TIMKoD – Lab 2 – Przybliżanie języka naturalnego – kontynuacja

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.
- 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 znakach, przejdziemy do źródeł używających całych słów jako symboli.

Przygotowanie do zajęć

- Do wykonania zadań potrzebne będą korpusy tekstowe, które można pobrać z katalogu "dane do zadań"w odpowiedniej sekcji kursu.
- Pliki są znormalizowane, zawierają jedynie 26 małych liter alfabetu łacińskiego, cyfry i spacje (czyli w sumie 37 znaków).
- 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 tys. słów, a posługuje się tylko 20% z tego zbioru, co daje tylko 6 tys. słów. Czy w Polsce panuje ubóstwo językowe? Sprawdź jaki procent "wiedzy" z Wikipedii umiałby przekazać przeciętny Polak, gdyby jego elokwencje przełożyć na grunt języka angielskiego.

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

Częstość słów – prawo Zipfa

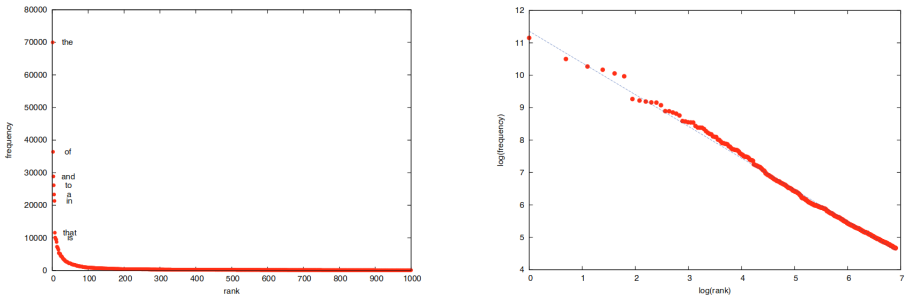
Rysunek 1: George Kingsley Zipf (1902 - 1950)

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 słowa w korpusie, $\propto$ oznacza proporcjonalność, r to ranga słowa odpowiadająca częstości jego występowania względem innych słów w korpusie, natomiast α to stała skalująca równa ok. 1.

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



Rysunek 2: Liczność słów w próbce miliona słów języka angielskiego na skali liniowej (lewa strona) oraz skali logarytmicznej (prawa strona)

Rozkład i zasada Pareto

Powyżej opisana zależność nie dotyczy jedynie liczności słów w języku. Prawo Zipfa jest dyskretną formą rozkładu Pareto, nazwanego tak na cześć włoskiego ekonomisty Vilfreda Pareto, który w 1906 poczynił sławną obserwację, że 80% procent ziemi we Włoszech jest w rękach 20% populacji kraju. Dlatego rozkład Pareto jest również znany jako zasada Pareto albo zasada 80/20, która mówi, że należy oczekiwać, że 20% badanych obiektów związanych jest z 80% pewnych zasobów. Rozkład ten występuje w wielu dziedzinach takich jak fizyka, biologia, czy nauki społeczne. Poniżej przedstawione są przykłady takich rozkładów:

- 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,
- ilość cytowań artykułów naukowych (również prawo Bradforda i prawo Lotka),
- pojemności złóż surowców,
- wielkość kraterów na księżycu, jak i meteorytów,
- intensywność rozbłysków słonecznych.

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 przybliżenia zerowego 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 preferencją 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.

2 Przybliżenie pierwszego rzędu △

Treść

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

3 Przybliżenia na podstawie źródła Markowa *10pt*◇

Treść

Wygeneruj przybliżenie języka angielskiego na podstawie źródła Markowa pierwszego rzędu na słowach (tzn. prawdopodobieństwo następnego słowa zależy od jednego poprzedniego słowa). ***(3pt)***

Implementacja powinna opierać się na łańcuchu Markowa, nie zaś na metodzie zaproponowanej przez Shannona, polegającej na wyszukiwaniu słów w tekście. Słowa wygenerowane w ten drugi sposób mogą pochodzić z rozkładu odbiegającego od rzeczywistego prawdopodobieństwa warunkowego. Zastanów się, dlaczego tak jest.

Następnie zrób to samo dla źródła Markowa drugiego rzędu (tzn. prawdopodobieństwo następnego słowa zależy od dwóch poprzednich słów). ***(3pt)***

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

Przypomnienie

Źródło Markowa generuje kolejne słowo j po n-gramie i (gdzie n jest stopniem źródła) z następującym prawdopodobieństwem warunkowym:

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

gdzie $P(i)$ jest prawdopodobieństwem n-gramu i, a $P(i,j)$ prawdopodobieństwem łącznym wystąpienia n-gramu i oraz słowa j.

Źródła

- 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
- http://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf
- http://www.inference.org.uk/itprnn/book.pdf