

Spam Filter

Semestrální úloha z RPH

Jakub Ambroz a Kateřina Kučerová



České vysoké učení technické

6. ledna 2021



Úvod

- ▶ Knihovna emails k rozdělení textu souboru na části
- ▶ Tokenizace pomocí "str.replace" a *regular expression* k odstranění html tagů
- ▶ Rozdělení na slova



Popis

- ▶ Náš filtr funguje na principu počítání četnosti výskytu slov ve spamech a v hamech. Pak počítá jestli je slovo v testovaném emailu častěji v spamech či v hamech
- ▶ Využití metadat o emailu stejným způsobem
- ▶ Spojení obou pravidel logickým and - odstranění většiny False Positives



Testovací sada

- ▶ Vytvoření náhodné trénovací a testovací. Napůl pomocí pythonu napůl příkazové řádky

```
print("Creating appropriate directories.")
os.system("mkdir " + save_path)
os.system("mkdir " + save_path+"/train/")
os.system("mkdir " + save_path+"/test/")
print("Saving files into data set folders")
for name, status in dictionary.items():
    if (random.randint(1, 100) <= percentage_used): # percentage portion that is used for training
        dict_train[name] = status
        #print("Train: " + name + " : " + status)
        os.system("cp " + path + "/" + str(number) + "/" +
                    name + " " + save_path+"/train/"+name)
    else:
        dict_test[name] = status
        #print("Test: " + name + " : " + status)
        os.system("cp " + path + "/" + str(number) + "/" +
                    name + " " + save_path+"/test/"+name)
utils.write_dict_to_file(save_path+"/train/!truth.txt", dict_train)
utils.write_dict_to_file(save_path+"/test/!truth.txt", dict_test)
print("Success!\n Subsets created:\n" +
      save_path+"/train/\n"+save_path+"/test/")
print("Used dataset: " + str(number))
print("Used portion for trianing " + str(percentage_used) + "% ")
```



Spolupráce: GitLab

The screenshot displays the Open Trello interface with four boards, each containing a list of tasks with due dates and estimated times.

- Future Board (6 items):**
 - #13: Figure out how to train simply (To Do)
 - #6: Prezentace v TeXu (Jan 4, 2021, 3h)
 - #10: Funkční verze k odevzdání (Monday, 1d 6h 12m)
 - #11: Odevzdat vole! (Monday, 10m)
 - #12: Nezapomenout odevzdat (Tuesday)
 - #15: Subset generator (Sunday, 1h)
- To Do Board (10 items):**
 - #1: Vytáhnout metadata z txt souboru (Tomorrow, 40m)
 - #2: Vytřít text emailu (Tomorrow, 40m)
 - #3: Training - remember sender (Tomorrow, 20m)
 - #4: Spočítat slova (Tomorrow, 20m)
 - #5: Najít spam slova (Saturday, 1h)
 - #13: Figure out how to train simply (Future)
 - #7: Kačka setup (Tomorrow, 10m)
 - Advanced setup
- Doing Board (1 item):**
 - #9: Todo-list!
- Meta Board (4 items):**
 - #7: Kačka setup (To Do, Tomorrow, 10m)
 - #8: Advanced setup (To Do, Tomorrow, 15m)
 - #11: Odevzdat vole! (Future, Monday, 10m)
 - #12: Nezapomenout odevzdat (Future, Tuesday)



Spolupráce: VSCode - Liveshare

```
1 #Jakub Ambroz
2 #03.12.2020
3 #RPH - uloha SPAM FILTER
4 #krok4: base filter
5 import utils Kucerova, Katerina
6 import corpus
7 import os
8 class BaseFilter():
9     """
```



Výsledky

Results on dataset 1

Username	TP1	TN1	FP1	FN1	q1	Pts1	Rank1
Best submission for dataset 1	461	153	0	0	1.0	4	1.0
ambrojak_kucerka7	415	148	5	46	0.8543247344461306	3	32.0

Results on dataset 2

Username	TP2	TN2	FP2	FN2	q2	Pts2	Rank2
Best submission for dataset 2	431	147	3	19	0.9218500797448166	4	1.0
ambrojak_kucerka7	308	147	3	142	0.7256778309409888	3	41.0

Results on dataset 3

Username	TP3	TN3	FP3	FN3	q3	Pts3	Rank3
Best submission for dataset 3	583	199	1	17	0.9666254635352287	4	1.0
ambrojak_kucerka7	479	198	2	121	0.8276283618581907	4	38.0