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 - krok 5

Od třídy `Corpus` odvoďte třídu `TrainingCorpus`, která bude obalovat korpus se známým ohodnocením emailů, tedy korpus, který je možno použít k učení filtru.

Testy [/b211/courses/b4b33rph/cviceni/spam/unit_testing] ke kroku 5:

- samostatně test5_trainingcorpus.zip [/b211/_media/courses/b4b33rph/cviceni/spam/test5_trainingcorpus.zip] nebo
- společně se všemi předchozími testy test5_all.zip [/b211/_media/courses/b4b33rph/cviceni/spam/test5_all.zip].

Třída `TrainingCorpus` není nijak povinná a její implementace není pevně dána. Implementujte jen ty metody, které se vám budou hodit. Dodávané testy kontrolují všechny níže uvedené metody - rozhodnete-li se nějaké z nich neimplementovat, smažte (nebo zakomentujte) příslušné testy ve třídě `TrainingCorpusTest`.

Příprava

- Rozmyslete si, co by měla třída vašeho trénovacího korpusu (`TrainingCorpus`) umět, aby vám usnadnila učení filtru. T.j. co musí umět navíc vzhledem k třídě `Corpus`.

Korpus trénovacích dat

Úkol:

- V modulu `trainingcorpus.py` vytvořte třídu `TrainingCorpus`.

K čemu nám to bude?

- Třída `TrainingCorpus` bude umožňovat jednodušší tvorbu učicích se filtrů, pokud se nakonec rozhodneme učení implementovat. Bude obalovat adresář s emaily včetně jejich správné klasifikace ze souboru `!truth.txt`.

Specifikace

Specifikace tohoto úkolu není pevná, záleží na vás, jaké metody se vám budou hodit. Následující berte jako inspiraci. (Unit testy ale tyto metody testují.)

- Třídu vybavte metodou `get_class()`, jejímž vstupem bude název souboru s emailem a výstupem buď kód `OK` nebo `SPAM`.
- Třídu vybavte metodami `is_ham()` a `is_spam()`, jejichž vstupem bude opět název souboru s emailem a výstupem pravdivostní hodnoty `True` nebo `False`.
- Třídu vybavte metodami `spams()` a `hams()`, které budou generátory a budou fungovat podobně jako metoda `emails()` ve třídě `Corpus`, tj. budou postupně vracet jména souborů a těla všech spamů, resp. všech hamů jako řetězce.
- …

Záleží jen na vás, zda některé z metod implementujete nebo zda se rozhodnete pro jiný přístup.

courses/b4b33rph/cviceni/spam/krok5.txt · Last modified: 2018/07/17 13:25 (external edit)