Warning

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

Formát dat

V této úloze se pracuje s množinou (množinami) emailových zpráv, které mohou být opatřeny meta-informací. Takové množině se často říká korpus [https://en.wikipedia.org/wiki/Text_corpus]. Meta-informacemi budou v našem případě údaje o tom, zda jeden každý email skutečně je nebo není spamem, a o tom, zda spam filtr odhadl, že jeden každý email je spamem.

Máte k dispozici 2 sady dat [/b211/_media/courses/b4b33rph/cviceni/spam/spam-data-12-s75-h25.zip] pocházející ze stejného zdroje.

Domluvme se následovně: emailový korpus pro nás bude představovat

- adresář, v němž budeme všechny soubory považovat za emaily, až na
- soubor `!truth.txt`, který bude obsahovat na každém řádku jméno souboru a informaci o tom, zda email uložený v daném souboru je či není spam, a
- soubor `!prediction.txt`, který bude obsahovat na každém řádku jméno souboru a informaci o tom, zda email uložený v daném souboru je spam filtrem považován za spam nebo ne. (Tento soubor samozřejmě v datech, které od nás dostáváte, nenajdete - tento soubor bude vytvářet váš spam filtr.)

Tyto dva soubory v adresáři být mohou, ale také nemusí:

1. Spam filtr sám při **klasifikaci emailů** žádný z těchto souborů nepotřebuje. Bude ale muset umět vytvořit soubor `!prediction.txt`, který bude obsahovat jeho odhady.
2. Při **vytváření (učení) spam filtru** budete potřebovat tzv. *trénovací korpus*, tj. adresář, v němž je soubor `!truth.txt`. Jinak byste nevěděli, které emaily jsou spam a které ne.
3. Budeme-li chtít **vyhodnotit kvalitu spam filtru**, budeme potřebovat oba soubory - `!truth.txt` i `!prediction.txt`. Porovnáním údajů v těchto dvou souborech zjistíme, jak moc se předpovědi filtru shodují se skutečností. (Vlastní emaily zde vlastně ani nebudeme potřebovat).

Obsah souborů s emaily

S obsahem souborů lze pracovat jako s prostým textem, aniž byste předpokládali, že soubory mají nějakou vnitřní strukturu.

Pokud ale chcete jejich strukturu využít (což samozřejmě můžete), vězte, že tyto soubory měly odpovídat normě RFC5322 [https://tools.ietf.org/html/rfc5322] (příp. RFC2822 [http://tools.ietf.org/html/rfc2822.html]). V ní se v článku 3.6 [https://tools.ietf.org/html/rfc5322#section-3.6] píše:

```
The only required header fields are the origination date field and
the originator address field(s).  All other header fields are
syntactically optional.
```

Nespoléhejte tedy na to, že všechny emaily budou mít subjekt, nebo další pole!!!

V tomtéž dokumentu se také dočtete:

```
It is important to note that the header fields are not guaranteed to
be in a particular order.  They may appear in any order, ...
```

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