# Laboratorium 3

Proszę przesyłać rozwiązania w pliku tekstowym nazwiskoLab3.py.
W bibliotece NLTK mamy dostęp do wybranych korpusów.
Poniżej informacja o wybranych korpusach dostępnych w NLTK:

| korpus | nazwa w nltk | info |
|---|---|---|
| Brown Corpus | brown | 1.15 M słów, ze znacznikami (tagi) |
| CoNLL 2000 | conll2000 | 270 k słów ze znacznikami |
| Gutenberg | gutenberg | 18 dokumentów |
| Inaugural Address Corpus | inaugural | przemówienia prezydentów USA |
| Movie Reviews | movie_reviews | 2k recenzji filmów z wartoscią wydźwięku |
| Penn Treebank | treebank | zdania z drzewami rozkładu |
| Reuters Corpus | reuters | 10 k dokumentów z kategoriami |
| SentiWordNet | sentiwordnet | zbiór synsetów |
| Stopwords Corpus | stopwords | 24 k słów w 11 językach |
| Wordnet 3.0 (English) | wordnet | 145k synsetów |

Dla korpusów okreslone są metody:
fileids(), fileids([categories]), categories(), categories([fileids]), raw(), words(), sents(). Sprawdzic ich znaczenia.
Zadania.

1. Sprawdzić nazwy plików w korpusie Gutenberg.
2. Sprawdzić nazwy plików w korpusie Inaugural Address Corpus.
3. Jakie kategorie występują w korpusie Movie Reviews.
4. Wydrukuj zdania występujace w korpusie Inaugural Corpus Address w pliku '1909-Taft.txt'.
5. Napisz skrypt, służący do sprawdzenia jak często w korpusie Brown w kategorii 'adventure' występują słowa 'mountains', 'ocean', 'Bungee jump'.

6. Wyszukaj 10 najczęsciej występujących słów w korpusie 'inaugural'.

7. Zdefiniuj funkcję, która dla tekstu zlicza jaki procent stanowią słowa spoza listy stopwords. Zastosuj funkcję do każdego z tekstów z NLTK Book.

8. Sprawdź jaki jest wydźwięk słów: journalist, writer, actor, singer.

9. Zaproponuj miarę podobienstwa dla rzeczowników i sprawdź, czy dla poniższej listy (w kolejnosci malejącego podobienstwa) uporządkowanie będzie zgodne z poniższym:

   - boy-lad
   - journey-voyage
   - coast-hill
   - monk-slave
   - food-fruit
   - journey-car

10. Dla korpusu Gutenberg podać dla kazdego tekstu następujące dane: nazwa pliku, srednia liczba znaków w słowie, srednia liczba słów w zdaniu, srednia liczba powtórzeń tego samego słowa.

11. Polecenie

    ```
    nltk.corpus.brown.tagged_words(tagset='universal')
    ```

    zwraca

    ```
    [('The', 'DET'), ('Fulton', 'NOUN'),  ('Grand', 'ADJ'),
    ('Jury', 'NOUN'), ('said', 'VERB'), ('Friday', 'NOUN'), ..
    ```

    Dla cr09 z korpusu Browna wylicz częstosci wystąpień słów z poszczególnymi znacznikami z listy uproszczonej. Jakie znaczniki są najczęstsze?