Inteligencja obliczeniowa

Laboratorium 7: Zgłębianie tekstu.

Zgłębianie tekstu to obecnie bardzo modna dziedzina. Internet jest pełny tekstów, danych, które nie są schematycznie uporządkowana. Wydobywanie informacji z tekstu jest trudne. Wymaga on z reguły wielostopniowej obróbki (usuwanie znaków, liczb, słów nic nie wnoszących do treści), by był zdatny do analizy.

Na laboratoriach spróbujemy wykorzystać kilka technik do podstawowego zgłębiania tekstu.

Zadanie 1

Ściągnij załączone na stronie pliki (w paczce zip)

- language.txt (strona z wikipedii definiująca język)
- computer.txt (strona z wikipedii definiująca komputer)
- programming-language.txt (strona z wikipedii definiująca język programowania)
- natural-computing.txt (strona z wikipedii definiująca algorytmy inspirowane naturą)
- life.txt (strona z wikipedii definiująca życie)
- genetic-algorithm.txt (strona z wikipedii definiująca algorytm genetyczny)

Korzystając z samouczka:
https://rstudio-pubs-static.s3.amazonaws.com/265713_cbef910aee7642dc8b62996e38d2825d.html
dokonaj analizy tych dokumentów w R.

a) Zainstaluj potrzebne paczki i załąduj dokumenty (rozdział **Loading Texts**)

b) Dokonaj obróbki textu (rozdział **Preprocessing**).

Proszę zwłaszcza zwrócić uwagę na akapit „**Removing stopwords**”. Jakie to słowa? Wyświetl odpowiednią komendą (wykomentowaną w samouczku). Niestety lista nie jest pełna. Jakie słowa-śmieci znajdą się na naszej liście słów kluczowych („also”? „can”? „used”?). W Nastepnym akapicie „**Removing particular words**” dodaj te słowa do usunięcia. Rób to tak długo, aż będziesz zadowolony z efektu (lista może być nieco dłuższa).

c) Stwórz macierz **dtm** (Document-Term-Matrix), jak podano w rozdziałach **Stage the Data, Explore your Data, Focus!**. Co oznaczają liczby w kolumnach, a co liczby w wierszach tej macierzy?

d) Przechodzimy do części badawczej. Jaki słowa występują najczęściej w dokumentach? Przedstaw wyniki w formie tekstowej i na wykresach. Rozdział: **Word Frequencies, Plot Word Frequencies**.

e) Sprawdź jaka korelacja zachodzimi między termami: computer, life, programming, language. Rozdział **Term Correlations**.

f) Stwórz chmury słów (word clouds). Ustaw min.freq na taką wartość, by chmura była interesująca. Rozdział **Word Clouds**.

g) Słowa (termy) są podobne, gdy często wystepują w tym samym dokumencie. Kuszące jest, by sprawdzić czy algorytmy do klasteryzacji pogrupują je w klastry odpowiadające dokumentom, w których wystepują.
Stwórz dendrogram i sprawdź, czy słowa siedzące na wspólnej gałęzi są ze sobą powiązane

tematycznie.

Nastepnie korzystając z algorytmu K-Means, dokonaj grupowania słów na kilka klastrów. Sprawdź jak algorytm działa dla k=2,3,4,5,6.

Zadanie 2

Wykorzystaj macierz **dtm** z poprzedniego zadania do znalezienia podobieństwa dokumentów. Wówczas łatwo będzie odpowiedzieć na takie pytania jak: „Czy artykuł o algorytmie genetycznym jest bardziej podobny do artykułu o życiu czy programowaniu?”

Do badania podobieństwa dokumentów wykorzystaj podobieństwo kosinusowe: https://en.wikipedia.org/wiki/Cosine_similarity