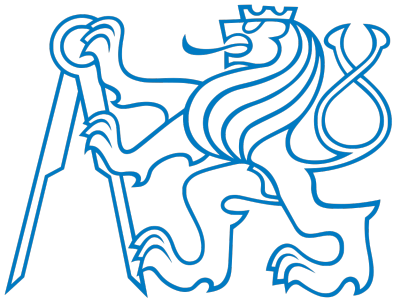
Spam Filter

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

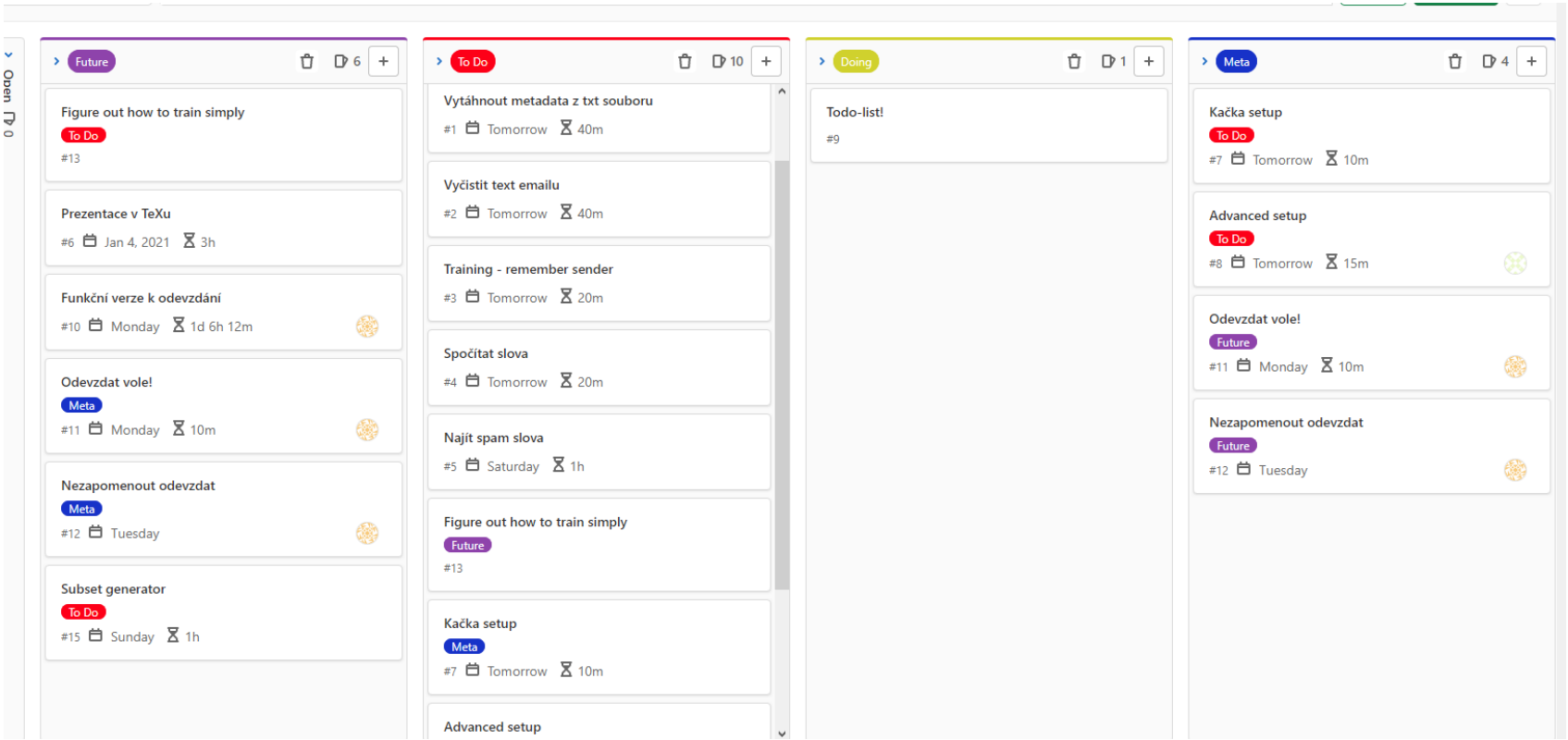
- Náš filtr funguje na principu počítání četnosti výskytu slov ve spamech a v hamech. Pak počítá jestli je slovo v testovaném emailu častěji v spamech či v hamech
- Využití metadat o emailu stejným způsobem
- Spojení obou pravidel logickým and - odstranění většiny False Positives

Testovací sada

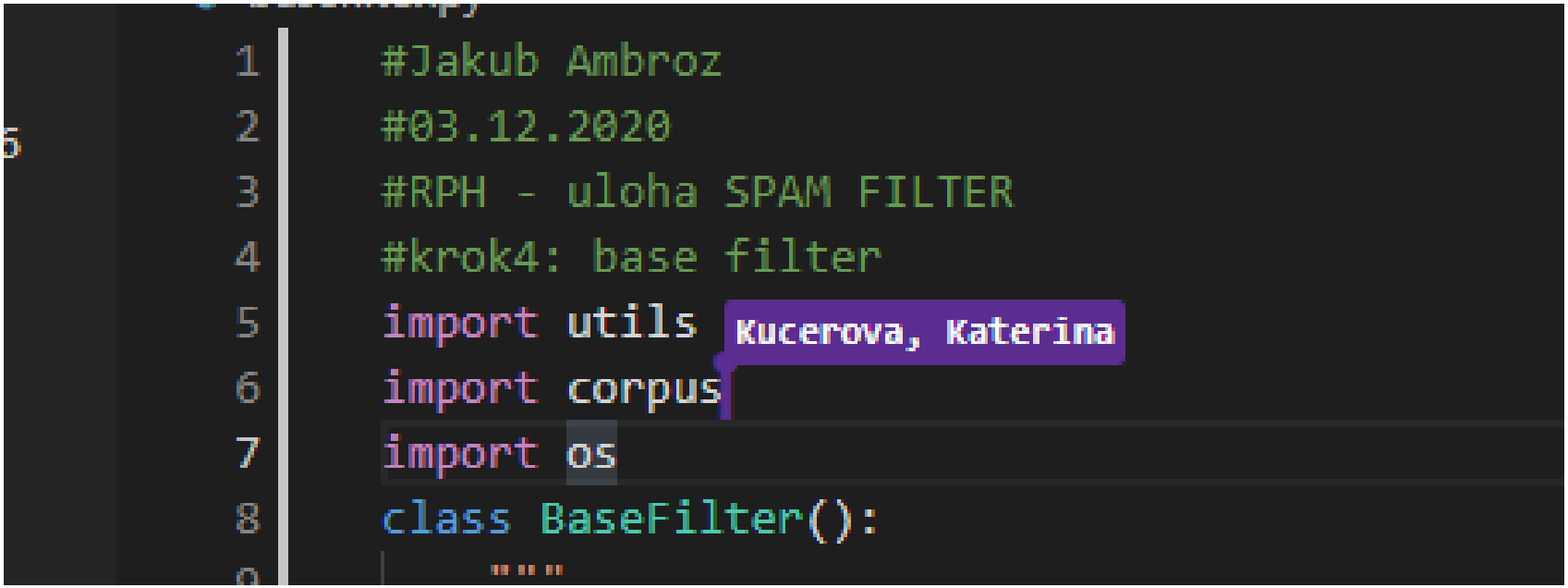
- Vytvoření náhodné trénovací a testovací. Napůl pomocí pythonu napůl příkazové řádky

```
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) <= percantage_used): # percantage 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("Succcess!\n Subsets created:\n" +
      save_path+"/train/\n"+save_path+"/test/")
print("Used dataset: " + str(number))
print("Used portion for trianing " + str(percantage_used) + "% ")
```

Spolupráce: GitLab



Spolupráce: VSCode - Liveshare



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