одређеним бројем идентификатора класа на основу њихове дужине, на пример: идентификатор **D04.345.295.750.650.700** замењен је помоћу одговарајућег регуларног израза са **D04.345,D04.345** што је еквивалентно појављивању двају концепата који припадају класи **D04.345** у том документу. Начин мапирања дужине концепата у одговарајући број понављања дат је у табели 1. Ту се види да се термини поистовећују са највише четири понављања (у случају да имају преко осам блокова) термина који означава неку класу. Након

примене ових корака, у сурогатима докумената сада се појављују само идентификатори класа, које треба једноставно пребројати.

број додатних блокова (преко 2)	0	1	2	3	4	5	6	7	8	9	10
број резултованих понављања	1	1	2	2	2	3	3	3	4	4	4

Табела 1. Мапирање броја слогова вишка и броја резултујућих идентификатора класа

Након што су тест скупови успешно направљени, за класификацију докумената припремљен је једноставан програм који као улаз захтева датотеку са нискама које означавају класе (прва два блока цифара) који су препознати. Програм једноставно пребројава класе које се јављају у сурогату документа и враћа ону која се најчешће јавља, а уколико постоји више класа које се у документу јављају са истом фреквенцијом, враћа се класа која се прва појављује, што је логично јер је у сурогатима задржан редослед појављивања термина. Такође, неопходно је да сваком документу буде додељен неки идентификатор, па се у случају да неки документ нема додељене идентификаторе, њему додељује један од најопштијих – **H02.403** – који означава медицину.

Када је низ идентификатора у сваком документу сведен на једну класу, она је узета за резултат класификације и прослеђена на евалуацију.

3. Евалуација и упоређивање резултата

Експерименти су извршени над текстовима из TREC Clinical Decision Support 2016 скупа (Подршка клиничком одлучивању).[11] Циљ је био да се документи класификују на основу термина који означавају концепте из онтологије MESH, а који се у њима јављају. Анотирање коришћено у евалуацији ручно је спровео експертски тим. Примењене су уобичајене метрике средње просечне прецизности, одзива[12] и F-

[11] TREC Clinical Decision Support 2016 (на вебу)

[12] Као погрешно негативно класификовани документи узети су релевантни документи који нису укључени у рангирање.

Тест скуп	Узимање дужине идентиф. у обзир	Прецизност @10	Средња просечна прецизност	Одзив	F-мера
1	да	0.58	0.0060	0.0696	0.0108
2	не	0.46	0.0057	0.0648	0.0103

Табела 2. Резултати евалуације класификације.

мере, као и метрика прецизност@10, која одражава успешност враћања релевантних докумената у првих 10 резултата за неки упит.

Табела 2 показује да је узимања дужине идентификатора у обзир донело мало побољшање резултата (5% побољшана прецизност и F-мера, 7% побољшан одзив, 12% прецизност@10).

Узимајући у обзир добијене вредности, одмах се може приметити да је прецизност необично ниска у односу на одзив. Примењивањем детаљније анализе података, примећује се јако велики број лажних погодака (false positive), па са тим опада и прецизност дате стратегије. Овај резултат не изненађује него је у складу са тренутним стандардима у пољу класификације медицинских докумената (Calì et al., 2017). Главни проблем у концептно оријентисаном проналажењу информација у биомедицинској сфери је велики број погрешно негативно класификованих докумената, што води изузетно ниском одзиву. Мала прецизност је тако у овом раду прихватљива јер је надомештена вишим одзивом и многи релевантни документи су враћени на највишим позицијама, са вредностима за прецизност@10 чак до 58%. Ипак и овде постоји простор за напредак.

4. Закључак

У овом раду је представљен приступ класификацији докумената, који се заснива на креирању минималних сурогата текста. У медицинским текстовима најпре се проналазе појмови од значаја који се потом замењују таксономским референцама. Тако екстраховане референце, користе за класификацију текстова помоћу простог алгоритма коришћењем класа из MeSH онтологије, а резултати се потом евалуирају у односу анотацију експертског тима.

Прелиминарни резултати показују погодност понуђеног приступа решењу овог комплексног задатка. Будући рад фокусираће се на смањење лажних погодака како би се унапредиле перформансе система.

Класификација заснована на онтологијама не зависи од домена у ком се примењује, али свакако зависи од расположивих ресурса, конкретно онтологије или таксономије која се користи за класификацију (Rakesh et al., 2001), тако да једном успостављен систем може наићи и на ширу примену. Када је у питању класификација (медицинских) докумената за српски језик, неопходно је најпре припремити ресурсе, при чему би свакако од користи била Међународна класификација болести – Шифарник болести МКБ 10. У оквиру истраживања на Рударско-геолошком факултету креирана је таксономија са овом класификацијом (Kolonja et al., 2016), где је одређен број термина повезан са енглеским и латинским еквивалентима, што омогућава проширење претраге назива концепата и њиховог проналажења у документима. Ипак, треба имати у виду богату морфологију српског језика која би захтевала допуну приступа коришћењем лексичких ресурса специфичних за област медицине за нормализацију текста пре класификације или индексирања, што би припомогло препознавању већег броја таксономских појмова у документима (Stanković et al., 2015).

5. Захвалност

Овај рад је настао у оквиру летње школе *Keyword Search in Big Linked Data* (Претраживање кључним речима велике количине података), организоване у оквиру COST акције Keystone од 21. до 25. августа 2017. године на Технолошком универзитету у Бечу.

Литература

Rakesh et al., Agrawal. “Multilevel Taxonomy Based on Features Derived from Training Documents Classification using Fisher Values as Discrimination Values”, U.S. Patent No. 6,233,575, 2001

Calì, Andrea, Dorian Gorgan и Martin Ugarte. *Semantic Keyword-Based Search on Structured Data Sources: COST Action IC1302 Second International KEYSTONE Conference, IKC 2016, Cluj-Napoca, Romania, September 8–9, 2016, Revised Selected Papers*, Vol. 10151, 2017

Dragoni, Mauro. "3rd KEYSTONE Summer School", 2017, URL http://ifs.tuwien.ac.at/keystone.school/slides/Dragoni_SemanticSearch.pptx

Elberrichi, Zakaria, Malika Taibi и Amel Belaggoun. "Multilingual Medical Documents Classification Based on MesH Domain Ontology". *International Journal of Computer Science Issues* Vol. 9 (2012)

Kolonja, Ljiljana, Ranka Stanković, Ivan Obradović, Olivera Kitanović и Aleksandar Cvjetić. "Development of terminological resources for expert knowledge: a case study in mining". *Knowledge Management Research & Practice* Vol. 14, no. 4 (2016): 445–456

Stanković, Ranka, Cvetana Krstev, Ivan Obradović и Olivera Kitanović. "Indexing of Textual Databases Based on Lexical Resources: A Case Study for Serbian". У *Semantic Keyword-based Search on Structured Data Sources*, Cardoso, Jorge, Francesco Guerra, Geert-Jan Houben, Alexandre Miguel Pinto и Yannis Velegrakis, 167–181. Cham: Springer International Publishing, 2015

Trieschnigg, Dolf, Piotr Pezik, Viv Lee, Franciska de Jong, Wessel Kraaij et al.. "MeSH Up: Effective MeSH Text Classification for Improved Document Retrieval". *Bioinformatics (Oxford, England)* Vol. 25 (2009): 1412–8