|  |  |
| --- | --- |
| http://students.uni-sofia.bg/wp/wp-content/uploads/2010/11/logo-su-s-nadpis.jpg | Софийски Университет „Св. Климент Охридски“  Факултет по Математика и Информатика |

**Курсова работа**

за курса по

**Изкуствен Интелект**

на тема

**„Класификация на съобщения от електронна поща“**

Тодор Борисов Михайлов, ф.н.24425

магистър, специалност „Извличане на Информация и Откриване на знания“

Дата 09.02.2014

Съдържание

[Въведение 3](#_Toc379691021)

[Обзор 3](#_Toc379691022)

[Решение 4](#_Toc379691023)

[Начин на работа 4](#_Toc379691024)

[Подготовка на класификацията 4](#_Toc379691025)

[Експерименти 5](#_Toc379691026)

[Weka 5](#_Toc379691027)

[Представяне на данните 7](#_Toc379691028)

[Резултати от тестовете 8](#_Toc379691029)

[Програмна реализация 9](#_Toc379691030)

[Изглед на приложението 10](#_Toc379691031)

[Заключение 11](#_Toc379691032)

[Използвани източници 12](#_Toc379691033)

[Класификация и обработка на документи и електронна поща 12](#_Toc379691034)

[**Информация за работа с Weka** 12](#_Toc379691035)

[Java 13](#_Toc379691036)

# Въведение

Класификацията на съобщения от електронна поща е проблем, който се увеличава с увеличаването на информацията, която хората получават всеки ден. Въпреки, че има различни канали за получаване на информация, електронната поща е най-разпространеният и един от най-ефективните начини за това. В ежедневието хората използват електронната поща в работата и в личния живот за комуникация. С всяка следваща година, броят съобщения, които средностатистическия интернет потребител получава се увеличава. С това се поражда и необходимостта за класифицирането и улесняването на поддръжката на голям брой съобщения. Моята идея е да разгледам класификация на електронна поща по теми, такива, каквито хората срещат в ежедневието. Например класификация на съобщенията като отделни групи – новини, поръчки на стоки от интернет, съобщения от форуми и социални мрежи, различни работни проекти и др.

# Обзор

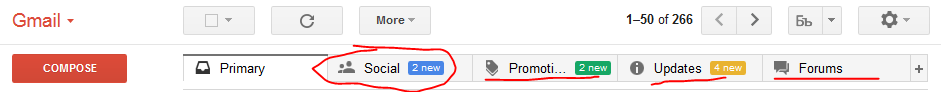
В научен аспект класификацията на съобщения от електронна поща е проблем, на който не е обръщано много внимание през годините. Причината за това е, че проблемът не е ясно дефиниран, тъй като е специфичен за всеки човек и начинът, по който той организира пощата си не е стандартен. Също така няма достатъчно публични корпуси с електронна поща, достатъчни за по-задълбочен анализ, както, например текстове и статии от определени области.

Единствената по-сериозна база от данни с електронна поща е Enron корпусът, който представлява колекция от пощенските кутии на корпоративни служители на американска енергиина компания, която бива разследвана за измами през 2003 и впоследствие кореспонденцията между служителите бива публикувана. Корпусът се състои от около 500 хиляди съобщения.

Има няколко статии на тази тема, които правят изводи за начинът, по който могат да се обработят съобщенията и са базирани на Enron корпусът.

Съществуват някои услуги и приложения, които предлагат възможност за обработка на електронна поща. Повечето от тях се ограничават до маркирането на съобщения като спам/неспам, важни/неважни, без да предлагат възможност за класификация по различни теми, проекти, по които работи човек в работата и т.н. Ето примерни услуги:

* http://www.sanebox.com/ - Услуга за приоритизиране на важна поща. За да я използва потребителят трябва да се регистрира на с e-mail адрес и да предостави парола за достъп. Съобщенията се свалят на техен сървър, класифицират се като важни/неважни, за по-късно и т.н. и се изпраща съобщение с информация за извършените класификации, ткаа, че потребителят може да види кои съобщения как са класифицирани и да обучи системата.
* Google Gmail - работят в насоката за класификация на съобщенията по дадени теми. Например има предефинирани „табове“, в които потребителят може да добави дадени съобщения и следващите от същия тип да се показват там.



Все още обаче няма възможност за добавяне на собствени класове.

# Решение

## Начин на работа

В текущата курсова работа е разгледан проблемът за класификация на електронна поща по теми. Подходите, които са избрани са базирани на лични наблюдения върху лична и работна електронна поща и са взети на предвид изводи от статии с подобни разработки за класификация на електронна поща.

Направени са експерименти за класифицирането на съобщенията, чрез наивен бейсов класификатор и Support Vector Machine. Данните, върху които са направени тестовете са от корпусът Enron и от собствена лична електронна поща.

Имплементирана е програмна реализация на класификатор на електронна поща, базиран на изводите и от тестовете и направен модел за вземане на обратна връзка от потребителя, чиято поща се класифицира.

При имплементацията на текущия класификатор са използвани методи споменати в секцията „Източници“, като основните подходи са взаимствани от труда [Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora](http://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1217&context=cs_faculty_pubs) на Ron Bekkerman, 2004.

По-долу са описани стъпките за обработка на електронна поща и избрания подход.

## Подготовка на класификацията

1. Премахват се папките които нямат някаква смислена тема.
   * Това са папки които са по подразбиране. Tакива папки в Gmail са:
     1. Trash
     2. Sent Mail
     3. All mail
     4. Spam
     5. Important
     6. Starred

Например, Gmail маркира голяма част от писмата като Important, без да има някакъв видим смисъл – по ското по ключови думи и подател. Обикновено там влизат и писма изпратени като subscription и др. и не бихме искали да тренираме класификатора с тях.

Папки с определена тематика могат да са „News“, „Travel“, “Business”, „Social“, “Projects” и др.

* + Ще премахнем и папки, които имат малък брой писма – под 5. Като по-късно , ако потребителят премести писма в тях може да ги използваме за трениране на класификатора.

1. Трениране на класификатора
   * Тренирането на съобщенията ще стане, като ще използваме папките за класове и ще добавяме последователно писмата в тях към класификатора
   * Ще използваме като feature-и полетата From, Subject, Body, събрани като един текст и разделени на думи – торба с думи. Няма да изключваме From, т.к. в много от случаите именно подателят е този по който се орпеделя класа на документа.

# Експерименти

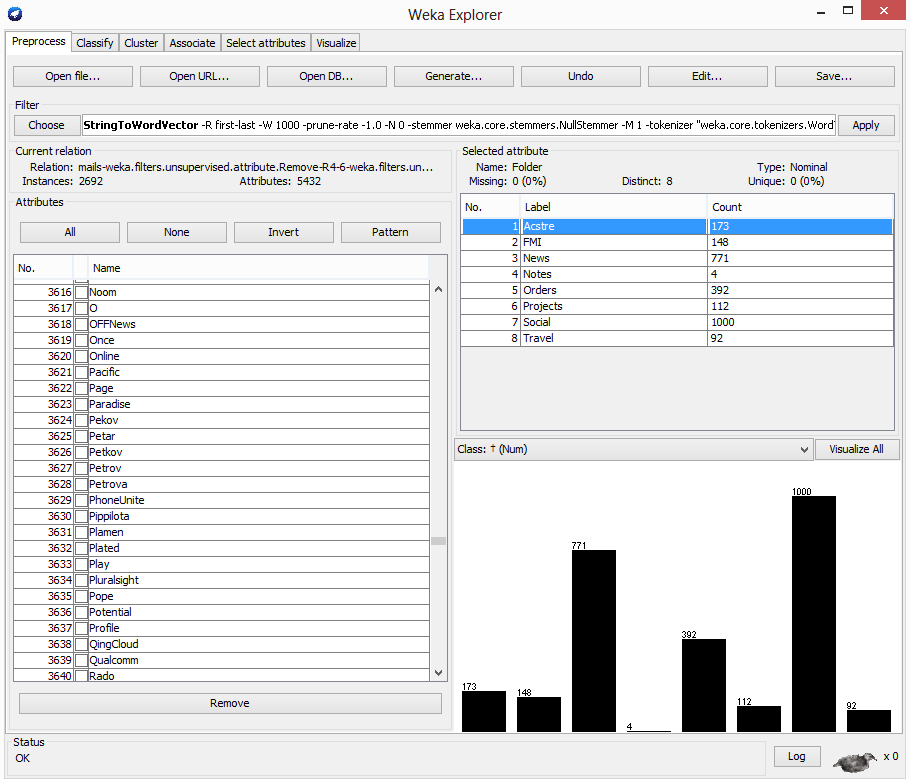
За експериментите е използвана Weka – съвкупност от различни класифициращи/клъстеризиращи и статистически алгоритми за обработка на данни.

Използван е 10 fold cross validation модел за валидация на резултатите – разделят се на 10 части, тренира се с 9 и се тества с 1.

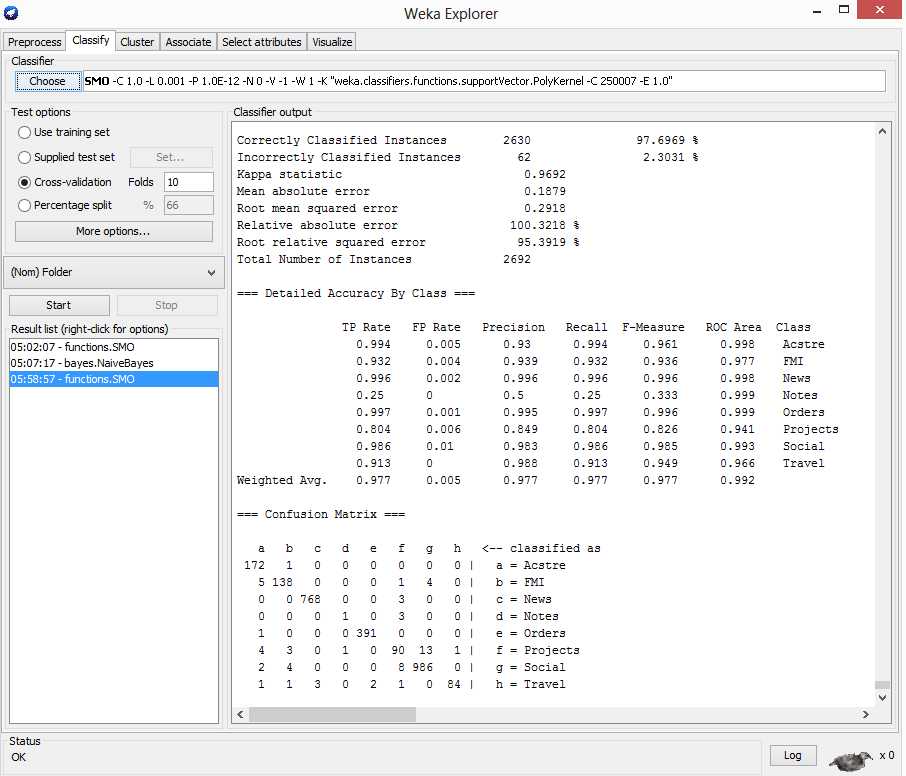
## Weka

Ето как изглежда инстументът Weka:

* Избор на feature-и



* Резултати от класификация



## Представяне на данните

Разгледани са 3 кутии на електронни пощи, като са извадени само така наречените topic folders – папки, в които документие са класифицирани по смисъл. Класифицирани са ръчно от собствениците на пощата:

* Електронна поща на Vincent Kaminsk, Head of Quantitative Modeling Group, Enron cospus

Смислени папки:

|  |  |
| --- | --- |
| bill | personal |
| bill\_williams\_iii | preschedule |
| california\_messages | rt\_cuts |
| el\_paso | rt\_strat |
| enron\_messages | schedule\_crawler |
| gwolfe | settlements |
| hr | symesees |
| human\_resources | tie\_meter\_multipliers |
| operations\_committee\_isas | timbelden |

* Електронната поща на, Michelle Lokay, Administrative Assistant, Enron corpus

Смислени папки:

|  |  |
| --- | --- |
| articles | personal |
| corporate | publications |
| enron\_travel\_club | systems |
| enron\_t\_s | training |
| hea\_nesa | tw\_commercial\_group |
| kim\_s\_files | personal |

* Лична електронна поща – tbmihailov

Смислени папки

|  |  |
| --- | --- |
| Acstre | Orders |
| FMI | Projects |
| News | Social |
| Notes | Travel |

## Резултати от тестовете

В таблицата е показана информация за броя съобщения, броя папки и резултата от класификация.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Информация за данните | | | Точност на резултатите от класификацията | |
| Поща | Брой съобщения | Брой папки | NB | SVM |
| tbmihailov | 2692 | 8 | 92.6077 | 97.9941 |
| lockay-m | 2493 | 11 | 79.6229 | 96.2294 |
| williams-w3 | 2769 | 18 | 95.1246 | 98.8443 |

Графично представяне на резултата

В случая SVM се е справил по-добре от Naïve Bayes. Затова сме го избрали за програмната реализация на проекта.

# Програмна реализация

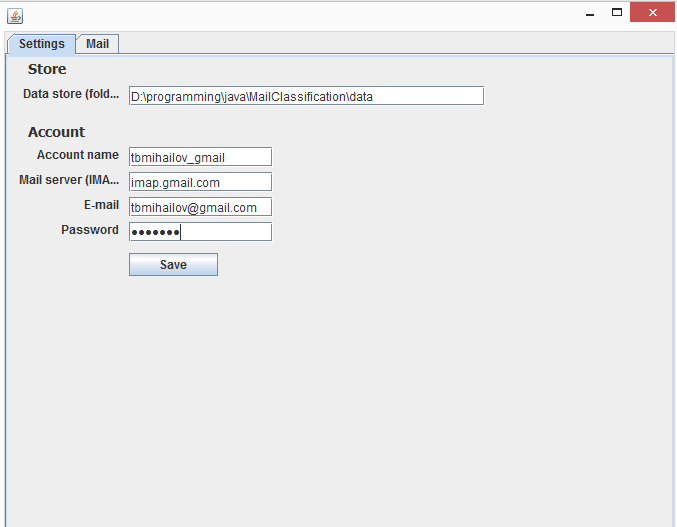
Реализирано е приложение за извличане на електронна поща по зададен акаунта, изграждането на модел за класификация с SVM и съхраняването на този модел. Приложението има следните възможности:

* Извличане на електронна поща през IMAP – Java през IMAP – поддръжа се изтегляне на папките и съобщенията към тях.
* Определяне на значещите папки, чрез използването на речник за незначещите – на базата на анализити и тестовете от първата част е изваден списък със не значещите папки и те се премахват – това е inbox, sent, deleted и др.
* Трениране на SVM класификатор със съобщенията – Използвана е библиотеката Weka, като през Java се създават инстанции на класификатора и филтрите за привеждане на данните към вид за обработка.
  + Създадени са класове за трениране и форматиране на данните.
* Съхраняване на модела във файл, с цел отваряне при следващата проверка на ел. поща
  + Създадена е възможност за записване на изграден модел, така, че да не бъде изграждан на ново при следващото стартиране
* Извличане на нови съобщения от електронната поща – извличат се съобщенията след датата на последното трениране
* Класифициране на новите съобщения и предоставяне на интерфейс за потребителя за потвърждение и корекция
  + Потребителят може да избира нови стойностите за класовете, ако класифицирането не е удовлетворяващо
* Преместване на съобщенията по класификационните папки – на базата на класификацията и на обратната връзка съобщенията се преместват в съответните папки, като се използва Java mail клиент и IMAP, за отразяване на промените на сървъра.

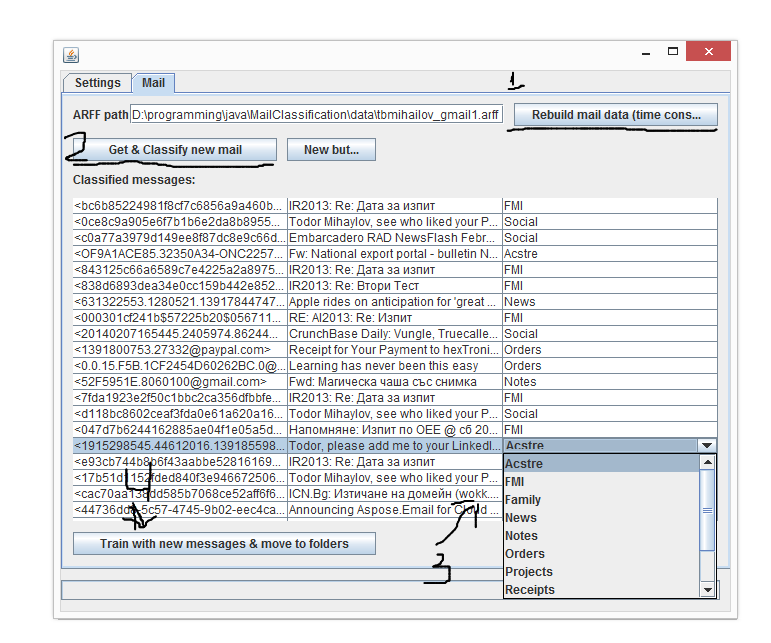
Приложението се стартира като конзолно приложение или като потребителски интерфейс.

## Изглед на приложението

1. Данни за потребителя



1. Извличане, класификация и обратна връзка от потребителя. Трениране с новите данни:



1. Първоначално изтегляне на данните и трениране с Weka – използва с Java API-а.
2. Изтегляне на нова поща и класифициране върху първоначално създадения модел.
3. Потребителят може да даде обратна връзка, ако нещо не е класифицирано, както се очаква от него
4. На базата на класификацията и на обратната връзка класификатор се тренира с новите данни

# Заключение

В текущата курсова работа е разгледан проблемът за класификация на електронна поща по теми. Можем да твърдим, че резултатът от курсовата работа е положителен, тъй като са проучени са съществуващи трудове по въпроса и са използвани изводи от тях и лични наблюдения за изграждане на опростен и завършен цикъл на класификация на съобщения с учител. Необходими са още проучвания по въпроса, тъй като разгледаните данни са много малко – 3 частни случая за електронни пощи, което е недостатъчно за формулиране на ясни правила за обработката на електронни съобщения.

Проблемът за класификация на електронна поща става все по приложим и необходим, с увеличаващата се информация и има смисъл да се правят бъдещи разработки.

Има смисъл да се разработи плъгин за популярните клиенти – Thunderbird, Outlook и др. и да се тества в повече случаи от повече потребители.

# Използвани източници

Използвани са източници за разучаване на средата Weka и разглеждане на трудове в сферата на категоризацията на документи и в частност електронна поща.

## Класификация и обработка на документи и електронна поща

* *A Comparative Study for Email Classification - Seongwook Youn and Dennis McLeod, University of Southern California – Сравнява класификацията на спам съобщения с различни методи – Support Vector Machine, Naïve Bayes, Невронна мрежа, Decision tree (J48)*
* *http://imsc-dmim.usc.edu/publications/Youn-CISSE.pdf*
* *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.15.2910&rep=rep1&type=pdf*
* *http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.206.3185*
* *http://www.cs.rpi.edu/~szymansk/theses/ozcaglar.08.ms.pdf*
* [*http://people.cs.umass.edu/~yanlei/publications/pakdd00.pdf*](http://people.cs.umass.edu/~yanlei/publications/pakdd00.pdf)
* http://management.haifa.ac.il/images/info\_people/ron\_bekkerman\_files/document%20classification%20on%20enron%20email%20dataset.pdf - read
* <https://www.cs.cmu.edu/~enron/> - Enron корпус
* http://arg.vsb.cz/arg/Enron\_Corpus/ - parsed enron corpus - read
* http://ceas.cc/2004/168.pdf - introducing the enron corpus - read
* http://ortho.clmed.ncku.edu.tw/~emba/2006EMBA\_MIS/3\_16\_2006/WekaIntro.pdf - weka intro - read
* http://imsc-dmim.usc.edu/publications/Youn-CISSE.pdf - read
* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.15.2910&rep=rep1&type=pdf - read

## **Информация за работа с Weka**

* http://www.aboutdatamining.com/tutorials/weka-tutorial-1-preparing-the-weka-arff-input-file/
* http://weka.wikispaces.com/Use+WEKA+in+your+Java+code
* Eclipse with weka tutorial: http://www.cs.umb.edu/~ding/history/480\_697\_spring\_2013/homework/WekaJavaAPITutorial.pdf
* http://www.unal.edu.co/diracad/einternacional/Weka.pdf - text classification and clustering
* http://ianma.wordpress.com/2010/01/16/weka-with-java-eclipse-getting-started/ - getting started with weka in eclipse
* http://weka.wikispaces.com/ARFF+files+from+Text+Collections -- Directory to Weka DataSet
* http://www.cs.columbia.edu/~kathy/cs4701/documents/jason\_svm\_tutorial.pdf - read
* http://weka.wikispaces.com/MessageClassifier
* http://bailando.sims.berkeley.edu/enron\_email.html - enron database mysql
* Explanation of SMO parameters in weka
* <http://list.waikato.ac.nz/pipermail/wekalist/2010-December/050570.html>

## Java

* --Code resources - http://stackoverflow.com/questions/4790844/how-to-get-the-list-of-available-folders-in-a-mail-account-using-javamail