Zespołowy projekt specjalnościowy 30.10.2019

Podział zadań

Paweł Lorek

- UWAGA: dużo rzeczy nie jest tutaj dopowiedzianych. Proszę dopytywać na zajęciach, tak by było jasno co trzeba robić.
- UWAGA: lista może ulegać zmianie
- Repozytorium: https://github.com/lorek/ZPS2019
- Każdy zespół jest zobowiązany dostarczyć min. jeden zestaw danych do klasyfikacji i umieścić go w `datasets/`
- Każdy zespół musi w pliku `docs/zespol{nr}_doc.txt` umieścić informacje:
  - jakie dane potrzebuje na wejściu (format, najlepiej info gdzie powinny się znajdować itp)
  - jak używa się ich narzędzia (co ono robi, jakie przyjmuje parametry, przykładowe wywołania itp)
- **Zespół 1**: Główne zadanie: przetworzenie danych (tekstów) na liczby. Jeden tekst = wektor o ustalonym wymiarze d (to może być parametr).

  Trzy metody:

  - Bag of Words (BoW) (można onegram lub bigram)
    `sklearn.feature_extraction.text.CountVectorizer`
  - TF-IDF
    `sklearn.feature_extraction.text.TfidfVectorizer`
  - Word2Vec. Proponuje uzycie bibliteki `gensim`.
    W katalogu `scripts` umieściłem dwa przykładowe pliki:
    * `word2vec_example.py` – tworzy sam słownik z wbudowanych zestawów danych
    * `word2vec_example_google.py` – używa już stworzonego/wyuczonego słownika, należy go wcześniej pobrać z https://github.com/eyaler/word2vec-slim/blob/master/GoogleNews-vectors-negative300-SLIM.bin.gz

Wszystkie poniższe metody redukcji wymiaru oraz klasyfikatory dostępnę są w bibliotece `sklearn`.

- **Zespół 2:** Techniki redukcji wymiaru
  - PCA (Principal component Analysis), Kernal PCA (z jakimś nieliniowym jądrem)
  - NMF (Non-Neagative Matrix Factorization)
  - ICA (Independent Component Analysis)
  - t-SNE (t-Distributed Stochastic Neighbor Embedding)
  - LDA (Lindear Discriminant Analysis – typowo klasyfikator, ale może być używany jako technika redukcji wymiaru), QDA (Quadriatic Discriminant Analysis)

    Uwaga: Skrypt powinien mieć opcję LDA i QDA, jednak sama prezentacja algorytmu to zadanie Zespołu 3.

- **Zespół 3:** Klasyfikacja
  - Naiwny klasyfikator Bayesa ("zwykly" oraz wersja GaussianNB, do implementacji tylko ta druga)
  - k-nearest neighbor (kNN)
  - Drzewa decyzyjne (np. C4.5), Random Forest
  - LDA (Linear Discriminant Analysis), QDA (Quadriatic Discriminant Analysis)

- **Zespół 4:** Klasyfikacja
  - SVM (Support Vector Machine), jądro liniowe i nieliniowe
  - Bagging, AdaBoost
  - Logistic Regression classifier

- **Zespół 5:** Testowanie narzędzi poprzednich zespołów. Cel: uzyskanie najlepszej klasyfikacji na każdym zestawie danych
  - Może jeszcze dodatkowo ... klasyfikator?