**Документация**

**Spam класификатор**

**Описание**

Конзолно приложение, написано на Pelr, което сваля писмата от зададена електронна поща и ги класифицира като спам съобщения или обикновени съобщения.

**Функционални части**

1. Сваляне на email-и от даден mail server

* Свалянето на email-ите се осъществява при стартиране на програмата.
* Конфигурацията за достъп до даден mail server се осъществява в config.ini файла в секцията [MailHostConfig], в която има следните параметри:

username=perl.spam.test@gmail.com

password=perl.test

mailhost=pop.gmail.com

port=995

* За свалянето на email-ите използваме модула Mail::POP3Client, а за извличането на body-то на съобщението използваме модула Email::MIME

1. Извличане на примерните данни, необходими за обучаване на спам филтъра, от файлове намиращи се на файловата система.

* Пътят до обучаващите файлове се конфигурира в config.ini файла в секцията [TrainingSetDirs], която има следните параметри:
  + spamDir=TrainingMailSets\spam – директория за примерите, които са спам(отирцателни примери)
  + hamDir=TrainingMailSets\ham - директория за примерите, които не са спам(положителни примери)
* Извличането на примерните данни за обучението става чрез следните функции:
  + extractMailWords – приема като параметри цялото съобщение и dictionary(optional), в което да бъдат записани отделните думи. Функцията игнорира думи с дължина по-малка от 3 симова и думи съставени само от цифри, защото те не носят достатъчно информация. Функцията връща dictionary-то с думите.
  + wordsStatisticInFile – приема име на файл, от който да четем съобщението и hash reference. Извиква extractMailWords с тези параметри и връща като резултат, резултата върнат от extractMailWords.
  + getHamAndSpamStatistics – обхожда всички файлове от обучаващото множество и за всеки файл извиква wordsStatisticInFile като резултатите се запазват в глобалните dictionary-та:

%hamWords;

%spamWords

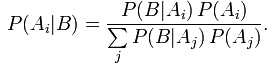
1. Филтриране на свалените email-и и класифицирането им.

* Филтрирането на съобщенията се извършва като за всяко съобщение, по алгоритъма на Бейсън, се пресмята вероятността му то да бъде спам или не.
* За спам се приемат съобщения, чиято вероятност е по-голяма или равна на 0.9.

Алгоритъмът се осъществява от следните функции:

* calcMailProb – приема масив с вероятностите на думите от съобщението.

Според тези вероятности, по разширената формула на Бейсън, се пресмята вероятността съобщението да е спам или да не е.



* calcWordsProb – приема масив с думите, чиито вероятности искаме да пресметнем.
  + За всяка дума от масива проверяваме колко пъти сумарно се среща в двата hash-a: hamWords и spamWords. Ако броят на срещанията са по-малко от 5, то не пресмятам вероятността на тази дума, тъй като тя няма да окаже влияние върху крайния резултат.
  + Броят на срещанията на думите от положителното множество се умножава по 2 с цел да намалим случаите, в които съобщенията, които не са спам се класифицират като такива.
  + Връща hash с думите и съответните им вероятности

**Използвани модули**

Mail::POP3Client

Email::MIME

Config::IniFiles

File::Slurp

List::Util