KIV/UIR - Semestrální práce pro ak. rok 2021/22

Automatická klasifikace dialogových aktů

Ve zvoleném programovacím jazyce navrhněte a implementujte program, který umožní v komixovém dialogu klasifikovat věty (nebo jejich části) do tříd podle jejich obsahu, např. rozkaz, otázka zjišťovací (wh-question), odpověď, apod. Tyto věty odpovídají tzv. dialogovým aktům a mají důležitou roli pro řízení dialogu, protože určují funkci věty v dialogu. Například funkce otázky je žádost o nějakou informaci, naproti tomu, funkcí sdělení je poskytnutí požadované informace. Při řešení budou splněny následující podmínky:

- Datová sada viz https://drive.google.com/drive/folders/1ZsEPcSh0MlFU-9iQrib-8eWI1bCv9ygp?usp=sharing obsahuje
  - Testovací (Test) a trénovací (Train) množiny
  - Trénovací množinu anotují sami studenti podle anotační příručky.
- Pro trénování implementovaných algoritmů bude NUTNÉ vybrané dokumenty ručně označkovat. Každý student ručně anotuje 40 vybraných komixů – *termín 15.4.2022*. Za dodržení termínu obdrží student **bonus 10b**.
- Přiřazení konkrétních komixů jednotlivým studentům spolu s návodem na anotaci a příklady bude uloženo spolu s daty na výše uvedené adrese.
- 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 věty (dialogové akty) 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 implementaci parametrizačních / klasifikačích algoritmů není možné používat hotové knihovní funkce!
  - Pro vlastní implementaci není potřeba čekat na dokončení anotace. Pro průběžné testování můžete použít testovací korpus (rozdělit na trénovací a testovací množinu).
  - 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 na vyžádání [až 20b navíc].