# TIMKoD – Lab 2 – Przybliżanie języka naturalnego – kontunacja

14 marca 2018

## Opis pliku z zadaniami

Wszystkie zadania na zajęciach będą przekazywane w postaci plików `.pdf`, sformatowanych podobnie do tego dokumentu. Zadania będą różnego rodzaju. Za każdym razem będą one odpowiednio oznaczone:

- Zadania do wykonania na zajęciach oznaczone są symbolem $\triangle$ – nie są one punktowane, ale należy je wykonać w czasie zajęć.
- Punktowane zadania do wykonania na zajęciach oznaczone są symbolem $\diamond$ – należy je wykonać na zajęciach i zaprezentować prowadzącemu, w wypadku nie wykonania zadania w czasie zajęć lub nieobecności, zadania staje się zadaniem do wykonania w domu ($\star$).
- Zadania do wykonania w domu oznaczone są symbolem $\star$ – są one punktowane, należy je dostarczyć w sposób podany przez prowadzącego i w wyznaczonym terminie (zwykle przed kolejnymi zajęciami).
- Zadania programistyczne można wykonywać w dowolnym języku programowania, używając jedynie biblioteki standardowej dostępnej dla tego języka.

## Cel zajęć

Na tych zajęciach kontynuujemy przybliżanie języka angielskiego. Jednak, zamiast tworzyć źródła operujące na znak, przejdziemy do źródeł używających całych słów jako symboli.

## Przygotowanie do zajęć

- Do wykonania zadań potrzebne będą korpusy tekstowe, które można pobrać z
  `http://www.cs.put.poznan.pl/kjasinska/lectures/timkod/data/lab1`
- Pliki są znormalizowane, zawierają jedynie 26 małych liter alfabetu łacińskiego, cyfry i spacje.
- Przygotuj funkcję do wczytywania pliku do pamięci (skopiuj z poprzednich zajęć).

# 1 Częstość słów △

## Treść

Zadanie polega na policzeniu częstości występowania słów w angielskim tekście. Jakie słowa występują najczęściej i jaki procent wszystkich słów stanowią?

Podobno przeciętny Polak zna 30 tysięcy słów, a posługuje się tylko 20. procentami z tego zbioru, co daje tylko 6 tysiący. Czy w Polsce panuje ubóstwo językowe? Sprawdź jaki procent "wiedzy" z Wikipedii umiałby przekazać przeciętny Polaka, gdyby jego elokwencje przełożyć na grunt języka angielskiego.

Policz jaki procent wszystkich słów stanowi zbiór 30. tysięcy najpopularniejszych słów, a jaki procent stanowi zbiór 6. tysięcy.

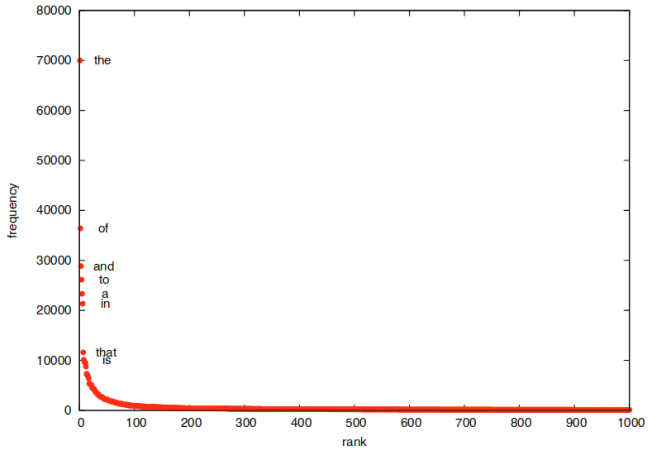
## Częstość słów – prawo Zipfa

Rysunek 1: George Kingsley Zipf (1902 - 1950)

Rozkład występowania słów jest obiektem badań lingwistyki statystycznej już od ponad 80 lat. Rozkład ten przybliża prosta formuła matematyczna znana jako prawo Zipfa:

$$f(r) \propto \frac{1}{r^{\alpha}}$$

gdzie $f(r)$ to częstość występowania w korpusie, $\propto$ oznacza proporcjonalność, $r$ to ranga słowa odpowiadająca częstości jego występowania względem innych słów w korpusie, $\alpha$ to stała skalująca, równa ok. 1.



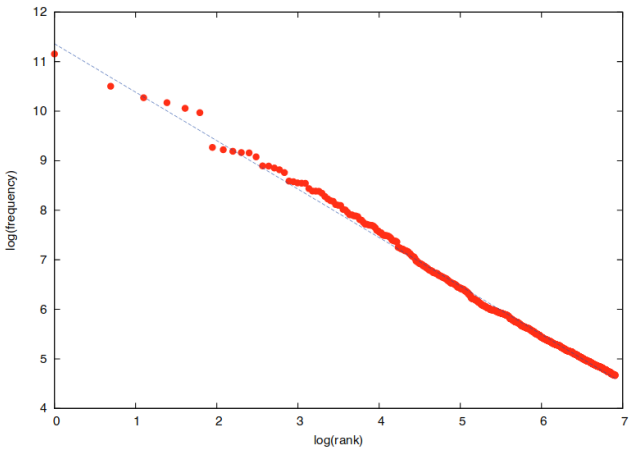

Rysunek 2: Liczność słów w próbce milionowa słów języka angielskiego

Fenoment ten odkryty przez Harvardzki lingwistę George Kingsley Zipf w 1936 roku dotyczy wszystkich języków i do dzisiaj nie znalazł jednoznacznego wyjaśnienia.

Jednak ta zależność nie dotyczy jedynie liczności słów w języku.

### Rozkład i zasada Pareto

Prawo Zipfa jest dyskretną formą rozkładu Pareto, nazwanego tak na cześć włoskiego ekonomisty Vilfreda Pareto, który w 1906 poczynił sławną obserwację, że ówcześnie 80% procent ziemi we Włoszech było w rękach 20%




Rysunek 3: Liczności słów z Rysunku 1 na skali logarytmicznej

populacji kraju. Dlatego rozkład Pareto jest bardziej znany jako zasada Pareta albo zasada 80-20, która mówi, że można się spodziewać, że 20% badanych obiektów związanych jest z 80% pewnych zasobów. A więc rozkład ten musi występować w wielu dziedzinach takich jak fizyka, biologi, naukach społecznych. I faktycznie występuje zadziwiająco często, przykładów rozkładów zbliżonych do rozkładu Pareto:

- wielkość posiadanego majątku przez ludzi,
- wielkość populacji miast (prawo Gibrata),
- ruch na stronach internetowych,
- wielkości przesyłanych plików przez internet,
- ilość sprzedanych egzemplarzy poszczególnych książek,
- częstość występowania poszczególnych nazwisk,
- liczba cytowań artykułów naukowych (również prawo Bradforda i prawo Lotka),
- pojemności złóż surowców,
- wielkość kraterów na księżycu, jak i meteorytów,
- intensywność rozbłysków słonecznych,

... i wiele innych. Jednakże obecnie uznaje się, że istnieje wiele różnych mechanizmów odpowiadających za występowanie rozkładu Pareto w tak wielu przypadkach. Poniżej rozważmy dwie teorie.

## Rozkład wykładniczy

Jedną z pierwszych teorii na częste występowanie rozkładu Pareto (zaproponowaną przez Benoit Mandelbrota) jest jego powiązanie z innym znacznie częściej występującym rozkładem wykładniczym.

Jeśli użyjemy naszego przybliżenia 0 rzędu z poprzednich zajęć (ciąg losowo generowanych znaków z równym prawdopodobieństwem) i policzymy częstości słów, okaże się, że rozkład słów również będzie przypominał prawo Zipfa.

Rzecz w tym, że na poprzednich zajęciach już doszliśmy do wniosku, że język naturalny jest bardzo daleki od losowych ciągów znaków.

## Preferencja przywiązania (Preferential attachment)

Innym popularnym wyjaśnieniem częstego występowania rozkładu Pareto jest proces nazywany preferencja przywiązania, w którym pewna ilość jakiegoś dobra jest rozdzielana pomiędzy odbiorców, proporcjonalnie do tego ile już posiadają. Ci, którzy posiadają dużo danego dobra, dostaną go więcej niż ci, którzy posiadają go mało. Bogacze stają się bogatsi, popularne staje się popularniejsze. W naszym kontekście słowo, które zostało użyte raz, ma większe szanse zostać użyte ponownie. Pojedyncze dyskusje, artykuły, książki często dotyczą konkretnego tematu. Jeśli w obrębie jednej dyskusji zostanie użyte 1 słowo, z dużym prawdopodobieństwem będziemy je powtarzać do momentu zmiany tematu.

**Zrodła**


- https://en.wikipedia.org/wiki/Zipf%27s_law
- https://en.wikipedia.org/wiki/Power_law
- https://en.wikipedia.org/wiki/Pareto_principle
- https://en.wikipedia.org/wiki/Principle_of_least_effort
- https://en.wikipedia.org/wiki/Preferential_attachment
- https://colala.bcs.rochester.edu/papers/piantadosi2014zipfs.pdf
- https://arxiv.org/pdf/cond-mat/0412004.pdf
- http://www.ling.upenn.edu/~ycharles/sign708.pdf

## 2 Przybliżenie pierwszego rzędu △

### Treść

Używając wyliczonych prawdopodobieństw w poprzednim zadaniu, wygenerują ciąg słów – przybliżenie pierwszego rzędu.

# 3 Przybliżenia na podstawie źródła Markova *10pt*◇

## Treść

Wygeneruj przybliżenie języka angielskiego na podstawie źródła Markova pierwszego rzędu na słowach (źródła, gdzie prawdopodobieństwo następnego symbolu zależy od 1. poprzedniego). ***(3pt)***

Implementacja powinna opierać się na łańcuchu Markova, nie zaś na zaproponowanej przez Shannona metodzie Monte Carlo, która to będzie generować słowa z prawdopodobieństwem odbiegającym od rzeczywistego prawdopodobieństwa warunkowego. Dlaczego?

Następnie zrób to samo dla źródła Markova drugiego rzędu (źródła, gdzie prawdopodobieństwo następnego symbolu zależy od 2. poprzednich). ***(3pt)***

Na koniec wygeneruj przybliżenie źródła Markova drugiego rzedu, zaczynając od słowa "probability". ***(4pt)***

## Przypomnienie

Źródło Markova generuje kolejny symbol z następującym prawdopodobieństwem:

$$P(j|i) = P(i,j)/P(i),$$

gdzie $P(i)$ jest prawdopodobieństwem $n$-gramu $i$, gdzie $n$ = stopień źródła, $P(i,j)$ jest prawdopodobieństwem wystąpienia $n$-gramu $i$ i po nim symbolu $j$, a $P(j|i)$ jest prawdopodobieństwem warunkowym wystąpienia symbolu $j$ zaraz po $n$-gramie $i$.