УНИВЕРЗИТЕТ У БЕОГРАДУ

ЕЛЕКТРОТЕХНИЧКИ ФАКУЛТЕТ

ОБРАДА ПРИРОДНИХ ЈЕЗИКА

Пројектни задатак

Ментор:

проф. др Бошко Николић

др Вук Батановић

Кандидати:

Јован Стевановић - 2019/3154

Владимир Сивчев - 2019/3172

Предраг Митровић - 2019/3173

Матија Лукић - 2019/3279

Београд, јун 2020.

САДРЖАЈ

САДРЖАЈ **2**

1. УВОД **3**

2. ПРИКУПЉАЊЕ ПОДАТАКА (I ФАЗА) **5**

3. АНОТИРАЊЕ ПОДАТАКА (II ФАЗА) **10**

4. ОБРАДА ПОДАТАКА (III ФАЗА) **15**

4.1. ПРЕПРОЦЕСИРАЊЕ ПОДАТАКА 15

4.1.1. Без препроцесирања *15*

4.1.2. Lower casing *16*

4.1.3. TF, IDF, TF-IDF *18*

4.1.4. Stemmer and stop words *21*

4.1.5. Frequency word filtering *22*

4.1.6. Bigrams and trigrams *24*

4.2. ТРЕНИРАЊЕ КЛАСИФИКАТОРА 26

4.2.1. Мултиномијални Бајесов наивни класификатор *27*

4.2.2. Бернулијев наивни класификатор *28*

4.2.3. Логичка регресија *28*

4.2.4. Метод потпорних вектора без кренела *29*

1. Увод

Пројектни задатак на предмету ***Обрада природних језика*** служи за примену стеченог теоријског знања на конкретном проблему. У оквиру овог пројектног задатка било је потребно **прикупити**, **анотирати** и **истренирати класификатор** коментара написаних на програмском језику *SQL*.

За израду овог рада било је потребно удружити неколико различитих технологија. Примарни скуп података је добијен кроз *API GitHub-a*, који је јавно доступан. *GitHub* тренутно представља најпопуларнији алат за верзионисање пројеката различитих величина. Кроз *API* може се приступи свим пројектима који су *open-source* статуса.

За прикуплање и анотирање података коришћено је неколико различитих алата. Наиме, неки од ових алата су специјално израђени као део пројекта, док су други већ постојећи који на један или други начин помажу за израду ове две фазе. Поставком пројекта је неопходно скупити и анотирати 3000 коментара.

Након прикупљања и анотирања података, прелазимо у трећу фазу која подразумева препроцесирање анотираних података и тренирање класификатора. Поставком пројекта је захтевано да се искористе следеће технике препроцесирања:

- *Lower casing* - пребацивање свих коментара на мала слова,
- *TF, IDF, TF-IDF* – подразумевају три методе које се базирају на фреквенцији појаве речи у коментарима,
- *Stemmer and stop words* – процесирање коментара коришћењем штемера и стоп речи,
- *Frequency word filtering* – филтрирање речи коришћењем учестаности њихове појаве, и
- *Bigrams and trigrams* – уместо да се процесирање врши на основу сваке појединачне речи, у овом случају правимо биграме и триграме од коментара.

*П*осле претпроцесирања коментара, поставком задатка је одређено да се примене следећи класификатори:

- *Мултиномијални Бајесов наивни класификатор*,

- *Бернулијем наивни класификатор,*
- *Логистичка регресија, и*
- *Класификација методом потпорног вектора без кернела.*

Последња фаза је подељена на две потфазе. Прва потфаза подразумева да се сви коментари који описују неку функционалност рачунају да су истог типа. Друга потфаза подразумева да се функционални тип коментара дели на три категорије које ближе описују циљ коментара.

2. Прикупљање података (I фаза)

Као што је већ речено, као главни и једини извор података за израду овог пројекта коришћен је *GitHub*. Сви пронађени пројекти су *OpenSource* статуса и коришћењем *GitExtensions* алата, исти пројекти се клонирају како би сада били доступни на локалном рачунару где се само прикупљање обавља.

Како би се подаци прикупили, првенствено је било неопходно пронаћи репозиторијуме из којих би се коментари извукли. Ово је било могуће одрадити на више начина, међутим у пројекту је донешена следећа одлука. Наиме, првенствено је била извршена претрага на *Google* систему, где су били претраживани било какви пројекти *SQL* типа, или да у себи садрже модуле који се баве неким *SQL*-ом. Резултати овакве претраге су навели на наредне сајтове, где постоје референце на разне пројекте који укључују *SQL*, а који се налазе на *GitHub*-у.

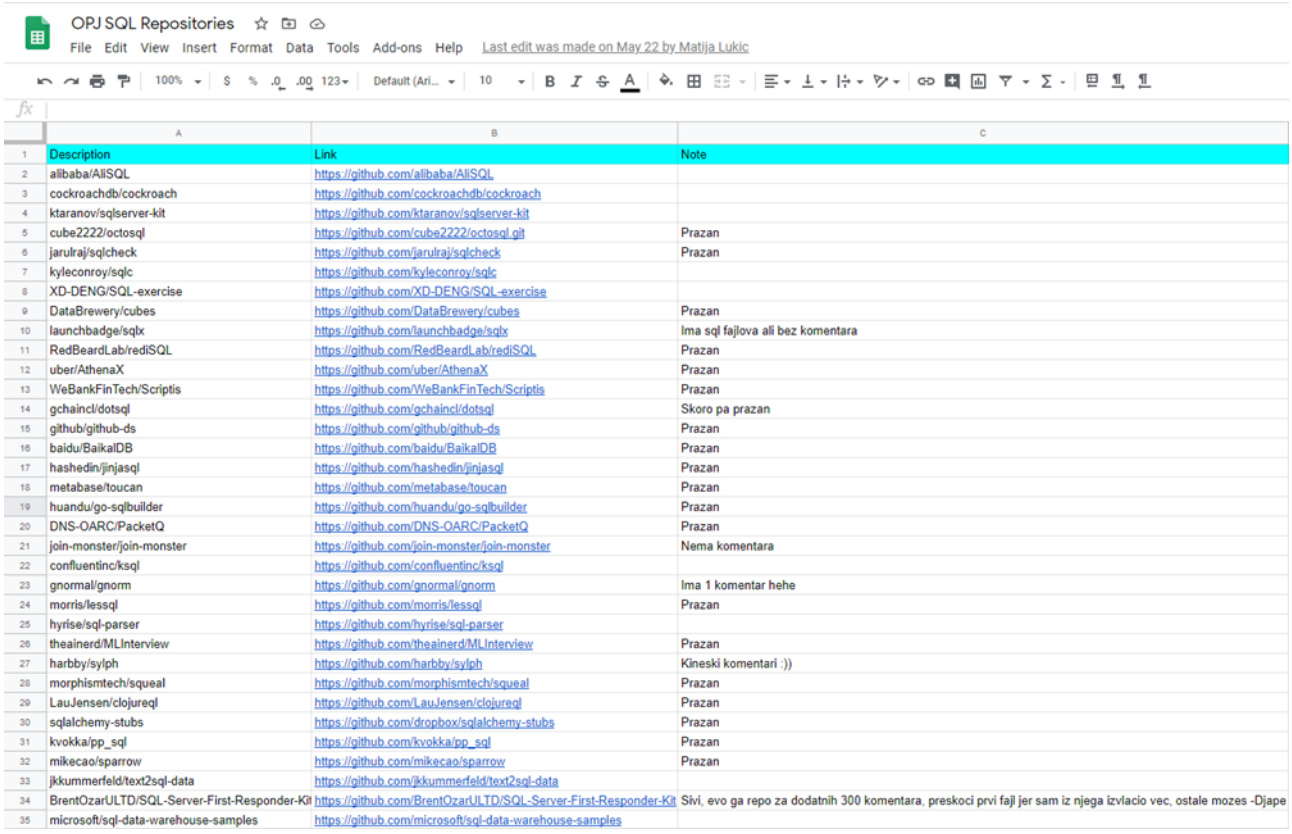
https://medium.com/issuehunt/50-top-projects-of-sql-on-github-in-2018-aabe0950a43a

https://awesomeopensource.com/projects/sql

https://awesomeopensource.com/projects/sql-server

Са претходних линкова је примарна претрага за репозиторијуме извршавана. Међутим, не садржи сваки референцирани репозиторијум оно што је у пројекту од интереса. Наиме, у овом тренутку се појављује неколико проблема.

С обзиром да је реч о тимском пројекту, била је неопходна синхронизација на то како ће се подаци прикупљати. Како би се избегло да више особа прикупљају податке са истог репозиторијума, направљен је *Google SpreadSheet* документ где су записивани до сада претраживани и разматрани репозиторијуми, уз додатне потенцијалне напомене за сваки репозиторијум. На слици 1 се може видети приказ овог документа, са првих 35 разматраних репозиторијума.



OPJ SQL Repositories

File Edit View Insert Format Data Tools Add-ons Help Last edit was made on May 22 by Matija Lukic

	A	B	C
1	Description	Link	Note
2	alibaba/AliSQL	https://github.com/alibaba/AliSQL	
3	cockroachdb/cockroach	https://github.com/cockroachdb/cockroach	
4	ktaranov/sqlserver-kit	https://github.com/ktaranov/sqlserver-kit	
5	cube2222/octosql	https://github.com/cube2222/octosql.git	Prazan
6	jarulraj/sqlcheck	https://github.com/jarulraj/sqlcheck	Prazan
7	kyleconroy/sqlc	https://github.com/kyleconroy/sqlc	
8	XD-DENG/SQL-exercise	https://github.com/XD-DENG/SQL-exercise	
9	DataBrewery/cubes	https://github.com/DataBrewery/cubes	Prazan
10	launchbadge/sqlx	https://github.com/launchbadge/sqlx	Ima sql fajlova ali bez komentara
11	RedBeardLab/rediSQL	https://github.com/RedBeardLab/rediSQL	Prazan
12	uber/AthenaX	https://github.com/uber/AthenaX	Prazan
13	WeBankFinTech/Scriptis	https://github.com/WeBankFinTech/Scriptis	Prazan
14	gchaincl/dotsql	https://github.com/gchaincl/dotsql	Skoro pa prazan
15	github/github-ds	https://github.com/github/github-ds	Prazan
16	baidu/BaikalDB	https://github.com/baidu/BaikalDB	Prazan
17	hashedin/jinjasql	https://github.com/hashedin/jinjasql	Prazan
18	metabase/toucan	https://github.com/metabase/toucan	Prazan
19	huandu/go-sqlbuilder	https://github.com/huandu/go-sqlbuilder	Prazan
20	DNS-OARC/PacketQ	https://github.com/DNS-OARC/PacketQ	Prazan
21	join-monster/join-monster	https://github.com/join-monster/join-monster	Nema komentara
22	confluentinc/ksql	https://github.com/confluentinc/ksql	
23	gnormal/gnorm	https://github.com/gnormal/gnorm	Ima 1 komentar hehe
24	morris/lessql	https://github.com/morris/lessql	Prazan
25	hyrise/sql-parser	https://github.com/hyrise/sql-parser	
26	theainerd/MLInterview	https://github.com/theainerd/MLInterview	Prazan
27	harbby/sylph	https://github.com/harbby/sylph	Kineski komentari :))
28	morphismtech/squeal	https://github.com/morphismtech/squeal	Prazan
29	LauJensen/clojureql	https://github.com/LauJensen/clojureql	Prazan
30	sqlalchemy-stubs	https://github.com/dropbox/sqlalchemy-stubs	Prazan
31	kvokka/pp_sql	https://github.com/kvokka/pp_sql	Prazan
32	mikecao/sparrow	https://github.com/mikecao/sparrow	Prazan
33	jkkummerfeld/text2sql-data	https://github.com/jkkummerfeld/text2sql-data	
34	BrentOzarULTD/SQL-Server-First-Responder-Kit	https://github.com/BrentOzarULTD/SQL-Server-First-Responder-Kit	Sivi, evo ga repo za dodatnih 300 komentara, preskoci prvi fajl jer sam iz njega izvlacio vec, ostale mozes -Djape
35	microsoft/sql-data-warehouse-samples	https://github.com/microsoft/sql-data-warehouse-samples	

Слика 1 – Синхронизациони документ за бирање репозиторијума

Као што се са слике 1 може видети, постоје одређени репозиторијуми који су празни. Такви репозиторијуми чак нису ни у мањини. Други проблем који је овде настао јесте потенцијална могућност на губитак у продуктивности (тако што би дати репозиторијуми улазили даље у наредне кораке процеса у пројекту, а да се на крају испостави да је то било непотребно је су репозиторијуми празни – немају коментаре који су од интереса). За ово је искоришћена *GitHub* опција која омогућава брзо површно (људско) скенирање репозиторијума које помаже у доношењу одлуке да ли овакав репозиторијум треба даље разматрати или га аутоматски искључити и прећи на други.

Поменута *GitHub* опција јесте *Find File* која омогућава да се репозиторијум претражује по фајловима, а у пројекту је наравно од интереса било извршавати претрагу на „*.sql*“, чиме се ефективно добијају сви фајлови са екстензијом *.sql*. Сада, могуће је летимично проћи кроз неколико фајлова и у зависности од процене човека, донети одлуку да ли је репозиторијум вредан даљег процесирања, или га треба одбацити са евентуалном напоменом да је празан.

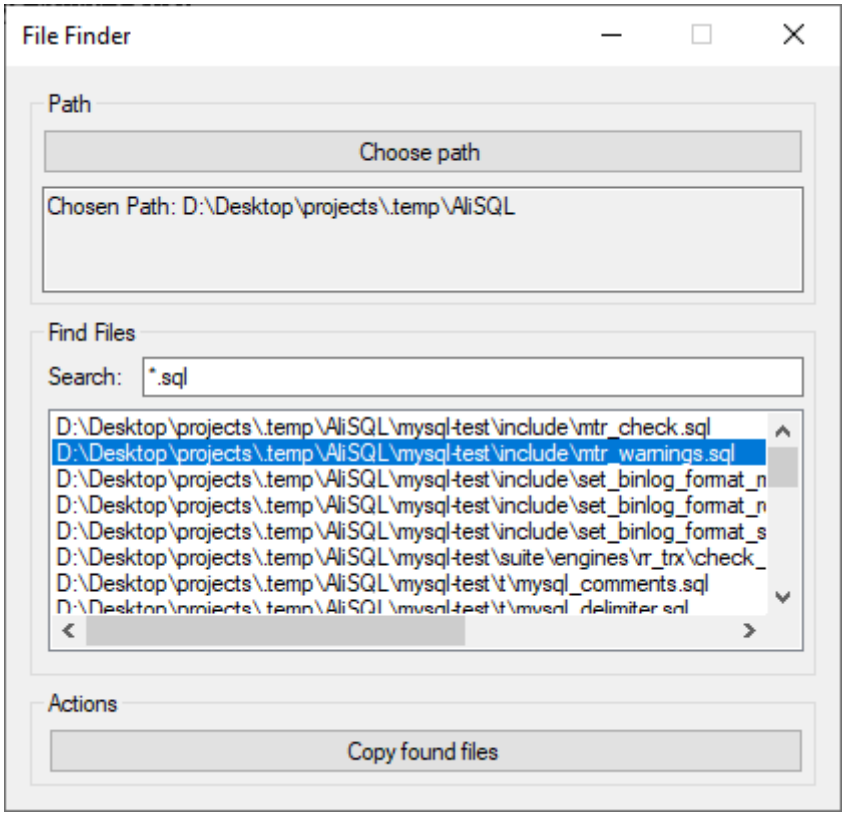
И празне репозиторијуме треба наводити у синхронизационом документу, тако да би наредне особе, уколико налете на исти, могле још брже да пређу на наредни репозиторијум, јер је тај већ био „процесиран“.

Када се пронађе репозиторијум који садржи коментаре од интереса, наредни корак у процесирању и прикупљању података јесте клонирање овог репозиторијума на локални рачунар како би било лакше обрадити репозиторијум даље. Наравно, и овај корак је могао другачије да тече, али је у пројекту донешена ова одлука из више разлога, а најбитнији јесте продуктивност, јер се користи већ направљени алат који има приступ одређеним деловима *GitHub API*-ја којем овај пројекат нема.

Већ је поменуто да је у ову сврху коришћен *GitExtensions* алат, мада било који други *Git* клијентски алат је потпуно равноправан овом. Једина опција која је овде од интереса јесте за клонирање датог репозиторијума.

Како сами пројекти могу имати произвољну структуру директоријума и датотека, потребно је на један или други начин доћи до *.sql* фајлова који потенцијално садрже потребне коментаре, али сада да аутоматизованији начин, а не ручном претрагом (ово је сада много лакше јер се фајлови налазе на локалном рачунару). Постоји више могућности како се ово може обавити, а у оквиру тима донета је пројектна одлука да се направе 2 нова алата који ће увећати продуктивност прикупљања података у овој фази.

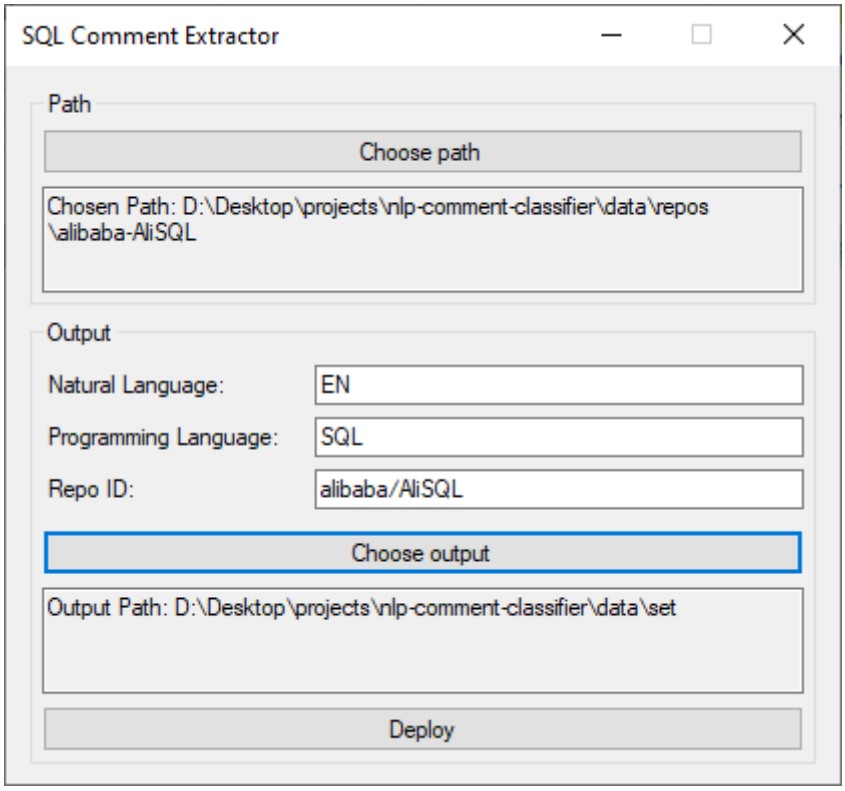
Наиме, први алат помаже у проласку кроз директоријуме клонираног пројекта и тражи фајлове по избору. На слици 2 се може видети изглед овог алата.



Слика 2 – Претраживач фајлова

Кроз овај алат могуће је одабрати основни фолдер у ком претрага започиње. Ово треба да буде основни (енгл. *Root*) директоријум клонираног пројекта у ком на даље треба пронаћи све потребне *.sql* фајлове. Уз помоћ оприје *Search* могуће је унети било какав израз по ком ће се претрага извршавати. Након притиска на тастер ентер, испод се излиставају сви пронађени фајлови из претраге. Такође, овај алат нуди могућности копирања пронађених фајлова на другу локацију притиском на опцију *Copy found files*. Сви пронађени фајлови ће се копирати на другу, изабрану локацију заједно са њиховим подфолдерима у оквирју самог пројекта. По конвенцији у оквиру тима, копирање се врши у оквиру репозиторијума самог текућег пројекта, тако да се сви фајлови из којих су коментари прикупљани такође налазе у оквиру овог репозиторијума.

Након копирања фајлова, неопходно је из њих извући коментаре и сложити их у један фајл по одређеном формату, који је описан у поставци пројекта. За ово је израђен други алат, чији се изглед може видети на слици 3.



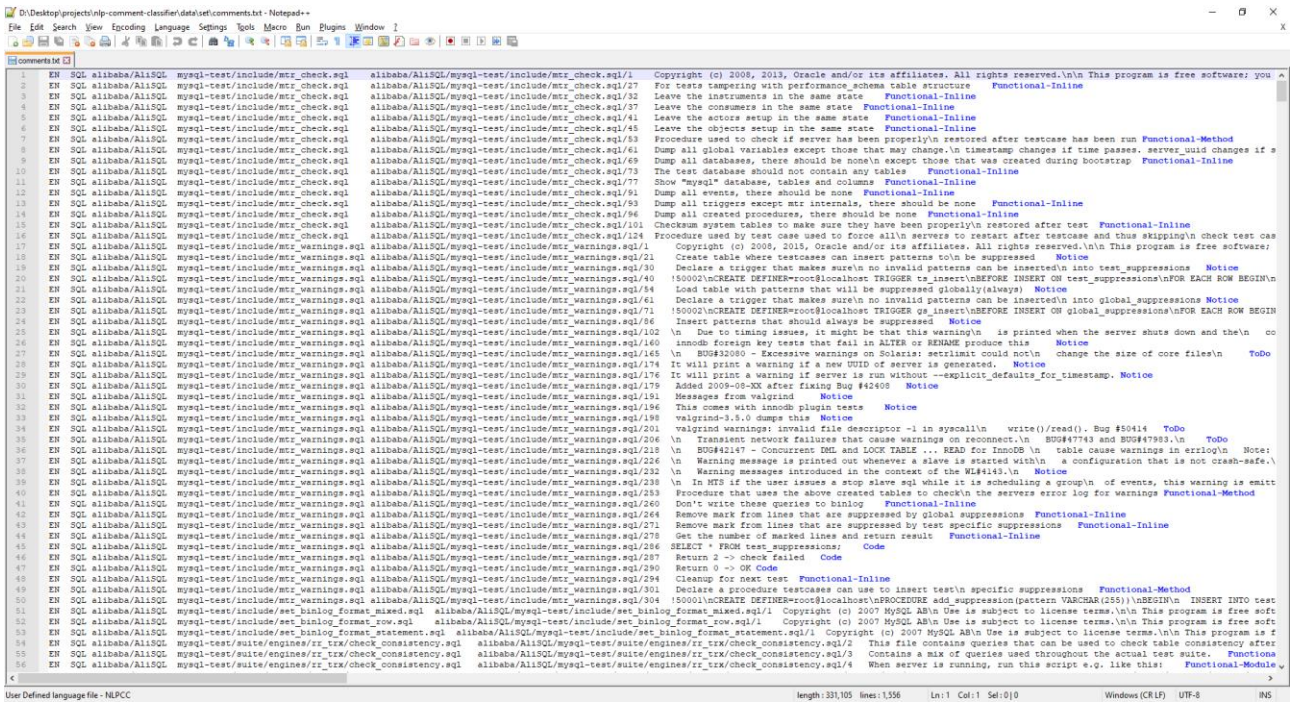
Слика 3 – Екстрактор коментара

Два наведена алата су тако пројектована да су компатибилна. Наиме, након што се фајлови копирају коришћењем првог алата, дестинациони фолдер копирања сада може да буде изворишни фолдер одакле се коментари извчаче, што се може подесити коришћењем опције *Choose path*. Након тога, треба евентуално подесити *Repo ID* тако да се уместо знака „-“ (конвенција усвојена у тиму пројекта) ставити знак „/“, како би сам *ID* био валидан. Потом, треба одабрати дестинациони фолдер где ће се направити два излазна фајла која ће садржади извучене коментаре из изворишта, по формату описаном у поставци пројекта. Наиме, уколико излазни фајлови већ постоје, алат само надовезује на претходни садржај тих фајлова (операција *append file*).

Притиском на опцију *Deploy*, алат пролази кроз све фајлове кроз читаву подструктуру датог изворишног фолдера, и у сваком пронађеном фајлу покушава да пронађе све врсте коментара који постоје у *.sql* језику.

3. АНОТИРАЊЕ ПОДАТАКА (II ФАЗА)

Након прикупљања података са горе описаним алатима, неопходно је анотирати којој класи припадају ови подаци. Овај посао је рађен ручно, и то коришћењем одређених окружења, као што је *VSCode, Nodepad++*, *Sublime* и слично. На слици 4 се може видети принцип анотирања у *Notepad++* окружењу.



Слика 4 – Анотирање у *Notepad++*

Са слике се може видети да су плавом бојом означене конкретне класе датих коментара по формату задатом у поставци пројекта. Наиме, сваки коментар је неопходно ручно обрадити на овај начин.

Ова фаза је донекле текла паралелно са фазом 2. Наиме, члан тима након што пронађе валидан репозиторијум и извезе коментаре описане у фази 1, такође треба и да прође кроз њих како би проверио да ли су потребне евентуалне дораде око коментара, као што је на пример, скупљање једнолинијских коментара нанизаних један иза другог у један коментар, јер уместо да је сам коментар написан у један вишелинијски коментар, он је био написан као више једнолинијских коментара, један за другим. У овом случају, потребно је скупити те коментаре и написати их као један вишелинијски коментар. У току овог пролаза могуће је паралелно и анотирати дати коментар.

Анотација сама по себи представља процес који је у одређеним ситуацијама једноставан и јасан, а такође у другим ситуацијама комплексан и не тако очигледан. Наиме, од класа које су описане у поставци пројекта, коментари типа *ToDo* и *Code* су они који су најједноставнији за уочавање и око њих уопште није било недоумица. Такође, коментари типа лиценца и ауторских права, информације о верзији, о аутору и сличних ствари су такође са лакоћом били разрешавани, а то су били коментари *General* класе.

Насупрот другим језицима где је то поприлично јасно, у *SQL* језику је тешко приметити коментаре класе *IDE*. Међутим, ипак су успешно били кроз прву фазу пронађени коментари који представљају шаблонске коментаре налик на *JavaDoc* анотатирање. Неколико примера око оваквих коментара:

```
name: GetAuthor :one
name: GetBook :one
name: DeleteBook :exec
name: BooksByTitleYear :many
```

Поред тога, постоје и још неки одређени коментари који би потенијално могли припадати класи *General* јер садрже информације о верзији и слично, међутим, у себи садрже име конкретне базе која је коришћена попут *MySQL* и слично, што означава да је овај коментар ипак био генерисан од стране коришћеног окружења, па је и сам коментар у том случају сврставан у класу *IDE*. Пример таквог коментара:

```
MySQL dump 10.13  Distrib 8.0.12, for osx10.14 (x86_64)\n Host: localhost  Database: cockroachtestdata \nServer version
```

Иако постоје одређене разлике у анотирању од стране сваког анотатора (што ће бити представљено касније), неке ствари су биле међусобно усаглашене током анотирања. На пример, донешена је пројектна одлука да се коментари који у себи садрже само линк ка одређеној *Web* страници сврстају у класу *Notice*. Наиме, разлог иза тога је био да обично линкови упућују програмера који гледа код ка неким додатним ресурсима који му помажу да лакше разуме о чему се у датом тренутку ради, или како би се нешто боље/сликовитије представило, објаснило и слично.

Класа *Notice* такође представља једну класу која у неким ситуацијама може да равноправно посматра као и класа *Functional-Inline,* па да постоји конфликт у томе која би се од те две класе користила. Наиме, ово је једна од ситуација коју сваки анотатор решава сам по себи, јер није био пронађен ни један разлог да се уведе одређени систем по ком би се одређена класа преферирала.

У одређеним ситуацијама, неки коментари су били дељени на 2 дела, тако да један део припада једној класи, а други другој.

Класа *Functional-Inline* у другим ситуацијама представља класу коју није толико тешко пропустити. Било какав опис кода који надолази након датог коментара јасно означава да тај коментар припада управо овој класи. Међутим, и овде постоје одређени изузеци. Поред тога, као што је већ речено, да ова класа може да се конфликтује са *Notice* класом, када постоји опис неког кода, ова класа се може конфликтовати са *Functional-Method* класом по овој логици. Међутим начин погледа како би се те две класе ефикасно разликовале јесте сам контекст. На пример, када би то био коментар у коду изнад дефиниције *Stored Procedure*, очигледно је да је реч о коментару типа *Functional-Method*. Међутим, овде може да постоји потенцијални проблем код класификатора касније. Наиме, уколико постоје два коментара који су веома слични, или чак можда идентични, а један се налази пре дефинисања одређене методе, док је други унутар ње, анотатор ће користити контекст како би одлучио који коментар припада једној од поменуте две класе. Међутим, сам класификатор нема такву информацију, па он види исти коментар (или сличан), сврстан у две различите класе (у тренутку када се ове класе разликују, а не када се све *Functional* класе посматрају као исте). Ово може да потенцијално умањи перформансе класификатора у трећој фази.

Слична дискусија може да се изведе и за *Functional-Module* класу. Међутим, с обзиром да *SQL* језик није објектно оријентисан (а ова класа углавном може да се користи за коментаре који описују читаву класу – читав модул), може се наслутити да таквих коментара и нема. Међутим, ово није истинито. Наиме, постоје одређени коментари који описују читав фајл, или можда део фајла који итекако припадају овој класи и никако ни не могу да се сврстају у неку другу, па је анотирање истих било једноставно, а такви коментари су се знатно разликовали од коментара других *Functional* класа, па ранија дискусија и не може тек тако да се примени на ову класу.

Поред сета за тренирање класификатора, додатно је још прикупљан одређени број коментара где је сваки члан тима засебно вршио анотацију, како би се срачунала сличност, односно сагласност анотатора. С обзиром да се ово десило након анотирања оригиналног сета, одређене одлуке које су заједно у тиму донете, потенцијално могу утицати на овај део, јер се анотатори, иако треба засебно, без комуникације да ово одраде, могу водити истим правилима као што су се водили кроз анотирање оригиналног сета. Међутим, ово ипак не представља проблем из наредног разлога. Циљ уношења оваквих пројектних одлука и јесте био да се анотатори синхронизују тако да анотација на крају буде што боља, што сагласнија, као да је један анотатор исти посао радио, како би се касније класификатору давали подаци који имају смисла. Али чак и са овим, и даље могу да постоје мале разлике у томе када 2 различита анотатора врше овај посао. Чак и са овим, у

издвојеном сету, сагласност између свака два анотатора у просеку износи негде око 90%, што указује на то да ипак постоји мали део анотирања који је субјективан и који може да утиче на класификатор. Међутим, ово не значи да је тај утицај лош. Наравно, није ни искључено да овај утицај може бити лош, али такође може и потенцијално позитивно да утиче на класификатор тако да не долази до преобучавања класификатора. У табели 1 се могу видети конкретни резултати рачунања сагласности.

Табела 1 – Сагласност анотатора	
Владимир Сивчев – Јован Стевановић	87.00%
Владимир Сивчев – Предраг Митровић	96.00%
Владимир Сивчев – Матија Лукић	95.33%
Јован Стевановић – Предраг Митровић	85.67%
Јован Стевановић – Матија Лукић	91.67%
Предраг Митровић – Матија Лукић	94.00%
Глобална сагласност	91.61%

У сваком случају, сам процес анотације је морао да тече ручно. Одређени коментари су такође били у потпуности уклањани из оригиналног и додатног сета, јер су или узроковали одређене проблеме, како за само тренирање касније, или једноставну неодлучност анотатора у томе којој класи би коментар припадао, или чак да ни једна класа није адекватна датом коментару. Таквих коментара је било веома мало, али су ипак постојали. У табели 2 су представљене расподеле коментара по класама.

Табела 2 – Расодела коментара по класама	
Functional-Inline	1259
Functional-Method	79
Functional-Module	31
ToDo	35
Notice	433
General	110
Code	838
IDE	235

4. Обрада података (III фаза)

У овом поглављу је детаљно описана фаза препроцесирања података, тренирање класификатора и израчунавање квалитета класификатора.

4.1. Препроцесирање података

Као што је наведено у уводу, у пројектној постави је захтевано више различитих типова препроцесирања. Сви резулати су подељени у две фазе. Прва фаза подразумева да се сви функцијски коментари посматрају као једна класа на излазу, док се у другој фази сваки од типова функцијских коментара посматра као засебна класа.

4.1.1. Без препроцесирања

Код овог типа, извршава се све оно што се извршава и пре свих осталих фаза препроцесирања. То подразумева да се уклоне све непотребне колоне, уклоне специјални карактери, *NA* вредности и дупликати.

Резултати евалуације без препроцесирања служе као референтна тачка за посматрање утицаја који имају на перформансе сви остали типови препроцесирања.

Резултати прве фазе:

	I	II	III	IV	V	VI	VII	VIII	IX	X	C	AVG
МНБК	86.75	87.75	88.41	85.76	86.42	87.09	88.74	87.09	85.43	84.77	/	86.82
БНБК	79.80	82.45	80.46	79.80	79.47	78.48	80.13	81.13	81.13	82.78	/	80.56
ПВ(Л1)	87.75	90.40	91.06	89.40	89.74	87.75	88.74	89.40	89.07	89.07	/	89.24
ПВ(Л2)	89.07	89.74	92.38	87.09	89.07	91.72	88.74	88.08	90.73	91.39	/	89.80
ПВ(О)	92.72	91.72	91.06	90.40	90.07	88.74	92.05	88.08	88.08	92.05	0.916	90.50
ЛР(Л1)	86.42	88.41	84.44	88.08	89.07	88.08	85.76	85.76	89.74	89.74	/	87.55
ЛР(Л2)	90.40	89.40	90.73	88.74	86.42	89.07	87.42	87.75	89.07	89.74	/	88.87
ЛР(О)	89.74	90.73	86.42	84.44	90.72	88.74	88.08	89.07	86.09	85.10	0.892	87.91

Резултати друге фаза:

	I	II	III	IV	V	VI	VII	VIII	IX	X	C	AVG
МНБК	85.10	84.11	83.77	85.43	85.10	83.44	82.78	83.44	86.42	82.12	/	84.17
БНБК	79.14	76.82	80.46	73.51	77.81	78.15	77.48	75.17	78.81	75.83	/	77.32
ПВ(Л1)	89.40	86.42	87.09	85.76	88.74	85.43	86.42	87.42	84.77	87.09	/	86.85
ПВ(Л2)	87.75	86.42	86.09	86.42	89.07	89.07	89.07	87.75	87.42	88.08	/	87.72
ПВ(О)	83.11	88.41	89.74	88.74	86.42	88.74	87.75	90.73	86.42	87.42	1.471	87.75
ЛР(Л1)	90.07	87.75	88.08	86.75	85.10	87.75	87.75	83.77	87.42	83.44	/	86.79
ЛР(Л2)	86.75	86.09	88.41	84.77	89.74	86.42	89.40	87.42	87.09	90.40	/	87.65
ЛР(О)	86.42	84.44	85.10	87.42	88.08	89.07	88.08	89.07	89.40	87.42	1.504	87.45

У првој фази примећујемо да су перформансе Бајесовог класификатора између 86-89%, што је заправо најнижи степен тачности који ћемо имати. Што је сасвим у реду, ако узмемо у обзир да немамо никаквог додатног препроцесирања. Сличну тачност имамо и код логичке регресије и потпорног вектора са Л1 функцијом грешке.

Још слабије пероформансе добијамо Бернулијевог класификатора где је степен тачности између 79-83%.

Највећи степен тачности, стабилно око 87%, имамо код логичке регресије и потпорног вектора са Л2 функцијом грешке.

У другој фази примећујемо сличне перформансе које су за отприлике 2% ниже, уколико кумулативно рачунамо разлику у односу на прву фазу.

4.1.2. *Lower casing*

Подразумева иницијалну обраду, након које се сви коментари пребацују на мала слова.

Резултати прве фазе:

	I	II	III	IV	V	VI	VII	VIII	IX	X	C	AVG
МНБК	88.41	83.44	90.40	85.43	87.09	86.42	89.40	88.41	84.11	87.42	/	87.05
БНБК	80.79	83.44	77.48	78.48	79.80	81.46	78.15	77.15	83.11	81.79	/	80.17

ПВ(Л1)	84.44	88.08	87.09	87.75	86.75	87.42	90.73	87.09	88.74	87.75	/	87.58
ПВ(Л2)	89.40	88.41	86.75	85.76	89.74	89.40	92.05	90.40	87.09	91.06	/	89.01
ПВ(О)	90.73	89.40	91.39	88.41	90.73	93.71	89.74	87.09	93.05	90.07	0.924	90.43
ЛР(Л1)	89.40	87.09	85.76	87.75	87.09	86.75	88.74	88.41	89.74	89.40	/	88.01
ЛР(Л2)	89.40	90.07	88.74	89.74	86.42	92.38	90.07	89.40	86.09	87.42	/	88.97
ЛР(О)	90.07	90.40	89.74	86.42	87.09	87.75	89.74	88.08	85.76	87.42	0.912	88.25

Резултати друге фазе:

	I	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**C**	**AVG**
МНБК	85.43	89.07	85.43	83.77	86.42	87.42	84.44	84.44	88.41	84.77	/	85.96
БНБК	77.15	79.47	77.48	80.79	80.46	76.16	78.81	80.79	78.15	75.17	/	78.44
ПВ(Л1)	86.42	87.75	87.42	86.42	87.09	84.77	89.40	87.42	87.42	85.43	/	86.95
ПВ(Л2)	87.42	90.73	87.09	89.74	86.75	86.09	88.74	85.10	84.44	88.74	/	87.48
ПВ(О)	87.75	86.09	88.08	88.08	86.42	91.39	88.41	87.75	83.15	89.07	1.503	87.62
ЛР(Л1)	85.10	86.42	85.43	82.12	87.09	87.75	85.76	85.10	85.76	88.08	/	85.86
ЛР(Л2)	88.41	87.75	85.76	89.40	88.41	87.75	87.09	84.77	89.07	83.44	/	87.19
ЛР(О)	90.07	87.75	87.75	89.74	87.42	88.41	86.75	85.10	88.08	86.75	1.483	87.78

У првој фази примећујемо да су перформансе Бајесовог класификатора између 86-89%, што је у просеку за 5% боља тачност у односу на иницијално тестирање. Што је сасвим у реду, ако узмемо у обзир да од додатног препроцесирања једино имамо претварање свих слова у мала.

У односу на иницијално тестирање имамо напретка код логичке регресије и потпорног вектора. Док за оне са Л2 функцијом грешке имамо опет највећи проценат тачности од свих класификатора.

Код Бернулијевог класификатора је степен тачности између 78-82%. Што је побољшање у односу на иницијално тестирање.

У другој фази примећујемо још сличније перформансе у односу на иницијално тестирање, што је свакако битно због стабилности решења.

4.1.3. *TF, IDF, TF-IDF*

Прва наведена метода подразумева да се као метрика узима учестаност појаве неког термина у тексту (***T**erm **F**requency*), друга метрика користи логаримску фунцкију учестаности појаве термина (***I**nverse **D**ocument **F**requency*), трећа представља комбинацију две горе наведене.

У сврху рачунања ових метрика користи се класа *TfIdfVectorizer* из пакета *sklearn.feature_extraction.text.*

Резултати прве фазе за *TF*:

	I	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**C**	**AVG**
МНБК	85.10	83.44	81.13	81.46	80.79	82.78	80.46	83.11	79.144	84.11	/	82.15
БНБК	80.13	79.47	76.82	77.15	81.46	82.12	79.80	78.81	79.14	81.79	/	79.67
ПВ(Л1)	86.42	89.74	86.42	86.42	86.75	90.73	89.40	89.74	88.71	90.40	/	88.44
ПВ(Л2)	89.07	90.07	89.74	91.72	88.74	93.38	89.07	91.72	87.75	90.07	/	90.13
ПВ(О)	87.42	90.07	88.08	91.06	88.74	87.42	86.42	90.07	90.40	89.40	0.898	88.91
ЛР(Л1)	80.46	80.13	84.44	83.79	87.75	86.09	86.09	86.42	83.44	83.44	/	84.01
ЛР(Л2)	84.11	84.11	87.09	87.75	83.44	87.75	89.40	82.45	86.75	84.11	/	85.70
ЛР(О)	87.42	85.76	86.75	87.09	88.09	86.09	84.07	86.09	87.09	86.42	1.033	86.56

Резултати прве фазе за *IDF*:

	I	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**C**	**AVG**
МНБК	88.08	88.08	88.74	86.09	87.09	86.75	89.40	86.42	86.75	89.74	/	87.72
БНБК	81.79	83.11	82.12	78.81	82.45	81.46	79.80	79.14	81.79	82.78	/	81.32
ПВ(Л1)	90.07	89.74	90.73	88.74	90.40	89.40	87.75	87.42	89.40	91.39	/	89.50
ПВ(Л2)	90.40	91.39	88.41	89.07	89.40	90.40	85.76	90.40	89.74	87.09	/	89.21
ПВ(О)	89.74	89.40	90.07	91.39	89.40	88.41	89.40	89.40	90.40	87.75	0.877	89.54
ЛР(Л1)	80.46	80.13	84.44	81.79	87.75	86.09	86.09	86.42	83.44	83.44	/	84.01
ЛР(Л2)	91.72	88.08	90.40	84.44	87.09	88.08	87.42	85.10	91.39	89.40	/	88.31
ЛР(О)	89.40	92.05	89.07	90.40	87.42	89.40	90.07	84.77	89.07	90.73	0.925	89.24

Резултати прве фазе за *TF-IDF*:

	I	II	III	IV	V	VI	VII	VIII	IX	X	C	AVG
МНБК	85.10	87.09	85.10	86.75	84.44	84.44	84.77	85.76	85.43	85.10	/	85.40
БНБК	80.13	77.48	83.44	80.46	81.46	79.80	81.46	82.45	80.13	81.13	/	80.79
ПВ(Л1)	89.70	87.09	86.75	89.40	90.07	90.40	87.09	85.43	88.08	89.07	/	88.25
ПВ(Л2)	88.41	90.07	93.05	89.40	93.71	89.07	89.40	87.75	90.40	89.07	/	90.03
ПВ(О)	90.40	87.09	88.74	85.76	89.74	89.74	86.09	88.41	89.40	89.74	0.897	88.51
ЛР(Л1)	83.11	84.77	82.78	82.45	81.79	85.43	87.42	82.45	83.44	82.12	/	83.58
ЛР(Л2)	84.44	85.76	84.77	87.75	86.75	86.09	86.42	87.42	90.09	87.42	/	86.69
ЛР(О)	85.76	87.09	87.42	83.77	83.44	88.41	82.45	85.76	86.09	87.42	1.039	85.76

Резултати друге фазе за *TF*:

	I	II	III	IV	V	VI	VII	VIII	IX	X	C	AVG
МНБК	78.48	84.44	81.13	82.12	77.48	79.80	79.14	80.79	81.79	81.46	/	80.66
БНБК	81.13	78.81	74.17	76.49	79.14	79.80	79.47	75.50	77.15	79.40	/	78.15
ПВ(Л1)	88.74	84.77	85.76	88.08	86.42	85.43	85.10	87.09	85.10	88.08	/	86.46
ПВ(Л2)	91.06	87.75	88.74	87.42	88.41	88.41	90.40	87.42	89.74	87.09	/	88.64
ПВ(О)	91.39	88.74	89.74	89.74	88.41	83.77	89.40	86.75	87.09	90.40	1.478	88.54
ЛР(Л1)	83.44	82.78	86.42	81.13	84.11	87.42	80.79	85.10	82.12	85.43	/	83.87
ЛР(Л2)	85.43	81.79	79.47	82.45	84.77	82.78	85.10	81.13	85.10	84.44	/	83.25
ЛР(О)	84.77	83.44	82.12	83.77	84.77	86.75	89.07	87.09	84.44	86.09	1.751	85.23

Резултати друге фазе за *IDF*:

	I	II	III	IV	V	VI	VII	VIII	IX	X	C	AVG
МНБК	84.77	85.10	83.77	82.78	85.76	86.42	84.44	84.44	85.43	86.42	/	84.93
БНБК	72.52	76.49	78.15	77.81	76.49	74.83	76.49	77.48	79.14	78.15	/	76.75
ПВ(Л1)	85.43	86.74	87.09	87.75	87.75	89.07	85.10	85.43	86.75	87.09	/	86.82
ПВ(Л2)	88.08	89.40	87.09	87.42	89.74	85.43	87.75	86.09	85.43	86.75	/	87.32
ПВ(О)	89.07	89.74	88.41	85.10	89.07	87.75	89.74.	83.77	88.74	88.74	1.497	88.01
ЛР(Л1)	87.42	90.07	87.09	87.42	89.07	85.76	86.09	89.40	86.09	88.08	/	87.65
ЛР(Л2)	85.43	81.79	80.79	82.78	83.11	80.79	85.43	81.79	86.09	82.45	/	83.05
ЛР(О)	89.74	89.07	88.08	88.74	87.09	87.09	89.40	86.75	89.74	85.43	1.462	88.11

Резултати друге фазе за *TF-IDF*:

	I	II	III	IV	V	VI	VII	VIII	IX	X	C	AVG
МНБК	81.46	84.44	79.47	84.77	81.79	84.11	83.77	79.80	80.46	81.13	/	82.12
БНБК	82.45	79.14	78.15	76.82	80.46	77.81	77.15	77.48	75.83	75.17	/	78.05
ПВ(Л1)	85.43	87.09	87.09	84.77	88.41	86.42	87.09	85.10	86.09	87.42	/	86.49
ПВ(Л2)	88.41	88.74	88.41	87.42	88.41	87.75	90.73	89.74	87.75	86.09	/	88.34
ПВ(О)	87.42	86.09	89.07	86.75	85.43	84.75	89.07	81.06	90.07	89.74	1.477	88.25
ЛР(Л1)	86.42	84.44	82.78	81.79	83.77	84.44	83.11	83.11	81.46	84.44	/	83.58
ЛР(Л2)	85.10	85.10	83.77	86.42	86.42	81.79	84.44	82.78	82.45	82.78	/	84.11
ЛР(О)	86.75	83.11	83.11	86.09	85.10	86.42	85.10	86.09	84.11	87.42	1.706	85.33

Код *TF* у комбинацији са Бајесовим класификатором, даје слабије перформансе у односу на Бернулијев класификатор, што нисмо имали као случај у ранијим тестирањима са препроцесирањем.

Што се тиче логичке регресије и потпорног вектора имамо да су перформансе сада стаблније и са мањом варијацијом.

Код *IDF* у комбинацији са Бајесовим класификатором сада немамо горенаведени феномен, а притом и побољшање од ~5% у односу на претходна тестирања. Резултати Бернулијевог класификатора су приближно слични као и код *TF* тестирања. За нијансу процентуално боље перформансе имамо у случају логичке регресије и потпорног вектора.

Код *TF-IDF* опет имамо случај да Бајесов класификатор има слабије перформансе у глобалу у односу на Бернулијев класификатор. Мали пад перформанси увиђамо и у случају логичке регресије и потпорног вектора.

У сва три случаја за другу фазу примећујемо пад од ~4%. Што представља већи пад него што смо имали у било ком од претходних типова препроцесирања.

4.1.4. *Stemmer and stop words*

Штемовање речи је метода која се користи за нормализацију речи, односно да се приликом процесирања не гледа реч у целини него корен саме речи. У ову сврху је коришћен *Портеров штемер*.

Стоп речи подразумевају листу речи које у процесирању текста немају тежину. То су обично везници, узречице... Зато се оне уклањају како не би уносили нечистоће приликом даљег процесирања.

Резултати прве фазе:

	I	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**C**	**AVG**
МНБК	82.78	83.44	86.09	85.43	84.77	87.75	86.75	87.42	84.44	85.10	/	85.40
БНБК	82.45	81.46	81.13	89.47	78.15	80.46	80.13	82.12	77.15	82.45	/	80.50
ПВ(Л1)	88.08	89.40	88.74	86.42	86.75	89.74	87.75	87.09	88.08	91.06	/	88.31
ПВ(Л2)	86.75	88.74	90.73	89.07	88.74	88.41	88.74	90.40	88.41	89.07	/	88.91
ПВ(О)	90.07	90.07	86.42	86.09	89.07	86.75	91.72	90.07	89.40	86.42	0.942	88.61
ЛР(Л1)	85.76	85.76	87.09	87.09	88.08	88.08	83.77	87.09	86.42	88.41	/	86.75
ЛР(Л2)	89.75	86.09	84.44	86.75	88.74	88.74	89.07	88.74	85.43	84.77	/	86.95
ЛР(О)	87.75	88.08	85.76	85.10	87.42	86.42	89.74	88.74	87.75	87.75	0.944	87.45

Резултати друге фазе:

	I	II	III	IV	V	VI	VII	VIII	IX	X	C	AVG
МНБК	84.77	81.79	85.76	82.45	84.44	79.47	83.77	80.79	85.10	83.44	/	83.18
БНБК	75.50	78.81	81.46	78.48	75.17	79.47	77.48	76.49	78.81	79.80	/	78.15
ПВ(Л1)	87.42	86.75	87.75	85.76	87.42	85.76	86.75	86.42	86.42	88.41	/	86.89
ПВ(Л2)	88.08	88.74	84.11	81.13	86.75	89.40	88.74	87.07	84.77	85.76	/	86.46
ПВ(О)	85.10	85.43	85.10	84.44	85.10	86.75	88.08	84.44	85.76	88.08	1.520	85.83
ЛР(Л1)	85.43	85.10	84.11	84.11	86.75	84.44	86.09	83.44	85.43	85.43	/	85.03
ЛР(Л2)	84.77	84.44	8609	84.11	87.75	82.78	87.75	88.41	80.79	85.43	/	85.23
ЛР(О)	86.42	87.09	87.42	87.42	85.43	84.77	85.43	89.04	85.76	86.42	1.519	86.52

Код Бајесовог класификатора да је проценат тачности приближан тачности коју имамо код истог класификатора, али са *IDF* препроцесирањем, што је до сада био најбољи резултат који смо успевали да добијемо. Док пад перформанси у односу на претходни скуп препроцесирања имамо у случају Бернулијевог класификатора.

Што се тиче логичке регресије и потпорног вектора такође имамо пад у односу на претходни скуп препроцесирања. Проценат тачности, кумулативно је нижи за приближно ~5%.

У другој фази тачност је идентична тачностима из прве фазе.

4.1.5. *Frequency word filtering*

Ова метода подразумева срачунавање фреквенције појаве сваке речи. Затим се врши филтрирање речи на основу тога колика има је појава у односу на средњу вредност.

Резултати прве фазе:

	I	II	III	IV	V	VI	VII	VIII	IX	X	C	AVG
МНБК	84.77	84.44	81.13	80.13	84.77	80.13	84.11	84.77	82.78	83.77	/	83.08
БНБК	76.82	74.50	78.15	82.78	79.14	83.11	82.12	80.46	81.79	77.15	/	79.60
ПВ(Л1)	85.10	90.07	90.07	82.72	85.10	86.42	88.08	86.09	87.09	91.39	/	88.21
ПВ(Л2)	88.74	89.07	89.40	90.73	90.07	89.74	90.07	88.74	90.07	90.07	/	89.67
ПВ(О)	88.08	88.74	90.73	87.75	90.40	90.40	91.39	87.42	87.06	88.08	0.916	89.40
ЛР(Л1)	83.44	82.78	85.76	83.77	82.45	84.77	82.78	85.10	83.44	84.11	/	83.84
ЛР(Л2)	87.75	87.42	86.42	87.42	86.42	87.09	84.11	84.44	86.09	86.42	/	86.36
ЛР(О)	84.77	87.09	87.75	86.09	87.09	86.75	84.11	86.09	89.07	87.75	1.048	86.66

Резултати друге фазе:

	I	II	III	IV	V	VI	VII	VIII	IX	X	C	AVG
МНБК	81.79	81.13	80.13	79.14	82.45	80.79	81.13	80.46	79.80	82.45	/	80.93
БНБК	76.82	78.15	78.81	75.50	76.49	79.14	79.47	76.49	76.49	74.50	/	77.19
ПВ(Л1)	87.42	87.42	87.75	82.45	87.09	88.08	84.77	84.11	84.44	86.75	/	86.03
ПВ(Л2)	86.42	87.42	87.75	89.74	87.42	89.74	86.75	90.40	88.41	87.42	/	88.15
ПВ(О)	86.42	88.41	86.75	87.09	87.42	90.07	87.09	88.08	84.44	87.75	1.475	87.35
ЛР(Л1)	84.77	85.76	82.78	81.46	85.43	82.78	85.76	85.10	86.09	83.77	/	84.37
ЛР(Л2)	84.77	84.44	83.44	82.11	85.43	86.09	81.13	81.46	83.11	83.11	/	83.51
ЛР(О)	85.10	85.43	86.09	84.44	85.10	84.77	85.43	86.75	83.77	82.78	1.703	84.97

У првој фази имамо тачност од 81 до 85 посто за Бајесов класификатор, и 75 до 84 посто за Бернулијев класификатор. У другој фази за ова два класификатора, сумарно, имамо пад од 4 посто за оба класификатора.

Метода потпорних вектора даје бољу тачност, па се она креће између 88 и 92 посто, док је логичка регресија слабија за ~5%.

Овај тип класификације, генерално гледано, је бољи у односу на методе са биграмима и триграмима, али не доноси значајан бољитак у односу на иницијално тестирање без препроцесирања.

4.1.6. *Bigrams and trigrams*

Препроцесирање је слично као код препроцесирања где се не врши обрада, већ се само поделе на унираме (први случај), само што се сада овде речи не деле на унираме, већ на биграме (две речи) и триграме (три речи).

Резултати прве фазе за биграме:

	I	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**C**	**AVG**
МНБК	79.14	83.77	81.79	82.12	76.49	79.47	83.11	79.47	77.48	84.77	/	80.76
БНБК	75.50	71.85	73.84	74.50	75.50	73.84	74.83	76.16	73.51	74.50	/	74.40
ПВ(Л1)	81.46	89.74	87.09	86.09	82.45	88.74	85.43	85.43	88.41	85.76	/	86.06
ПВ(Л2)	87.75	87.75	86.75	87.09	89.42	85.10	88.41	88.08	87.42	87.42	/	87.22
ПВ(О)	88.74	85.76	89.07	87.42	87.42	88.74	86.09	87.42	86.42	85.43	0.966	87.25
ЛР(Л1)	84.44	84.44	84.44	85.10	83.77	84.77	84.11	87.09	86.05	83.44	/	84.83
ЛР(Л2)	86.09	84.44	83.77	86.09	84.11	81.79	83.44	88.08	84.44	87.09	/	84.93
ЛР(О)	87.75	84.44	85.76	86.42	84.11	8.42	83.44	89.07	87.42	87.09	1.002	86.23

Резултати прве фазе за триграме:

	I	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**C**	**AVG**
МНБК	70.20	73.18	74.50	75.17	69.87	73.51	74.83	68.21	71.85	75.83	/	72.72
БНБК	72.19	69.54	70.53	68.87	66.56	71.85	67.88	69.54	70.53	71.52	/	69.90
ПВ(Л1)	86.42	86.42	82.78	84.11	87.42	86.75	85.76	85.76	85.10	87.42	/	85.79
ПВ(Л2)	84.77	86.09	83.11	84.44	84.44	85.76	86.09	85.10	84.77	83.77	/	84.83
ПВ(О)	87.75	84.44	85.76	84.77	87.42	85.43	85.43	84.44	84.44	87.42	1.006	85.73
ЛР(Л1)	80.79	81.46	75.83	80.46	81.13	77.15	79.80	82.12	78.15	81.13	/	79.80
ЛР(Л2)	83.44	81.79	79.14	80.46	82.78	82.45	83.44	82.12	81.46	84.11	/	82.12
ЛР(О)	83.77	82.45	79.80	83.44	79.80	83.77	81.13	81.79	82.12	82.78	1.056	82.09

Резултати друге фазе за биграме:

	I	II	III	IV	V	VI	VII	VIII	IX	X	C	AVG
МНБК	79.47	81.13	82.78	81.46	82.12	77.15	79.80	83.11	81.79	81.46	/	81.03
БНБК	71.52	71.19	71.52	70.86	73.84	74.17	71.19	71.85	73.51	72.85	/	72.25
ПВ(Л1)	84.44	86.42	87.75	85.43	82.78	85.10	85.76	85.10	83.44	85.76	/	85.20
ПВ(Л2)	88.08	87.09	88.74	86.42	86.75	83.77	84.44	82.78	82.10	87.42	/	86.06
ПВ(О)	82.12	84.77	87.42	84.77	88.74	88.41	88.41	86.75	86.09	88.41	1.580	86.59
ЛР(Л1)	82.12	82.45	80.79	82.78	82.12	82.45	79.14	82.78	85.76	83.44	/	82.38
ЛР(Л2)	83.44	83.44	83.44	80.79	84.77	84.44	80.79	79.14	86.75	83.77	/	83.08
ЛР(О)	82.78	84.11	84.70	87.42	86.42	86.09	82.12	83.42	83.44	88.74	1.660	85.23

Резултати друге фазе за триграме:

	I	II	III	IV	V	VI	VII	VIII	IX	X	C	AVG
МНБК	72.85	71.52	72.19	70.86	70.20	72.52	70.53	70.53	77.15	74.83	/	72.32
БНБК	68.87	66.89	67.55	64.09	64.24	67.55	71.19	67.55	66.56	66.23	/	67.15
ПВ(Л1)	84.44	84.44	81.46	81.79	82.79	80.46	83.77	84.11	83.44	83.77	/	83.05
ПВ(Л2)	85.76	83.77	84.77	84.11	82.45	82.45	79.80	81.13	84.77	82.78	/	83.18
ПВ(О)	85.76	81.46	80.79	84.44	84.77	85.10	82.10	84.44	80.79	83.44	1.632	83.31
ЛР(Л1)	77.15	79.14	79.14	81.13	76.82	81.79	80.79	76.49	76.49	79.14	/	78.81
ЛР(Л2)	82.12	79.80	79.80	80.79	81.46	79.80	83.11	79.47	79.80	80.79	/	80.70
ЛР(О)	79.80	81.46	77.15	81.79	80.79	80.46	84.77	79.80	84.11	85.76	1.747	81.59

Као што се из горенаведеног прилога види биграмска, а поготово триграмска техника је давала најслабије резултате од свих типова препроцесирања.

Код биграмског препроцесирања и Бајесовог класификатора, у обе фазе, добијају се резултати од 79 до 84 посто. Бернулијев класификатор, као и у другим случајевима, даје још лошије пероформансе, које се крећи између 72 и 76 посто.

Пад у случају логичке регресије и потпорног вектора у просеку у односу на друге типове препроцесирања је мањи него у случају прва два класификатора, па у обе фазе износи 88 и 91 процената тачности.

Код триграмског препроцесирања и Бајесовог класификатора, у обе фазе, добијају се резултати од 68 до 74 посто. Бернулијев класификатор, као и у другим случајевима, даје још лошије пероформансе, које се крећи између 70 и 73 посто.

Пад у случају логичке регресије и потпорног вектора у просеку у односу на друге типове препроцесирања је мањи него у случају прва два класификатора, па у обе фазе износи 85 и 88 процената тачности.

4.2. Тренирање класификатора

У овом одељку су детаљно описани сви класификатори који су коришћени у обради података.

Приликом обучавања свих класификатора се користи *10 слојна стратификована унакрсна валидација*.

Која на примеру Бајесовог класификатора изгледа:

```
sss = StratifiedShuffleSplit(n_splits=10, test_size=0.1)
index = 1
for train_index, test_index in sss.split(x, data_frame['Type']):
    x_train, x_test = x[train_index], x[test_index]
    y_train, y_test = data_frame['Type'][train_index], data_frame['Type'][test_index]

    mnb = MultinomialNB()
    mnb.fit(x_train, y_train)
    score = mnb.score(x_test, y_test)

    print("Score {}.: {:.3f}".format(index, score), end=" ")
    if index == 5:
        print()
    index += 1
```

Код логичке регресије и методе потпорних вектора користи се *угњеждена унакрсна валидација за оптимизацију хиперпараметара*. На примеру логичке регресије то изгледа:

```
def optimize_c_parameter(data_frame):
    models_param = {
        'max_iter': [10000],
        'C': [1]
    }
```

```
    nested_cv_search = NestedCV(model=LogisticRegression(), params_grid=models_param,
                                outer_kfolds=5, inner_kfolds=5,
                                cv_options={'sqrt_of_score': True,
'randomized_search_iter': 30})

    x = preprocessing.get_data_set()
    nested_cv_search.fit(x, data_frame['Type'])

    optimized_c_value = np.mean(nested_cv_search.outer_scores)
    print("Optimized C: {:.3f}".format(optimized_c_value))
    compare_regularisation_functions(data_frame, 'l2', optimized_c_value)
```

4.2.1. Мултиномијални Бајесов наивни класификатор

Мултиномијални Бајесов наивни класификатор преставља представља класификатор који се често користи приликом класификације текстова. Назив наивни потиче од тога што се не узима у обзир постојање завиности између речи у тексту. Већ се све речи тумаче као одвојене, дискретне, целине.

Као такав, Бајесов класификатор показује стабилно добре перформансе код технике коришћене при решавању проблема, односно *bag-of-words* технике. Где нисмо узимали семантику речи већ смо их посматрали као засебне целине.

Код већине тест скупова, Бајесов класификатор је показивао боље перофмансе од Бернулијевог класификатора. Гледано као целину, проценат боље тачности се кретао од 5 до 10 посто.

Најнижи проценат код униграм техника имамо у случају тестирања без препроцесирања. Док уопштено гледано најнижи проценат имамо код препроцесирања са триргамима.

Највиши проценат имамо код *IDF* и Портеровог штемера, где смо у просеку имали око 87% процената тачности. Што је и више него задовољавајуће ако узмемо у обзир једноставност и брзину технике.

У сврху реализације ово задатка се користи следећи пакет:

```
from sklearn.naive_bayes import MultinomialNB
```

Тренирање се врши на следећи начин:

```
mnb = MultinomialNB()
mnb.fit(x_train, y_train)
```

Процена квалитета класификатора се извршава на следећи начин:

```
score = mnb.score(x_test, y_test)
```

4.2.2. Бернулијев наивни класификатор

Бернулијев наивни класификатор функционише по истом принципу као и горенаведени Бајесова наивни класификатор, са малом оптимизацијом у специјалним случајевима.

У скупу података који су коришћени за тестирање, Бернулијев наивни класификатор је изузев пар случајева показивао лошије перформансе у односу на све остале класификаторе. Тако да је резултатски гледано најнеповољнији избор за класификатор.

Ипак иако резултатски најгори, за поједине комбинације препроцесирања, класификатор је давао сасвим добре перформансе од приближно 80%.

Најнижи проценат код униграм техника имамо у случају тестирања без препроцесирања. Док уопштено гледано најнижи проценат имамо код препроцесирања са триргамима.

У сврху реализације ово задатка се користи следећи пакет:

```
from sklearn.naive_bayes import BernoulliNB
```

Тренирање се врши на следећи начин:

```
bnb = BernoulliNB()
bnb.fit(x_train, y_train)
```

Процена квалитета класификатора се извршава на следећи начин:

```
score = bnb.score(x_test, y_test)
```

4.2.3. Логичка регресија

Моделом логистичке регресије описујемо везу између предиктора који могу бити непрекидни, бинарни, категорички, и категоричке зависне променљиве. На пример зависна променљива може бити бинарна - на основу неких предиктора предвиђамо да ли ће се нешто десити или не. Оцењујемо заправо вероватноће припадања свакој категорији за дат скуп предиктора.

Логичка регресија је, као што се види из претходних примера, за све типове препроцесирања давала, увек међу најбољим резултатима.

Тестирање је вршено са Л1 и Л2 функцијама грешке, сумарно гледано, Л2 фунцкија се показала као генерално боља. Такође, код логичке регресије смо имали и оптимизацију

хиперпараметра *C*. Оптимизација је утицала да перформансе буду стабилније, тј. са мањом осцилацијом и у просеку за нијансу боље него без оптимизација.

Као и код осталих класификатора, најлошије перформансе су добијене код триграмског препроцесирање. Док су најбољи и са најмањом осцилацијом показали код *IDF* техинке.

У сврху реализације ово задатка се користи следећи пакет:

```
from sklearn.linear_model import LogisticRegression
```

Тренирање се врши на следећи начин:

```
lr = LogisticRegression(penalty=rf, C=c, solver=solver, max_iter=10000)
lr.fit(x_train, y_train)
```

Процена квалитета класификатора се извршава на следећи начин:

```
score = lr.score(x_test, y_test)
```

4.2.4. Метод потпорних вектора без кренела

Метод потпорних вектора без кренела представља основну варијатну алгоритма. Без кернела, подразумева се да као фунцкију раздвајања узимамо линерну функцију. Кернел може бити готово било која функција.

Представља непробабалистички класификатор јер је излаз модела само класификациона одлика, а не и вероватно припарности класи.

Метода потпорних вектора је, као што се види из претходних примера, је уз логичку регресију давао најбоље резултате.

Тестирање је вршено са Л1 и Л2 функцијама грешке, сумарно гледано, Л2 фунцкија се показала као генерално боља. Такође, код методе потпорних вектора смо имали и оптимизацију хиперпараметра *C*. Оптимизација је утицала да перформансе буду стабилније, тј. са мањом осцилацијом и у просеку боље него без оптимизација.

Као и код осталих класификатора, најлошије перформансе су добијене код триграмског препроцесирање. Док су најбољи резултати, и са најмањом осцилацијом, добијени код *IDF* и *TF-IDF*. Слично као што смо добили код логичке регресије.

У сврху реализације ово задатка се користи следећи пакет:

```
from sklearn.svm import LinearSVC
```

Тренирање се врши на следећи начин:

```
svc = LinearSVC(penalty=rf, C=c, dual=rf == 'l2', max_iter=15000)
svc.fit(x_train, y_train)
```

Процена квалитета класификатора се извршава на следећи начин:

```
score = svc.score(x_test, y_test)
```