Warning

This page is located in archive. Go to the latest version of this course pages. Go the latest version of this page.

Spam filter: Specifikace

Na této stránce jsou popsány požadavky, jimž **musíte vyhovět!** Jejich nedodržení by mohlo vést k tomu, že vaše řešení úlohy nebude uznáno jako platné!

- Spam filtr musí fungovat v **Pythonu 3.8.x**, který bude nainstalován na strojích pro automatické hodnocení! Jinak se může stát, že vaše kódy nebudou fungovat správně!
- Je dán **předepsaný formát dat a metadat [/b211/courses/b4b33rph/cviceni/spam/data]**, která budete buď číst nebo vytvářet. Když jej nedodržíte, váš filtr nebude rozumět datům, která mu budeme předkládat při testování, nebo my nebudeme rozumět výstupům, které váš filtr vytvoří.
- Úloha má definovány 3 kontrolní body, jejichž požadavky jsou detailně popsány níže:
  1. Specifikace pro sp_eval [#odevzdanixx_sp_eval] (individuální úloha, plníte každý sám).
  2. Specifikace pro sp_filt [#odevzdanixx_sp_filt] (týmová úloha, řešíte ve dvojicích).
  3. Specifikace pro sp_prez [#odevzdanixx_sp_prez] (týmový úloha, řešíte ve dvojicích).

1. ODEVZDÁNÍ: xx_sp_eval

Individuální úloha!

Účelem prvního kontrolního bodu je zajistit, že všichni máte k dispozici funkci, která správně ohodnotí kvalitu filtru. Předmětem testování v této fázi bude pouze funkce `compute_quality_for_corpus()` (a kód, který tato funkce využívá), jejíž detailnější specifikaci najdete v kroku 3 [/b211/courses/b4b33rph/cviceni/spam/krok3#funkce_compute_quality_for_corpus].

Míra kvality filtru

"Kvalita filtru", podle níž budou přidělovány body, se bude počítat podle následujícího vzorce:

<latex>

```
$$q = \frac{TP + TN}{TP + TN + 10 \cdot FP + FN}$$
```

</latex>

Pozitivní případy (*P*) zde odpovídají mailům, které filtr označil jako spam, negativní (*N*) pak těm, které filtr označil jako korektní emaily. *FP* tedy označuje počet korektních emailů označených jako spam, *FN* označuje počet spamů označených jako korektní email. Zdůrazňujeme, že hodnoty *TP, FP, TN, FN* ve výše uvedeném vzorci představují **počty případů**, nikoli procenta.

Co odevzdat?

- **ZIP archív s vaším modulem `quality.py` a případně se všemi moduly, které tento modul importuje.**
- **Tyto soubory musí být v kořeni archívu**, nesmí být vnořeny v žádném adresáři!
- Pokud jste postupovali podle návodu, měl by váš ZIP obsahovat nejspíše soubory `quality.py`, `confmat.py`, `utils.py`, příp. další.
- **Neodevzdávejte** testy k jednotlivým krokům!

2. ODEVZDÁNÍ: xx_sp_filt

Týmová úloha!

Cílem druhého kontrolního bodu je otestovat a ohodnotit vámi vytvořený filtr. Předmětem testování bude třída `MyFilter`, jejíž detailnější specifikace najdete v kroku 6 [/b211/courses/b4b33rph/cviceni/spam/krok6#trida_myfilter].

Časové požadavky

- Vytvoření objektu filtru (metoda `__init__()`) by měla být rychlá operace trvající maximálně sekundy, spíše mnohem méně.
- Učení filtru (metoda `train()`) na několika stovkách emailů by nemělo trvat déle než desítky sekund, maximálně jednotky minut. Ve většině případů to lze zvládnout mnohem rychleji. V BRUTE bude nastaven limit na učení 5 minut.
- Aplikace filtru na nová data (metoda `test()`) by už měla být také rychlá operace trvající maximálně jednotky sekund. V BRUTE bude nastaven limit na predikci 2 minuty.

Velikost uploadu

- ZIP archiv, který obsahuje pouze kód, nebude větší než jednotky kB.
- Chcete-li se svým filtrem odevzdat i nějaká předpočítaná data, můžete, ale pak by velikost ZIP archivu neměla překročit 1 MB. Pokud myslíte, že vás tento limit příliš omezuje, proberte své

řešení se svým cvičicím.

- Kód nesmí stahovat jakákoli data z internetu.

Filtr nesmí

- Snažit se navázat spojení s jiným počítačem, jiným procesem a pod.
- Využívat knihovny třetích stran jako nltk, scikit-klearn, pandas, atp., protože na hodnoticím stroji nemusí být nainstalované. Využijte jen standardní knihovnu Pythonu. Pokud byste externí knihovny skutečně k něčemu potřebovali, **vždy to proberte se svým cvičícím**, který posoudí, zda to je z hlediska předmětu přínosné. Zkusíme následně vašemu přání vyhovět, ale nezaručujeme, že se to podaří.
- Být nepatřičně zvídavý, např. prohlížet obsah disku.
- …

Co odevzdat?

- Odevzdávat budete **ZIP archív s vaším modulem `filter.py` a všemi soubory, které váš filter potřebuje.**
- **Tyto soubory musí být v kořeni archívu**, nesmí být vnořeny v žádném adresáři!
- Pokud jste postupovali podle návodu, měl by váš archív obsahovat následující soubory:

  1. `filter.py` . Implementace vašeho filtru.
  2. `basefilter.py` . Pokud jste našli nějakou funkcionalitu, která je společná všem filtrům a extrahovali jste ji např. do třídy `BaseFilter` v modulu `basefilter.py` , od které třída vašeho filtru dědí, musíte do archívu zařadit i soubor `basefilter.py` .
  3. `corpus.py` a `trainingcorpus.py` . Pravděpodobně jste si uvědomili, že v metodě `train()` vašeho filtru s výhodou využijete třídu `TrainingCorpus` , zatímco v metodě `test()` využijete zase třídu `Corpus` . Pokud je využíváte, musíte je odevzdat v archivu spolu s vaším filtrem.
  4. `utils.py` . Vaše třída `TrainingCorpus` pravděpodobně využívá funkci `read_classsification_from_file` z modulu `utils.py` , musíte tedy do archívu zahrnout i tento soubor.
  5. případné další soubory, které váš filtr potřebuje k činnosti.

- **Neodevzdávejte** moduly, které váš filtr přímo nevyužívá, např. moduly `quality` nebo `confmat` .
- **Neodevzdávejte** testy k jednotlivým krokům!

3. REPORT A PREZENTACE: xx_sp_prez

Týmová úloha!

V této úloze je vaším úkolem splnit krok 7 [/b211/courses/b4b33rph/cviceni/spam/krok7], tj. vytvořit report a prezentaci k vašemu Spam filtru.

Co odevzdat?

Budete odevzdávat ZIP archiv s

- PDF obsahující váš report (povinná část) a
- PDF obsahující prezentaci vašeho spam filtru pro vaše spolužáky (nepovinná, ale doporučená část).

courses/b4b33rph/cviceni/spam/specifikace.txt · Last modified: 2021/12/29 11:53 by xposik