LATVIJAS UNIVERSITĀTE

DATORIKAS FAKULTĀTE

**SENTIMENTU ANALĪZE TEKSTIEM KRIEVU VALODĀ**

KVALIFIKĀCIJAS DARBS

Autors: Pēteris Girgensons

Studenta apl. Nr.: pg13013

Darba vadītājs: M.vad.zin. Pēteris Paikens

RĪGA 2015

# Anotācija

Šī programma sniedz portāla “tvnet.lv” moderatoriem iespēju pamanīt agresīvi un vardarbīgi noskaņotus komentārus.

Sistēma sastāv no 2 moduļiem: servera daļas un pārlūkprogrammas daļas.

Atslēgvārdi: noskaņojuma analīze, mašīnmācīšanās, tekstu apstrāde, naivais Baiesa klasifikators, krievu valoda

# Abstract

This program gives the moderators of “rus.tvnet.lv” portal a possibility to detect aggressive and violent comments.

The sysyem consists of two modules: server side and browser side.

KEYWORDS: sentiment analysis, machine learning, text evaluation

**Satura rādītājs**

[Anotācija 2](#_Toc419803389)

[Abstract 3](#_Toc419803390)

[DEFINĪCIJAS UN SAĪSINĀJUMI 6](#_Toc419803391)

[Ievads 7](#_Toc419803392)

[Sistēmas mērķis 7](#_Toc419803393)

[Darba mērķi un uzdevumi 7](#_Toc419803394)

[Darba struktūra 7](#_Toc419803395)

[PROGRAMMATŪRAS PRASĪBU SPECIFIKĀCIJA 8](#_Toc419803396)

[Ievads 8](#_Toc419803397)

[Nolūks 8](#_Toc419803398)

[Darbības sfēra 8](#_Toc419803399)

[Definīcijas 8](#_Toc419803400)

[Saistība ar citiem dokumentiem 8](#_Toc419803401)

[Vispārējs apraksts 8](#_Toc419803402)

[Lietotāji 8](#_Toc419803403)

[Lietotājstāsti 9](#_Toc419803404)

[Lietotāja raksturiezīmes 12](#_Toc419803405)

[Vispārējie ierobežojumi 12](#_Toc419803406)

[Funkcionālās prasības 12](#_Toc419803407)

[Ārējā saskarne 18](#_Toc419803408)

[Lietotāja saskarne 18](#_Toc419803409)

[Sakaru saskarne 18](#_Toc419803410)

[Nefunkcionālās prasības 18](#_Toc419803411)

[Izmantojamība 18](#_Toc419803412)

[Drošība 18](#_Toc419803413)

[Pārnesamība 18](#_Toc419803414)

[PROGRAMMATŪRAS PROJEKTĒJUMA APRAKSTS 19](#_Toc419803415)

[Nolūks 19](#_Toc419803416)

[Saistība ar citiem dokumentiem 19](#_Toc419803417)

[Definīcijas 19](#_Toc419803418)

[KVALITĀTES NODROŠINĀŠANA 25](#_Toc419803419)

[Konfigurācijas pārvaldība 26](#_Toc419803420)

[TESTĒŠANAS DOKUMENTĀCIJA 27](#_Toc419803421)

[PROGRAMMATŪRAS PROJEKTA ORGANIZĀCIJA 31](#_Toc419803422)

[DARBIETILPĪBAS NOVĒRTĒJUMS 32](#_Toc419803423)

[IZMANTOTĀ LITERATŪRA 33](#_Toc419803424)

[Pielikumi 34](#_Toc419803425)

[Secinājumi 37](#_Toc419803426)

# DEFINĪCIJAS UN SAĪSINĀJUMI

Lietotājs

Administrators

PHP **-** atklātā pirmkoda skriptu valoda, kura sākotnēji bija paredzēta servera puses lietojumosdinamiska tīmekļa lapu ģenerēšanai.

CSS - stila lapas kaskadēšana (angļu val. – *cascading style sheets*).

JavaScript - skriptu valoda, kas balstīta uz prototipu koncepta. Pamatā tiek izmantota tīmekļa vietņu skriptu izpildīšanai uz klienta datora.

Jquery - JavaScript valodas papildus bibliotēka.

Ajax – (angļu val. - asinhroni JavaScript un XML)

HTML – (angļu val. - HyperText Markup Language) hiperteksta iezīmēšanas valoda; valoda pārlūkprogrammā attēlojamas informācijas glabāšanai

SQL – (angļu val. - [Structured Query Language](http://termini.lza.lv/term.php?term=Structured%20Query%20Language&list=Structured%20Query%20Language&lang=EN)) [strukturēta vaicājumvaloda](http://termini.lza.lv/term.php?term=struktur%C4%93ta%20vaic%C4%81jumvaloda&list=struktur%C4%93ta%20vaic%C4%81jumvaloda&lang=LV); vaicājumu valoda, kas tiek izmantota datubāžu datu izgūšanai, dzēšanai un atjaunošanai

UTF-8

BootStrap - mājaslapu un sistēmu dizaina izstrādes ietvars

PPS - programmatūras prasību specifikācija

PPA - programmatūras projektējuma apraksts.

Git -

GitHub

Repozitorijs – vieta, kur tiek glabāti dati.

XAMPP - programmatūru paka, kas sastāv tikai un vienīgi no brīvām atvērtā pirmkoda programmām un ir paredzēta lielu, dinamisku tīmekļa lapu veidošanai.

Apache

Treniņdati

# Ievads

Mēs dzīvojām digitālajā gadsimtā, kad jebkāda veida dati ir publiski pieejami internetā.

Milzīgā datu apjoma dēļ ir ļoti grūti veikt jebkādus secinājumus no manuāli apstrādātās informācijas. Bet jebkura iespēja pat daļēji automatizēt tekstu apstrādi var palīdzēt pareģot vēlēšanu rezultātus, izprast resursu un akciju tirgus likumsakarības vai izsekot pircēju vēlēšanos, kas ir ļoti svarīgi mārketinga speciālistiem.

Tekstu analīze var arī palīdzēt filtrēt bīstamu informāciju, kā politiskās provokācijas un aicinājumi uz vardarbību, kā arī cilvēku pašcieņu pazemojošus komentārus.

Sentimenta jeb noskaņojuma analīze kļūst par svarīgu tematu zinātniskos pētījumos un sociālo mediju apstrādē. Tas ir tekstu apstrādes paveids, kas tiek izmantots, lai noteiktu autora attieksmi pret noteikto tēmu. Kaut gan noskaņojuma analīze reti sasniedz izcilu precizitātes līmeni, tās lietderība ir ļoti augsta.

Kvalifikācijas darbā aprakstītas programmas funkcionalitāte un prasības.

Programmatūras izveidošanai tika izmantotas PHP un JavaScript valodas.

## Sistēmas mērķis

Šīs programmas mērķis ir atvieglot „rus.tvnet.lv” portāla moderatoru dzīvi, ļaujot viņiem ātri pamanīt naidīgus un pazemojošus komentārus

## Darba mērķi un uzdevumi

Kvalifikācijas darba mērķis ir atspoguļot autora Latvijas Universitātē un praksē

apgūtās zināšanas.

## Darba struktūra

Kvalifikācijas darbs iekļauj programmatūras prasību specifikāciju, programmatūras

projektējuma aprakstu, testēšanas dokumentāciju, kvalitātes nodrošināšanu, konfigurāciju pārvaldību, programmatūras projekta organizāciju, darbietilpības novērtējumu un būtiskākos pirmkoda fragmentus.

# PROGRAMMATŪRAS PRASĪBU SPECIFIKĀCIJA

## Ievads

1. **Ievads** - īsi aprakstīts šī dokumenta un programmas mērķis un būtība.

2. **Vispārējs apraksts** - īsumā aprakstītas programmas funkcijas, lietotāji, ierobežojumi un atkarības.

3. **Funkcionālās prasības** - visas programmas prasības, saskaņā ar kurām programmas tālāk tiks projektēta, izstrādāta, testēta, akceptēta un ieviesta ekspluatācijā.

4. **Nefunkcionālās prasības** - programmas nefunkcionālās prasības, saskaņā ar kurām programmas tiks projektēta, izstrādāta, testēta, akceptēta un ieviesta ekspluatācijā, kā arī definē programmas pasūtītāja atbildību.

5. **Ārējās saskarnes prasības** - sadaļa ietver saskarnes prasības un ierobežojumus.

### Nolūks

Programmatūras prasību specifikācija, jeb PPS, tika izstrādāta, lai pasūtītājs varētu precīzi definēt nepieciešamās sistēmas prasības.

### Darbības sfēra

Programma ir paredzēta lielu portālu moderatoriem, kuri vēlas noteikt, kuri viņu portāla komentāri ir agresīvi, pazemo cilvēkus vai aicina uz vardarbību.

### Definīcijas

Definīcijas un saīsinājumi sīkāk aprakstīti sadaļā *„Definīcijas un saīsinājumi”.*

### Saistība ar citiem dokumentiem

Programmatūras prasību specifikācijas noformēšanā ievērotas standarta LVS 68:1996 “Programmatūras prasību specifikācijas ceļvedis” prasības.

### Vispārējs apraksts

Programma ir atkarīga tikai no portāla, kurā atrodas komentāri.

### Lietotāji

Portāla moderators

Sistēmas administrators

## Lietotājstāsti

|  |  |
| --- | --- |
| Nosaukums | Reģistrācija |
| Stāsts | Es, kā sistēmas lietotājs, vēlos reģistrēties sistēmā, lai varētu pilnvērtīgi sākt lietot tās funkcionalitāti. |
| Detaļas | Lai piereģistrētos, lietotājam ir jāievada savs lietotājvārds un parole. |
| Akceptēšanas kritēriji | Pēc datu ievades un pogas „Piereģistrēties”  nospiešanas, tiek pārbaudīts vai ievadītie dati ir korekti. Pie nekorektu datu ievades tiek izvadīti informatīvi paziņojumi un lietotājs netiek izveidots. |

|  |  |
| --- | --- |
| Nosaukums | Autorizācija |
| Stāsts | Es, kā piereģistrēts lietotājs, vēlos autorizēties sistēmā, lai varētu izmantot visas tās iespējas. |
| Detaļas | Lai autorizētos, lietotājam ir jāievada lietotājvārds un parole. |
| Akceptēšanas kritēriji | Ja ievadīti dati ir korekti un ja šāds lietotājs patiešām eksistē, tiek izveidotā jauna sesija un unikāls sesijas identifikators. |

|  |  |
| --- | --- |
| Nosaukums | „Chrome spraudņa izmantošana” |
| Stāsts | Es, kā autorizēts lietotājs, atvērot „rus.tvnet.lv” rakstu, gribu uzreiz redzēt, kuri komentāri ir negatīvi. |
| Detaļas | „Chrome spraudnis” parsē komentārus un pārbauda katru komentāru, vai tas ir negatīvs vai neitrāls. |
| Akceptēšanas kritēriji | Veiksmīgas izpildes gadījumā neitrāli komentāri pārlūkprogrammā tiek iekrāsoti zaļajā krāsā, bet negatīvi komentāri tiek iekrāsoti sarkanajā krāsā. |

|  |  |
| --- | --- |
| Nosaukums | Komentāra klases maiņa |
| Stāsts | Es, kā autorizēts lietotājs, vēlos manuāli mainīt komentāra klasi gadījumā, ja es nepiekrītu programmas darbībai. |
| Detaļas | Zem katra iekrāsota komentāra tiek ievietota poga ar tekstu „Tomēr ir negatīvs”, ja programmas rezultāts ir neitrāls, un „Tomēr ir neitrāls”, ja programmas rezultāts ir negatīvs. |
| Akceptēšanas kritēriji | Ja dotā komentāra klase netika mainīta, tiek izvadīts paziņojums. |

|  |  |
| --- | --- |
| Nosaukums | Treniņdatu papildināšana |
| Stāsts | Es, kā autorizēts lietotājs, vēlos pievienot jaunos treniņdatus, lai uzlabotu klasifikatora precizitāti. |
| Detaļas | Tiek ievadīts jaunā komentāra teksts un tiek izvēlēta tā klase. |
| Akceptēšanas kritēriji | Ja jauns komentārs tika pievienots teksta failam, tiek izvadīts paziņojums par veiksmīgi izpildītu darbību. |

|  |  |
| --- | --- |
| Nosaukums | Treniņdatu dzēšana |
| Stāsts | Es, kā autorizēts lietotājs, vēlos nodzēst noteiktus treniņdatus, lai uzlabotu klasifikatora precizitāti. |
| Detaļas | No visu dotās kopas komentāru saraksta tiek izvēlēti dzēšamie komentāri un tiek apstiprināts brīdinājums ar pazinojumu „Vai tiešām vēlaties dzēst šos komentārus?” |
| Akceptēšanas kritēriji | Ja izvēlētie komentāri tika nodzēsti no teksta faila, tiek izvadīts paziņojums par veiksmīgi izpildītu darbību. |

|  |  |
| --- | --- |
| Nosaukums | Populārāko rakstu analīze |
| Stāsts | Es, kā autorizēts lietotājs, vēlos atrast visus negatīvus komentārus no šobrīd populārākiem „rus.tvnet.lv” rakstiem. |
| Detaļas | Tiek izgūti komentāri no ”rus.tvnet.lv” populārākajiem rakstiem. Ar klasifikatora palīdzību tiek noteiktās to klases. |
| Akceptēšanas kritēriji | Veiksmīgas izpildes gadījumā tabulas veidā tiek atlasīti negatīvi komentāri no šobrīd populārākajiem rakstiem. |

|  |  |
| --- | --- |
| Nosaukums | Šķērsvalidācija |
| Stāsts | Es, kā autorizēts lietotājs, vēlos veikt programmas šķērsvalidāciju, lai pārbaudītu tās precizitāti. |
| Detaļas | Komentāri no abām datu kopām tiek samaisīti kopā. Visi komentāri tiek sadalīti desmit daļās. Viena daļa tiek paņemta, kā testu daļa, bet pārējās deviņas, kā treniņdati. Klasifikators tiek uztrenēts uz pārējām deviņām daļām un pēc tam mēģina noteikt katra testa komentāra klasi. Process tiek atkārtots, mainot testa datus. |
| Akceptēšanas kritēriji | Veiksmīgas izpildes gadījumā tabulas veidā tiek izvadīti šķērsvalidācijas rezultāti. |

### Lietotāja raksturiezīmes

Jebkuram sistēmas lietotājam ir jābūt pamatiemaņām darbā ar datoru, izvēlēto interneta pārlūkprogrammu un elementārai izpratnei par tīmekļa vietnēm. Citas specifiskas raksturiezīmes nav nepieciešamas, lai lietotājs spētu pilnvērtīgi izmantot šo programmu.

### Vispārējie ierobežojumi

Programma nav daļa no citas programmatūras, taču, tā kā sistēma ir realizēta

tīmeklī, funkcionalitāti var ierobežot interneta pārlūkprogramma, interneta savienojums un

servera noslodze.

### 

### Funkcionālās prasības

|  |
| --- |
| Autorizācija |
| Lietotājs var autorizēties sistēmā |
| Ievade: |
| Lietotājvārds  Parole |
| Apstrāde: |
| Pārbauda, vai lietotājs ar šādiem datiem ir piereģistrēts  Pārbauda, vai ievadīti dati ir korekti  Pārbauda, vai obligāti aizpildāmi lauki ir aizpildīti  Kļūdas gadījumā izvada kļūdas paziņojumu |
| Izvade: |
| Paziņojums par veiksmīgu autorizāciju |

|  |
| --- |
| Reģistrācija |
| Lietotājs var piereģistrēties, lai varētu pilnvērtīgi lietot sistēmu |
| Ievade: |
| Lietotājvārds  Parole |
| Apstrāde: |
| Saglabā lietotāja datus datu bāzē  Pārbauda, vai ievadīti dati ir korekti  Pārbauda, vai obligāti aizpildāmi lauki ir aizpildīti  Kļūdas gadījumā izvada kļūdas paziņojumu |
| Izvade: |
| Paziņojums par veiksmīgu reģistrāciju |

|  |
| --- |
| Komentāru klases noteikšana ar „Chrome” paplašinājumu |
| „Chrome” paplašinājums iezīmē negatīvus komentārus sarkanus, bet neitrālus kā zaļus. |
| Ievade: |
| Raksta tīmekļa adrese |
| Apstrāde: |
| Dokumenta parsēšana un komentāru izgūšana  Katra komentāra klases noteikšana |
| Izvade: |
| Katra komentāra iezīmēšana sarkanajā vai zaļajā krāsā atkarībā no klases |

|  |
| --- |
| Komentāra klases maiņa |
| Gadījumā, ja lietotājs nepiekrīt programmas rezultātam, ir iespējams manuāli izmainīt komentāra klasi |
| Ievade: |
| Komentāra teksts  Pareizā klase |
| Apstrāde: |
| Komentārs tiek pievienots pareizās klases treniņdatu kopai  Ar papildinātiem treniņdatiem funkcija „Komentāru klases noteikšana ar „Chrome” paplašinājumu” tiek palaista vēlreiz  Kļūdas gadījumā izvada kļūdas paziņojumu |
| Izvade: |
| Katra tekošās tīmekļa adreses raksta komentāra iezīmēšana sarkanajā vai zaļajā krāsā atkarībā no klases |

|  |
| --- |
| Treniņdatu pievienošana |
| Ir iespējams manuāli papildināt treniņdatu kopu |
| Ievade: |
| Jauno datu saturs  Paredzētā klase |
| Apstrāde: |
| Komentārs tiek pievienots dotās klases treniņdatu kopai  Kļūdas gadījumā izvada kļūdas paziņojumu |
| Izvade: |
| Paziņojums par veiksmīgu rezultātu |

|  |
| --- |
| Treniņdatu dzēšana |
| Ir iespējams manuāli nodzēst komentāru no treniņdatu kopas |
| Ievade: |
| Dzēšamie komentāri  Dzēšamā komentāra klase |
| Apstrāde: |
| Komentārs tiek dzēsts no dotās klases treniņdatu kopas  Kļūdas gadījumā izvada kļūdas paziņojumu |
| Izvade: |
| Paziņojums par veiksmīgu rezultātu |

|  |
| --- |
| Populārāko rakstu analīze |
| Ir iespējams atrast visus negatīvus komentārus no šobrīd populārākiem „rus.tvnet.lv” rakstiem |
| Apstrāde: |
| Tiek parsēti komentāri no šobrīd populārākiem rakstiem  Tie tiek saglabāti XML failā  Katra komentāra klase tiek noteikta ar klasifikatora palīdzību  Kļūdas gadījumā izvada kļūdas paziņojumu |
| Izvade: |
| Tiek izvadīti visi negatīvi komentāri no pārbaudītajiem rakstiem. Tiek izvadīta tabula, kas satur šādus laukus: komentāra izveidošanas laiks, komentāra autora lietotājvārds, raksta virsraksts ar saiti uz to, komentāra saturs |

|  |
| --- |
| Naidīgāko komentētāju atrašana |
| Ir iespējams atrast komentētājus ar vislielāko negatīvu komentāru skaitu no šobrīd populārākiem „rus.tvnet.lv” rakstiem |
| Apstrāde: |
| Tiek parsēti komentāri no šobrīd populārākiem rakstiem  Tie tiek saglabāti XML failā  Katra komentāra klase tiek noteikta ar klasifikatora palīdzību  Tiek atrasti komentētāji ar vislielāko negatīvu komentāru skaitu  Kļūdas gadījumā izvada kļūdas paziņojumu |
| Izvade: |
| Tiek izvadīti visnaidīgākie komentētāji no pārbaudītajiem rakstiem |

|  |
| --- |
| Pēdējo komentāru analīze |
| Ir iespējams izgūt pēdējos 50 komentārus no portāla „rus.tvnet.lv” šajā dienā rakstītajiem rakstiem |
| Apstrāde: |
| Tiek parsēti komentāri no šajā dienā rakstītajiem rakstiem  Tie tiek sakārtoti pēc izveidošanas laika  Tie tiek saglabāti XML failā  Katra komentāra klase tiek noteikta ar klasifikatora palīdzību  Kļūdas gadījumā izvada kļūdas paziņojumu |
| Izvade: |
| Tiek izvadīta tabula, kas satur šādus laukus: komentāra izveidošanas laiks, komentāra autora lietotājvārds, raksta virsraksts ar saiti uz to, komentāra saturs, komentāra klase, poga, kas ļauj mainīt komentāra klasi. |

|  |
| --- |
| Jaunās datu klases izveidošana |
| Lietotājs var izveidot jaunu datu klasi. Piemēram, „politiski komentāri” vai „komentāri par sportu” |
| Ievade: |
| Datu klases nosaukums  Pirmie šīs kopas komentāri |
| Apstrāde: |
| Tiek izveidots jauns teksta fails  Teksta failā tiek ierakstīti pirmie šīs kopas komentāri  Kļūdas gadījumā izvada kļūdas paziņojumu |
| Izvade: |
| Tiek izvadīta tabula, kas satur šādus laukus: komentāra izveidošanas laiks, komentāra autora lietotājvārds, raksta virsraksts ar saiti uz to, komentāra saturs, komentāra klase, poga, kas ļauj mainīt komentāra klasi. |

|  |
| --- |
| Datu klases krāsas maiņa |
| Lietotājs var piešķirt izvēlēto krāsu jaunai datu klasei. Sarkanā un zaļā krāsas ir rezervētas primārai kopai. |
| Ievade: |
| Jaunās klases krāsa |
| Apstrāde: |
| Tiek rediģēts dotās klases teksta kopas teksta fails  Teksta failā tiek ierakstīts izvēlētās krāsas nosaukums angļu valodā  Kļūdas gadījumā izvada kļūdas paziņojumu |
| Izvade: |
| Tiek izvadīts paziņojums par veiksmīgu darbību. |

### Ārējā saskarne

### Lietotāja saskarne

Visas sistēmas pogas, darbības un paziņojumi ir latviešu valodā. Sistēmai jābūt lietojamai bez tās papildus uzstādīšanas un jāpietiek tikai ar tīmekļa pārlūkprogrammu.

Par vizuālo sistēmas noformējumu atbild kaskadētu stila lapu valodas (CSS) nosacījumi, lietotāja izvēlētā interneta pārlūkprogramma un ierīces skatlauka (viewport) platums. Sistēma tiek veidota pēc responsīvā dizaina pamatnosacījumiem, tātad programmas vietnes daļa spēs pielāgoties un attiecīgi attēlot saturu, atbilstoši ierīces izmēriem. Programmas vietnes daļas responsivitāti pamatā nodrošina „Bootstrap” klienta pušu tehnoloģiju ietvars. Vietnes dizains, visos tai eksistējošajos skatos, ir vienkāršots un nemainīgs.

### Sakaru saskarne

Saskarnes veids, kādā lietotājs un sistēma savā starpā komunicē, tiek nodrošināts ar HTTP protokola palīdzību.

### Nefunkcionālās prasības

### Izmantojamība

### Drošība

Visi dati, kas tiek sūtīti izmantojot „POST” metodi, tiek filtrēti caur CSRF (XSRF) filtriem.

Lietotāju paroles datubāzē tiek glabātas vienīgi šifrētā veidā, izmantojot „MD5” šifrēšanas algoritmu. Tāpat arī visas darbības ar sistēmas lietotāju parolēm, notiek šifrētā veidā.

### Pārnesamība

Sistēmas pārnesamības sarežģītība nav augsta. Nepieciešamības gadījumā, sistēmu būtu viegli pārvietot uz citu serveri

# PROGRAMMATŪRAS PROJEKTĒJUMA APRAKSTS

1. Ievads - īsi izklāstīts šī dokumenta un sistēmas, uz kuru dokuments attiecas, mērķis un būtība.

2. Dekompozīcijas apraksts - sīkāk izklāstīts par sistēmas moduļiem un to savstarpējām sakarībām.

3. Datu plūsmu diagrammas - grafiski attēlotas sakarības starp sistēmas funkcijām, moduļiem un lietotājiem.

4. Datu bāzes projektējums – (iespējams)

5. Lietotāja ekrāna formu projektējums - izklāstīts ekrānformu projektēšanas process.

### Nolūks

Programmatūras projektējuma apraksta, jeb PPA, izveides nolūks ir tehniski konkretizēt programmatūras prasību specifikācijā (PPS) esošo informāciju, atvieglojot sistēmas izstrādes procesu.

### Saistība ar citiem dokumentiem

Programmatūras projektējuma apraksts tika izveidots, cenšoties balstīties uz Latvijas Valsts standarta LVS 72:1996, „Ieteicamā prakse programmatūras projektējuma aprakstīšanai” vadlīnijām.

### Definīcijas

Definīcijas un saīsinājumi sīkāk aprakstīti sadaļā „DEFINĪCIJAS UN SAĪSINĀJUMI”.

Datņu apraksts

Failu saraksts

NaiveBayesClass.php – šajā failā tiek aprakstīta sistēmas galvenā klase „sentiments”

Cross-val.php – kross-validācijas skripts

Negative\_data.txt – šajā failā atrodas negatīvās klases treniņdati

Neutral\_data.txt – šaja failā atrodas neitrālās klases treniņdati

Stop-words.txt – šaja failā atrodas „stopvārdi” – reti lietojami vārdi, kurus var neņemt vērā

Check\_label.php – programma, kas nosaka viena konkrētā komentāra klasi

Added\_negative.txt – lietotāja pievienotie negatīvās klases treniņdati

Added\_neutral.txt – lietotāja pievienotie neitrālās klases treniņdati

„Chrome” paplašinājums

Manifest.json – paplašinājuma galvenais fails

Content.js – paplašinājuma programmas pirmkods

Noskaņojuma analīzes veidus var sadalīt četrās grupās: analīze atbilstoši nosacījumiem, analīze ar vārdnīcu, pārraudzītā mācīšanās, nepārraudzītā mācīšanās.

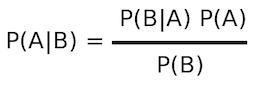
Pirmais veids ir analīze atbilstoši nosacījumiem (rule based analysis). Daudzas komercsistēmas lieto šo pieeju, kaut gan tā pieprasa nozīmīgus laika un resursu patēriņus, jo augstas precizitātes sasniegšanai ir jāapraksta pietiekami daudz noteikumu. Pie tam parasti noteikumi ir ļoti atkarīgi no projekta temata, tādēļ, mainot tēmu programma ir jāizstrādā pilnīgi no jauna. Neskatoties uz to visu, šī pieeja paliek ļoti populāra un, ja ir pietiekams datu apjoms, ļoti precīza, kaut gan no algoritmu viedokļa neļoti interesanta pētīšanai.

Analīze ar vārdnīcu. Tiek pielietotas tonalitātes vārdnīcas (affective lexicons). Vienkāršajā veidā vārdnīca ir vārdu saraksts ar katra vārda sentimentu koeficientu.

Nepārraudzītā mācīšanās (nonsupervised machine learning) - visinteresantākais, bet tajā pašā laikā vismazāk precīzs un veids. Šo metodi ir arī visgrūtāk īstenot. Mēs ļaujam programmai patstāvīgi atrast tekstu likumsakarības.

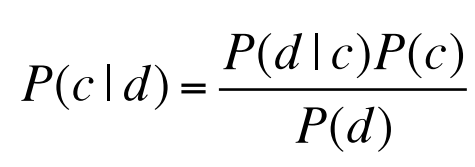
Pārraudzītā mācīšanās (supervised machine learning) ir visizplatītākā pētījumos metode. Lai realizētu šo pieeju, ir nepieciešams savākt treniņdokumentu kopu klasifikatora apmācīšanai. Katrs documents ir jāapzīmē ar pareizo klasi jeb sentimentu. Pēc tam jāizvēlās klasifikācijas algoritmu. Tad būs iespējams izmantot iegūto modeli. Es izmantošu tieši šo metodi.

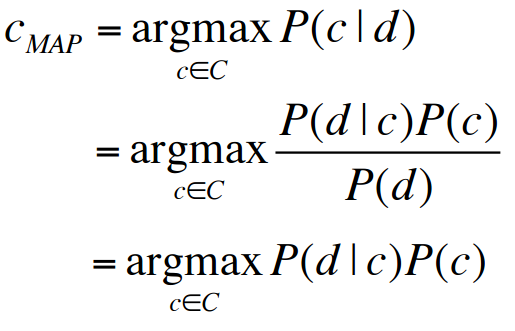
Naivais Baiesa klasifikators ir vienkāršs, bet ļoti efektīvs. Tas ir bieži izmantots tekstu analīzēs, un tas efektīvi strādā gan ar mazām, gan ar lielām datu kopām. Naivais Baiesa klasifikators ir balstīts uz Baiesa teorēmas. Klasifikators ir salīdzinoši viegli pielietojams.



Pēc Baiesa teorēmas var secināt, ka A notikuma varbūtība, ja B notikums ir patiess ir vienāda ar B notikuma varbūtības, ja A notikums ir patiess sareizinātu ar A notikuma varbūtību, un tas viss ir dalīts ar B notikuma varbūtību.

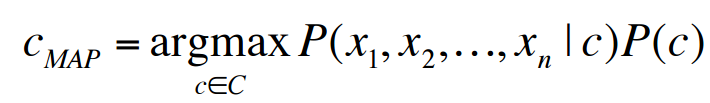
Mūsu gadījumā mēs varam pārrakstīt Baiesa teorēmu šādi: dokumentam d un klasei c, vārbūtība, ka dotais dokuments pieder klasei c ir





MAP – „maximum a posteriori” – visvairāk iespējamā klase

Saucēju var atmest, jo pie visām klasēm tas paliks nemainīgs.



P(d) izteikts, kā vārdu x1, x2, x3 ... xn kopa.

Klasifikators ir balstīts uz „vārdu somas” (bag of words) principa. Tas nozīmē, ka mēs varam uztvert tekstu, kā vārdu kopu, neņemot vērā gramatiku un vārdu kārtību, bet ņemot vērā vārdu biežumu.

Kāpēc naivais?

Klasifikators ir veidots ar diviem pieņēmumiem:

vārdu somas princips

ir pieņemts, ka katra vārda varbūtība ir neatkarīga, ja ir dota klase c

**Algoritma apraskts.**

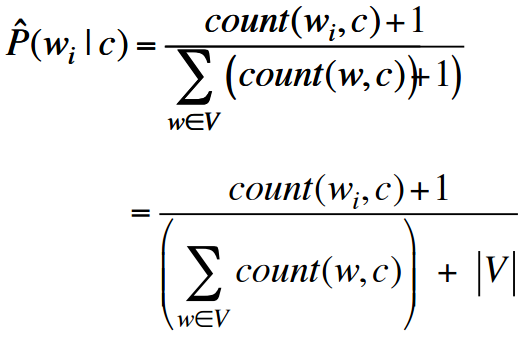
Vispirms ir jāiegūst no interneta nepieciešamie treniņdati – tekstu kopas, pēc kurām mūsu klasifikators varēs noteikt dotā teksta klasi.

Tika izveidota klase „sentiments”. Šai klasei ir definētas trīs galvenās metodes add, classify un tokenize.

Lai pievienotu dokumentu treniņdatiem , ir jāpadod dotais dokuments, kā parametrs funkcijai add. Funkcija add izsauc funkciju tokenize, kura sadala dokumentu atsevišķos vārdos. Pēc tam funkcija add pārtaisa katru vārdu augšējā reģistrā. Tiek apskatīts katrs vārds un ar bibliotēkas phpmorphy palīdzību tiek atrasta katra vārda normālforma. Dotais vārds tiek saglabāts mainīgajā index pie vajadzīgās klases.

Lai noteiktu, pie kādas klases pieder dotais dokuments, ir jāpadod šis dokuments kā arguments funkcijai classify. Tā veic tādu pašu vārdu apstrādi, kā add – tokenizācija, pārveidošana augšējā reģistrā un normālformu iegūšana. Pēc tam tiek iegūta katras klases vārbūtība pēc šāda principa -

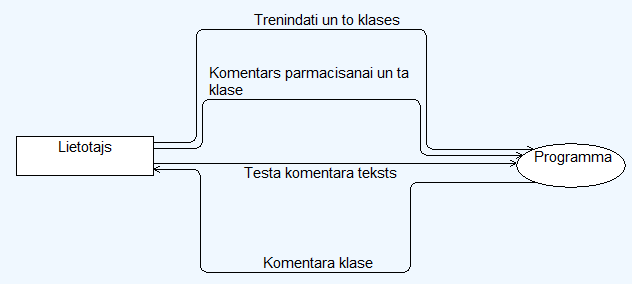
šis vārds šīs klases treniņdatos dalīts ar visiem vārdiem šajā klasē. Vēlāk tiek piereizināti prioru vārbūtības. Dažreiz tiek sastapti tādi vārdi, kas netika sastapti treniņdatos. Tādā gadījumā, lai novērstu reizināšanu ar nulli, tiek pielietota Laplasa izlīdzināšana (Laplace or add-1 smoothing).

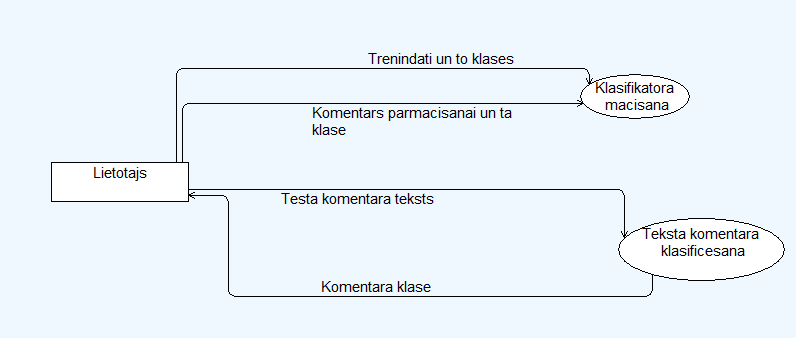


Lai saglabātu datu precizitāti, varbūtības tiek glabātas logaritmu formā.

**Šķērsvalidācija**

Lai noteiktu programmas precizitāti tika izveidota šķērsvalidācija. Treniņdati tika sajaukti un sadalīti 10 daļās. Ejot pēc kārtas viena datu daļa tika pieņemta kā testu datu kopa, bet pārējie 90% tika pieņemti kā klasifikatora treniņdati. Tika noteikta katra pārbaudījuma precizitāte, un pēc tam tika paņemta precizitāšu vidējā aritmētiskā vērtība.





# KVALITĀTES NODROŠINĀŠANA

Lai nodrošinātu augstu programmatūras kvalitāti, tika veiktas sekojošas darbības:

Programmatūras dokumentācijas veikšana pēc Latvijas valsts republikas

standartiem:

LVS 68:1996 Programmatūras prasību specifikācijas ceļvedis

LVS 72:1996 Ieteicamā prakse programmatūras projektējuma aprakstīšanai

Programmas kods tika rakstīts pēc vienota stila, panākot vieglu lasāmību un pārskatāmību.

Veidojot pirmkodu, tika izmantoti Latvijas Universitātes Datorikas fakultātes

kursā „Kvalifikācijas darbs” aprakstītos programmēšanas stila ieteikumi, iekļaujot:

Lietojiet komentārus programmas sākumā, lai dokumentētu programmas būtiskās īpašības

Lietojiet komentārus programmas tekstā, lai uzlabotu lasāmību

Rakstiet īsas programmas, funkcijas (apmēram vienu lappusi garas)

Izvēlieties paskaidrojošus un pielāgotus vārdus mainīgā veidam. Izvēlieties

pēc iespējas īsus mainīgo vārdus, tomēr, lai būtu saprotama jēga;

Lokālie mainīgie, īpaši skaitītāji var būt ar vienburta vārdiem – i, j, k utt.

Funkciju vārdus vēlams veidot no darbības vārda, kam, iespējams, seko lietvārds

Klašu vārdus sāciet ar lielo burtu

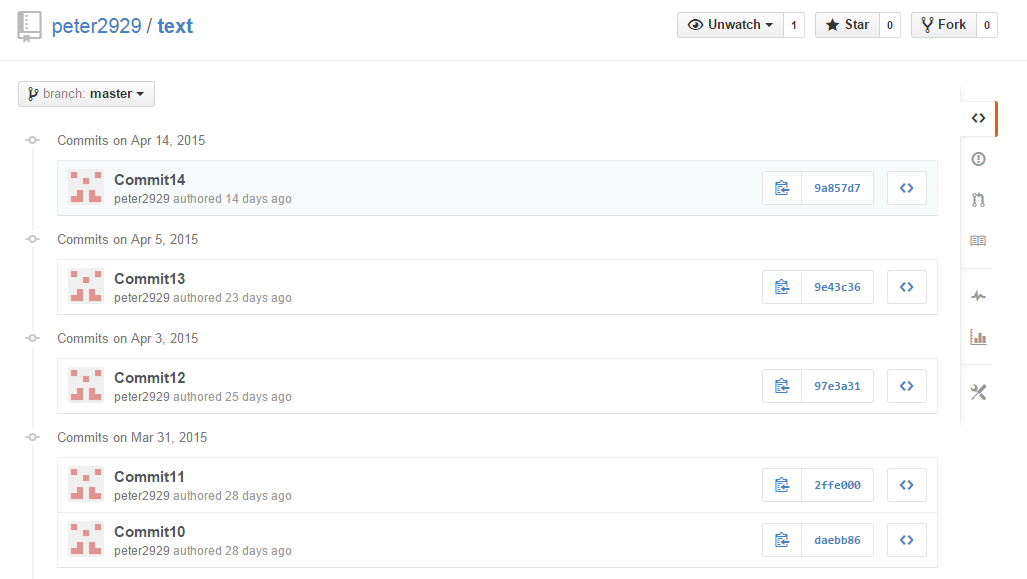
Aizvietojiet izteiksmes, kas atkārtojas, ar funkciju izsaukumiem

# Konfigurācijas pārvaldība

Programmatūras pirmkoda versiju kontrolei tika izmantota versiju kontroles sistēma Git, kodu glabājot savā “github” repozitorijā, papildus, dublējot failus arī uz ārējo cieto disku.

Izmaiņas tika “iestumtas” (git push) apmēram reizi nedēļā.

Katram failam ir iespējams apskatīt visas faila glabātās versijas, kopā ar komentāriem.



# TESTĒŠANAS DOKUMENTĀCIJA

Šis dokuments ir paredzēts, lai aprakstītu sistēmas testēšanas norisi un tās iegūtos rezultātus. Šī detalizētā vienībtestēšana, tika veikta projekta izstrādes beigās, taču, katras funkcijas vienībtestēšana, tika veikta arī projekta izstrādes procesa laikā. Tika mēģināts veikt testēšanu, lielākoties, pamatojoties uz Latvijas valsts standarta LVS 70:1996, “Programmatūras testēšanas dokumentācija” ieteikumiem.

Par sekmīgu testa rezultātu uzskata tādu, kas pilnībā atbilst sagaidāmajam rezultātam. Ja ir novirzes no sagaidāma rezultāta, tās ir jāatzīmē testēšanas rezultātos un radušās kļūdas jāizlabo.

**Testējamās raksturiezīmes**

Veicot sistēmas testēšanu tika pārbaudītas sekojošas raksturiezīmes:

Lietotāju saskarne – elementu izskats, lietotāja puses skriptu izpilde.

Funkcionalitāte – datu ievade, kļūdainu datu un izņēmu situāciju apstrāde.

Integrācija – lietotāja tiesību pārbaude.

Drošība – aizsardzība pret SQL injekcijām un starpvietņu skriptēšanu.

Par sekmīgu testa rezultātu uzskata tādu, kas pilnībā atbilst sagaidāmajam rezultātam. Ja ir novirzes no sagaidāma rezultāta, tās ir jāatzīmē testēšanas rezultātos un radušās kļūdas jāizlabo.

**Vienībtesti**

**Lietotāja reģistrācija sistēmā**

|  |  |  |  |
| --- | --- | --- | --- |
| **Testa** **nr.** | **Apraksts** | **Sagaidāmais rezultāts** | **Rezultāts** |
| 1. | Netiek aizpildīts kāds no obligāti aizpildāmajiem laukiem | Tiek pieprasīts aizpildīt visus laukus | Sekmīgs |
| 2. | Tiek ievadītas paroles, kuras nesakrīt | Tiek izvadīts paziņojums  „Paroles nesakrīt!” | Sekmīgs |
| 3. | Tiek ievadīta parole ar mazāk nekā 8 simboliem | Tiek izvadīts paziņojums „Parolei jāsatur vismaz 8 simbolus!” | Sekmīgs |
| 4. | Dati tiek ievadīti korektos  formātos un garumos | Dati tiek akceptēti, un  lietotāja dati tiek saglabāti  sistēmā | Sekmīgs |
| 5. | Tiek ievadīts lietotājvārds, ar mazāk nekā 3 simboliem | Tiek izvadīts paziņojums „Lietotājvārdam jāsatur vismaz 3 simbolus!” | Sekmīgs |
| 6. | Tiek ievadīts jau eksistējošs lietotājvārds. | Tiek izvadīts paziņojums „Šis lietotājvārds jau ir aizņemts!” | Sekmīgs |

**Lietotāja pieslēgšanās sistēmā**

|  |  |  |  |
| --- | --- | --- | --- |
| **Testa** **nr.** | **Apraksts** | **Sagaidāmais rezultāts** | **Rezultāts** |
| 1. | Netiek ievadīts lietotājvārds vai parole | Paziņojums, ka lietotājvārds  vai parole nav pareizi ievadīti | Sekmīgs |
| 2. | Dati tiek ievadīti pareizi | Dati tiek akceptēti, un  lietotājs tiek pieslēgts sistēmai | Sekmīgs |

**Treniņdatu pievienošana un dzēšana**

|  |  |  |  |
| --- | --- | --- | --- |
| Testa nr. | Apraksts | Sagaidāmais rezultāts | Rezultāts |
| 1. | Tiek ievadīti treniņdati ar klasi “Negatīvs” | Paziņojums, ka dati veiksmīgi pievienoti klases “Negatīvs” datu kopai | Sekmīgs |
| 2. | Tiek ievadīti treniņdati ar klasi “Neitrāls” | Paziņojums, ka dati veiksmīgi pievienoti klases “Neitrāls” datu kopai | Sekmīgs |
| 3. | Ievaddatu garums ir mazāks par 2 simboliem | Tiek izvadīts paziņojums „Lūdzu ievadiet treniņdatus!” | Sekmīgs |
| 4. | Tiek nodzēsts izvēlētais komentārs | Tiek izvadīts paziņojums „Komentārs tika nodzēsts!” | Sekmīgs |

**Komentāru klases noteikšana ar „Chrome” paplašinājumu**

|  |  |  |  |
| --- | --- | --- | --- |
| Testa nr. | Apraksts | Sagaidāmais rezultāts | Rezultāts |
| 1. | Klases noteikšana | Negatīvi komentāri tiek iekrāsoti sarkanā krāsā, bet neitrāli – zaļajā. | Sekmīgs |
| 2. | Komentāra klases maiņa | Nospiežot pogu pie komentāra, lapa tiek atjaunota un dotais komentārs tiek pārkrāsots prētējā krāsā. | Sekmīgs  Piezīme: ja komentārs ir ļoti īss mēdz gadīties, ka klase paliek nemainīta |
| 3. |  |  | Sekmīgs |

**Šķērsvalidācija**

|  |  |  |  |
| --- | --- | --- | --- |
| Testa nr. | Apraksts | Sagaidāmais rezultāts | Rezultāts |
| 1. | Šķērsvalidācijas precizitātes un atsaukuma noteikšana | Tiek noteikti šķērsvalidācijas precizitāte un atsaukums | Sekmīgs |
| 2. | Grafiskās diagrammas iezīmēšana | Tiek izveidots grafiks, kas parāda šķērsvalidācijas precizitātes atkarību no treniņdatu apjoma | Sekmīgs |
| 3. |  |  | Sekmīgs |

# PROGRAMMATŪRAS PROJEKTA ORGANIZĀCIJA

Sistēmas izstrādē tika izmantots iteratīvās izstrādes modelis, visas darbības veicot pakāpeniski –

programmu prasību specifikācijas izstrādei seko programmatūras projektējuma apraksts, pēc

kura tiek veikta sistēmas izstrāde un testēšana.

Kad sistēmas izstrāde tika pabeigta, tika veikta testēšana, kurā tika pārbaudīta sistēmas un

funkciju darbība, padodot funkcijām dažādus - gan akceptējamus, gan neakceptējamus

ievaddatus. Testēšanas rezultāti ir apkopoti testēšanas dokumentācijas sadaļā.

# DARBIETILPĪBAS NOVĒRTĒJUMS

Tika paredzēts, ka dokumentācijai (programmatūras prasību specifikācijai, programmatūras projektējuma aprakstam un testēšanas dokumentam) būs nepieciešams aptuveni viens personmēnesis, bet produkta izstrādei – trīs personmēneši.

# IZMANTOTĀ LITERATŪRA

LVS 68:1996 Programmatūras prasību specifikācijas ceļvedis

<http://estudijas.lu.lv/mod/resource/view.php?id=131427>

LVS 72:1996 Ieteicamā prakse programmatūras projektējuma aprakstīšanai.

[http://estudijas.lu.lv/mod/resource/view.php?id=131428](http://estudijas.lu.lv/mod/resource/view.php?id=131428%20)

“Learning Classifiers from Imbalanced, Only Positive and Unlabeled Data Sets” ” [tiešsaiste -pārbaudīts 07.05.2015]. Pieejams:

<https://www.cs.iastate.edu/~yetianc/cs573/files/CS573_ProjectReport_YetianChen.pdf>

„Machine Learning from Imbalanced Data Sets 101” [tiešsaiste - pārbaudīts 07.05.2015]. Pieejams:

<http://pages.stern.nyu.edu/~fprovost/Papers/skew.PDF>

„On the Class Imbalance Problem” [tiešsaiste - pārbaudīts 07.05.2015]. Pieejams:

<http://sci2s.ugr.es/keel/pdf/specific/congreso/guo_on_2008.pdf>

**Jason Rennie** „*Tackling the Poor Assumptions of Naive Bayes Text Classifiers*” [tiešsaiste – pārbaudīts 07.05.2015]. Pieejams:

<http://people.csail.mit.edu/jrennie/talks/icml03.pdf>

*“Machine Learning - Lecture 4: The Naive Bayes Classifier”* [tiešsaiste – pārbaudīts 07.05.2015]. Pieejams:

<http://www.sussex.ac.uk/Users/christ/crs/ml/lec02b.html>

**Jason D. M. Rennie , Lawrence Shih , Jaime Teevan , David R. Karger**

“Tackling the Poor Assumptions of Naive Bayes Text Classifiers” [tiešsaiste – pārbaudīts 07.05.2015]. Pieejams:

<http://www.aaai.org/Papers/ICML/2003/ICML03-081.pdf>

*“Sentiment Analysis of Micro blogs using OpinionMining Classification Algorithm”* [tiešsaiste – pārbaudīts 07.05.2015]. Pieejams: <http://www.academia.edu/5320342/Sentiment_Analysis_of_Micro_blogs_using_Opinion_Mining_Classification_Algorithm>

# Pielikumi

**Pirmkoda paraugs**







# Secinājumi

Programmatika sekmīgi izstrādāta atbilstoši tās prasību specifikācijai, iekļaujoties 4 mēnešu laika periodā. Darba izstrādes gaitā tika apgūti projekta pārvaldības un dokumentu izstrādes pamati. Tika gūtas praktiskas iemaņas darbā ar Git sistēmu.

Tika iegūtas teorētiskās un praktiskās zināšanās par mašīnmācīšanās metodēm, konkrēti par tekstu apstrādes veidiem. Tika izstrādāts Naivā Baiesa klasifikators, kura šķērsvalidācija uzrādīja 98% precizitāti.

Tika apgūtas prakstiskās iemaņas darbā ar AJAX tehnoloģiju.