Eksploracja danych w Internecie

(Kod przedmiotu: 02 64 5054 00)

Laboratorium 4-5

Zadanie: Zadaniem jest skorzystanie z technik przekształcenia tekstu w wektory cech (*bag of words, tf-idf*) w celu klasteryzacji tekstu (*k-means*). W zadaniu tym wykorzystaj dowolny web scrapper (np. *Scrapy*) do utworzenia zbioru uczącego.

1. Utwórz zbiór (bazę) danych:
   a) wybierz 5 niepowiązanych ze sobą tematycznie stron internetowych (np. serwis sportowy, programistyczny, o grach, dla dzieci, finansowy, filmowy, literaturowy, itp.),
   b) następnie pobierz po 100 podstron (z tej samej domeny) dla każdej z nich.

   Efektem jest zbiór obiektów (np. zbiór plików lub csv), gdzie każdy z nich posiada:

   - tekst (oczyszczona zawartość strony),
   - źródłowy adres URL (adres podstrony, np. *https://www.pythongasm.com/introduction-to-scrapy/*),
   - kategoria (nazwa domeny, np. *pythongasm.com*).

kategoria	URL	tekst
pythongasm.com	https://www.pythongasm.com/introduction-to-scrapy	...
pythongasm.com	https://www.pythongasm.com/major-differences-between-python-2-3/	...
...	...	...

2. Tekst zamień na wektor cech korzystając z worka słów (*bag-of-words*) oraz odwrotną częstość w dokumentach (*tf-idf*).

kategoria	URL	tekst	bow	tfidf
pythongasm.com	https://www.pythongasm.com/introduction-to-scrapy	...	[3, 0, 0, 1, …]	[0.45, 0.20, …]
pythongasm.com	https://www.pythongasm.com/major-	...	[0, 5, 1, 0, …]	[0.08, 0.36, …]

	differences-between-python-2-3/			
...	...	...	...	...

3. Zastosuj algorytm *k-means* osobno dla *bow* i *tfidf*, a następnie wyświetl wyniki w formie tabeli i porównaj. Czy oba algorytmy dają identyczne wyniki?

Uwagi:

- Zadanie powinno składać się z osobnych modułów (jeżeli nie jest zrobione w jupyter notebook), aby możliwe było uruchomienie wybranego z nich.
- Wyniki (wyjście programu) powinny być czytelne i jasno opisane, aby nie było wątpliwości, co jest zaprezentowane.
- W kodzie źródłowym powinny znajdować się komentarze, wskazujące w którym miejscu jest implementacja poszczególnych punktów zadania.

Przydatne linki:

- Scrapy: scrapy.org
- Bag of words (bag of n-grams – zobacz parametr analyzer): https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html
- Tf-idf: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html
- K-means: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- Confusion matrix: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html