KIV/UIR - Semestrální práce pro ak. rok 2019/20

Klasifikace dokumentů

Ve zvoleném programovacím jazyce navrhněte a implementujte program, který umožní klasifikovat textové dokumenty do tříd podle jejich obsahu, např. počasí, sport, politika, apod. Při řešení budou splněny následující podmínky:

- Použijte data z českého historického periodika „Posel od Čerchova“, která jsou k dispozici na https://drive.google.com/drive/folders/1mQbBNS43gWFRMHDYdSkQug47cuhPTsHJ?usp=sharing. V původní podobě jsou data k dispozici na http://www.portafontium.eu/periodical/posel-od-cerchova-1872?language=cs.
- Pro vyhodnocení přesnosti implementovaných algoritmů bude NUTNÉ vybrané dokumenty ručně označkovat. Každý student ručně anotuje 10 stran zadaného textu – *termín 31.3.2020*. Za dodržení termínu obdrží student **bonus 10b**.
- Přiřazení konkrétních textů jednotlivým studentům spolu s návodem na anotaci a příklady je uloženo spolu s daty na výše uvedené adrese, konkrétně:
  - **0** - vzorová složka (takhle by měl výsledek vypadat)
  - **1, 2, .. , 15, 101, 102, ..** - data k anotaci
  - **prirazeni_souboru_studentum.xlsx** - určení, jaké soubory má jaký student anotovat. Až budete mít anotaci hotovou, doplňte sem informaci.
  - **Anotační příručka** - návod, jak články anotovat.
  - **Klasifikace dokumentů - kategorie.xlsx** - seznam kategorií k anotaci s příklady.
  - **sem_prace20.pdf** - Zadání semestrální práce
- implementujte alespoň tři různé algoritmy (z přednášek i vlastní) pro tvorbu příznaků reprezentující textový dokument.
- implementujte alespoň dva různé klasifikační algoritmy (klasifikace s učitelem):
  - Naivní Bayesův klasifikátor
  - klasifikátor dle vlastní volby
- funkčnost programu bude následující:
  - spuštění s parametry:
    název_klasifikátoru
    soubor_se_seznamem_klasifikačních_tříd,
    trénovací_množina, testovací_množina,
    parametrizační_algoritmus, klasifikační_algoritmus,
    název_modelu
    program natrénuje klasifikátor na dané trénovací množině, použije zadaný parametrizační a klasifikační algoritmus, zároveň vyhodnotí úspěšnost klasifikace a natrénovaný model uloží do souboru pro pozdější použití (např. s GUI).

  - spuštění s jedním parametrem:
    název_klasifikátoru
    název_modelu
    program se spustí s jednoduchým GUI a uloženým klasifikačním modelem. Program umožní klasifikovat dokumenty napsané v GUI pomocí klávesnice (resp. překopírované ze schránky).

- ohodnoťte kvalitu klasifikátoru na dodaných datech, použijte metriku přesnost (accuracy), kde jako správnou klasifikaci uvažujte takovou, kde se klasifikovaná třída nachází mezi anotovanými. Otestujte všechny konfigurace klasifikátorů (tedy celkem 6 výsledků).

Poznámky:

- pro vlastní implementaci není potřeba čekat na dokončení anotace. Pro průběžné testování můžete použít korpus současné češtiny, který je k dispozici na http://ctdc.kiv.zcu.cz/ (uvažujte pouze první třídu dokumentu podle názvu, tedy např. dokument 05857_zdr_ptr_eur.txt náleží do třídy „zdr“ - zdravotnictví).

- další informace, např. dokumentace nebo forma odevzdávání jsou k dispozici na CW pod záložkou *Samostatná práce*.

Bonusové úkoly:

- vyzkoušejte již nějakou hotovou implementaci klasifikátoru (scikit-learn, Weka, apod.) a výsledky srovnejte s Vaší implementací [až 10b navíc].

- vyzkoušejte shlukování (klasifikaci bez učitele, např. k-means) a výsledky porovnejte s výsledky klasifikace s učitelem [až 10b navíc].

- implementujte navíc klasifikační algoritmus založený na neuronové síti typu MLP s využitím knihoven Keras a Tensorflow [až 10b navíc].

- vyzkoušejte klasifikaci anglických dokumentů, korpus Reuters je k dispozici na adrese http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html [až 20b navíc].