Warning

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

Spam filtr

Filtrování spamu je velmi praktická úloha s velkým reálným dopadem. Je to představitel klasifikačního problému, s nimiž se setkáváme ve strojovém učení, a lze na ní ukázat jejich typické rysy. Vyžaduje práci s nestrukturovanými daty (textem). Lze ji řešit velkou škálou metod - od jednoduchých až po velmi propracované.

- Úvod do problému filtrování spamu [/b211/courses/b4b33rph/cviceni/spam/uvod].
- Hodnocené části úlohy a jejich bodování [/b211/courses/b4b33rph/cviceni/spam/hodnoceni].
- Specifikace [/b211/courses/b4b33rph/cviceni/spam/specifikace], jichž se **musíte** držet.
- Doporučené kroky (1 [/b211/courses/b4b33rph/cviceni/spam/krok1], 2 [/b211/courses/b4b33rph/cviceni/spam/krok2], 3 [/b211/courses/b4b33rph/cviceni/spam/krok3], 4 [/b211/courses/b4b33rph/cviceni/spam/krok4], 5 [/b211/courses/b4b33rph/cviceni/spam/krok5], 6 [/b211/courses/b4b33rph/cviceni/spam/krok6]), jichž se držet nemusíte, pokud nejsou součástí výše uvedených specifikací.
- Data [/b211/courses/b4b33rph/cviceni/spam/data] potřebná k řešení úlohy.

Přínos úlohy pro studenty

- Procvičit si **objektové programování v Pythonu** a práci s **nestrukturovanými daty** (textem).
- Vyzkoušet **řešení složitější úlohy**, která vyžaduje spolupráci několika tříd.
- Seznámit se s prvky automatického testování softwaru - s **unit testy**.
- Na praktické úloze se seznámit se **základními principy strojového učení** (jak ohodnotit kvallitu filtru, trénovací/testovací data, …).
- Zažít **pocit, že zvládnu vyřešit netriviální úlohu**.
- V závěrečném **srovnání** zjistit, zda lépe fungují **člověkem navržené filtry** nebo **filtry schopné sami se učit z dat**.

Cíle úlohy

Vaším prvořadým úkolem v této úloze nebude vytvoření co nejkvalitnějšího spam filtru, jak by se mohlo zdát. Metody, které by vám v tom mohly pomoci, ještě neznáte a dozvíte se o nich v dalším

průběhu studia. Vaším úkolem je především

- pochopit, oč v této úloze jde, úlohu analyzovat a dekomponovat,
- vytvořit si sadu funkcí či tříd v jazyce Python, které vám umožní snadno použít spam filtr (až nějaký budete mít) a vyhodnotit jeho kvalitu, příp. porovnat dva a více spam filtrů, a
- vytvořit jednoduchý (třeba i velmi primitivní) spam filtr, který bude možno použít v rámci vytvořeného “frameworku”.

courses/b4b33rph/cviceni/spam/start.txt · Last modified: 2020/11/09 14:31 by xposik