Projekt 2 - Antispam

Michal Ormoš (xormos00@stud.fit.vutbr.cz)

14. prosince 2017

Abstrakt

Táto krátka správa predstavuje dokumentáciu k druhému projketu pre predmet BIS. Popisuje zdroje používane pri spracovaní tohto zadania a v krátkosti predstavuje riešenie daného problému.

1 Využívané neštandardné knižnice

`email` - Parsovanie formátu .eml
`numpy`, `collections` - matematické funkcie a datatypy
`sklearn`, `pickle` - strojové učenie a ukladanie slovníkov/modelov Implementácia prebehla v programovacom jazyku python3.

2 Získavanie dát pre strojové učenie

Dáta pre trénovanie modelu pre anglický jazyk boli získané zo zdrojov `ling-spam corpus`[1] a `Euron-spam corpus`[2], kde boli rozdelené rovnomerne na emaily vyžiadanej(ham) pošty a nevyžiadanej(spam) pošty, pričom implementácia trénovania modelu zohľadňovala túto skutočnosť. Kedže žiadny zdroj pre český/slovenský spam som neobjavil, do trénovacích dát som ručne prihodil pár svojich nedôležitých osobných ham/spam emailov vyexportovaných zo svôjho osobného email účtu.

Pri získavaní dát z textu robia najväčší problém znaky, ktoré netovria žiadne rozumné slová a neprinášajú do textu žiadnu hodnotu. Tieto znaky sa snažíme predspracovaním z textu odstrániť. Jedná sa o znaky ako napríklad interpunkcia, bodky na konci viet, čísla, atd. Texty z `ling-spam corpus` sú už takto z časti predspracované, neobsahujú bodky konca viet a sú zlemmatizované. Tedá slová sú vo svojom základnom tvare. Ostatné nepotrebné znaky si musíme odstrániť sami a to vytvorením slovníka.

3 Vytvorenie slovníku

Vytvorením slovníka si spomedzi všetkých emailov slova v ich základnom slovesnom tvare zoradíme podľa početnosti (počtu výskytov v texte).

Pomocou tochto získaného poradia už jednoducho odstránime spomínané nepotrebné znaky, rovnako ako aj slová ktoré pozostávajú len z jednoho písmena

4 Trénovanie modelu

Pre trénovanie modelu som využil všetky dostupné funkcie knižnice `sklearn`. Experimentoval som s trénovaním dát pomocou trénovacích metód `LinearSVC` a `MultinomialNB`, a ako lepšia a presnejšia možnosť sa pre naše účely ukázala `MultinomialNB`. Natrénovaný model a slovník som uložil spolu s programom a dáta sa už pri ďalších spusteniach viac netrénujú.

5 Testovacie dáta

Pre testovacie dáta som znova použil zdroj `CSmining group` avšak, narozdiel od trénovacích dáta som použil emaily z corpusu `Spam email datasets`[3]. Jeho časť `TRAIN_DATA` obsahovala 4327 emailov v pomere:

Ham	Spam	Sum
2949	1378	4327

Mnou nameraná úspešnosť bola:

Ham	Spam	Fail	Sum
2871	1443	13	4327

6 Príklad spustenia

```
$ make
$ ./antispam ham.eml spam.eml deletefile.eml
ham.eml - OK
spam.eml - SPAM
deletefile.eml - FAIL - FAILED TO OPEN FILE
```

Reference

[1] http://csmining.org/index.php/ling-spam-datasets.html

[2] http://nlp.cs.aueb.gr/software_and_datasets/Enron-Spam/index.html

[3] http://csmining.org/index.php/spam-email-datasets-.html