**  
  
Eksploracja danych tekstowych - laboratoria 5: Normalizacja danych tekstowych**

*Do realizacji zadań wykorzystaj środowisko jupyter notebook. Po zakończeniu zadań wyślij odpowiedzi na adres michal.zarnecki@cdv.pl.*

1. Wybierz i pobierz zbiór danych zawierających dane tekstowe ze strony: https://www.kaggle.com/datasets?tags=13204-NLP

2. Dokonaj analizy wybranego zestawu danych. Wyznacz:

1. listę 20 najczęściej występujących słów
2. listę 20 najrzadziej występujących słów
3. średnią/maksymalną/minimalną liczbę słów
4. średnią/maksymalną/minimalną liczbę słów rozpoczynających się wielką literą
5. średnią/maksymalną/minimalną liczbę znaków interpunkcyjnych
6. średnią/maksymalną/minimalną liczbę fraz numerycznych

3. Skorzystaj z notatnika Jupyter "Eksploracja danych tekstowych - Jupyter 5-6 Normalizacja i tokenizacja tekstu.ipynb" i skopiuj kod dotyczący analizy i normalizacji tekstu, następnie zastosuj go na swoim zbiorze danych.

4. Dokonaj normalizacji tekstu (lematyzajca, usunięcie liczb, usunięcie znaków interpunkcyjnych, usunięcie rzadko występujących słów).