Popularność słów w języku angielskim
405219, Piotr Ludynia
AGH, Wydział Informatyki Elektroniki i Telekomunikacji
Rachunek prawdopodobieństwa i statystyka 2020/2021
Kraków, piątek 28 stycznia 2022 r. – wieczór

Ja, Piotr Ludynia deklaruje, ze przygotowałem przedstawiony do oceny projekt samodzielnie i żadna jego część nie jest kopia pracy innej osoby.

## Streszczenie raportu

Raport powstał w oparciu o analizę danych dotyczących popularności słów w języku angielskim.
Przedstawia analizę zależności długości słów i ich częstości użycia

## Opis danych

Dane do projektu pochodzą ze strony https://www.kaggle.com/rtatman/english-word-frequency
Zawierają one 333333 najpopularniejszych słów angielskich w internecie.
Zbiór danych został wydobyty z plików firmy google dostępnych na stronie
https://norvig.com/ngrams
Zawierają 2 kolumny. Jedną będącą słowem, a drugą będącą liczbą jego wystąpień.

# Analiza danych

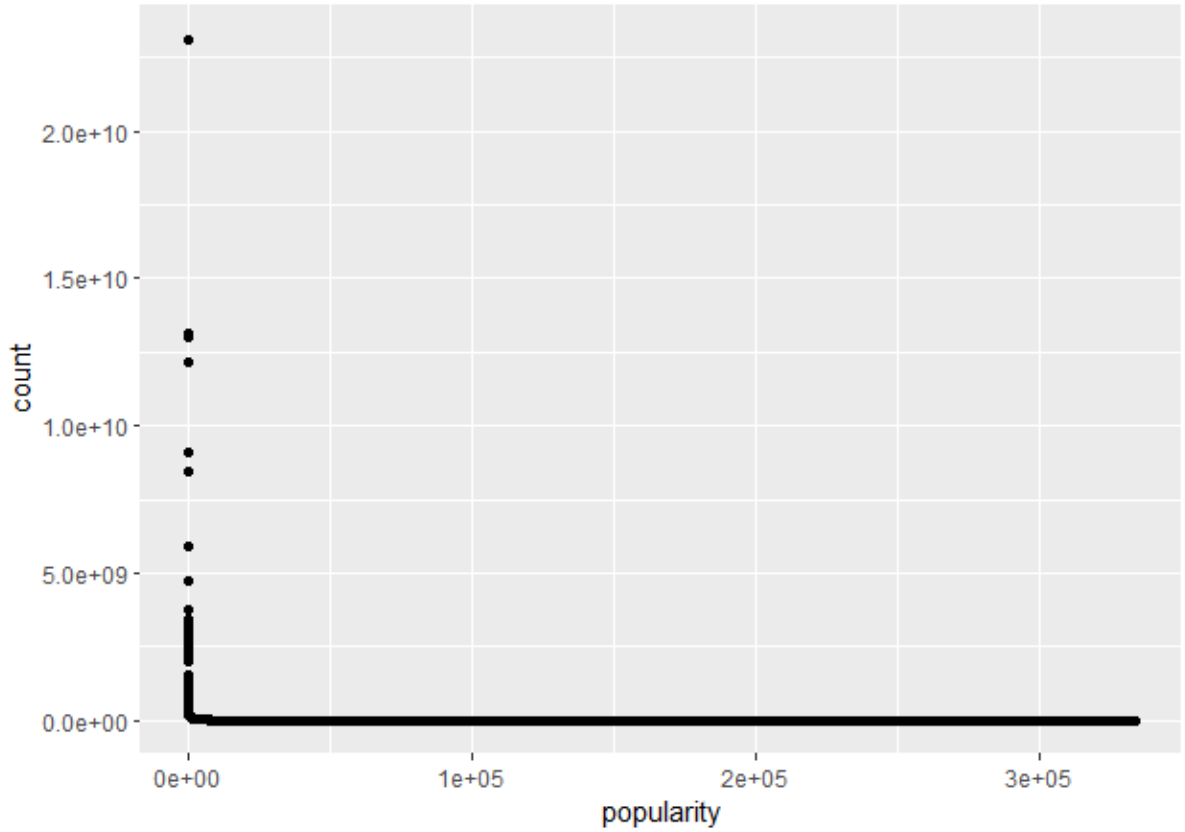
Poniższe kroki przedstawiają analizę danych

## Obróbka danych

- Program wczytuje pakiet ggplot2.
- Następnie następuje wczytanie danych z pliku unigram_freq.csv i wypisanie kilku informacji o zbiorze danych: wymiarów danych, kolumn, typów i kilku pierwszych wyrazów w kolumnach.
- Dane zostają posortowane malejąco według popularności(liczbie wystąpień).
- Dodane zostają następnie 2 kolumny.
- Popularity - oznaczającą kolejność od najpopularniejszego do najmniej popularnego słowa .
- Lenght - długość słowa wyrażoną w liczbie liter.

## pierwszy wykres

Zostaje dodany pierwszy wykres ggplot będący zależnością między pozycją w rankingu popularności a częstością wystąpień




Jest on nieczytelny ze względu na prawo zipfa dotyczące danych tworzonych przez ludzi. Nie jest ono tematem tego projektu dla tego wystarczy tylko w skrócie wyjaśnić, że ranga(nasz numer w rankingu popularności) jest odwrotnie proporcjonalny do częstości wystąpień słowa próbie.

## Co teraz?

- Wykres nie dał ciekawych wniosków jednak nie należy się poddawać. Wykorzystane zostają pozostałe dane, czyli długość słowa.
- Znalezione zostaje słowo o największej liczbie znaków – tutaj jest to 38.
- Powstają 3 wektory o długości 38
    - word_length – liczba liter w słowie
    - word_count  - liczba słów danej długości
    - word_popularity – liczba wystąpień słów o danej długości
- Pętla for przechodzi przez wszystkie słowa zbierając potrzebne informacje – uzupełnia dane w wektorach

## Przygotowanie do kolejnych kroków

- Zadeklarowana zostaje funkcja var_with_mean przyjmująca wartość średnią i wektor i zwracająca wartość wariancji elementów wektora.
    - Nie można użyć funkcji var(), ponieważ przyjmuje ona wektory obciążone wagami, z których sama wylicza wartość oczekiwaną.
    - Wektor 38 długości słowa od 1 do 38 to po prostu ciąg arytmetyczny. Taka funkcja nie wystarczy. Funkcja var_with_mean 'ręcznie' oblicza estymator wariancji.
    - W celu późniejszego rysowania wykresów zostaje stworzony nowy dataframe zawierający wspomniane wektory.
- Następuje obliczenie kwantyli rozkładu chi kwadrat oraz rozkładu testu Denta dla przedziału ufności 0.95

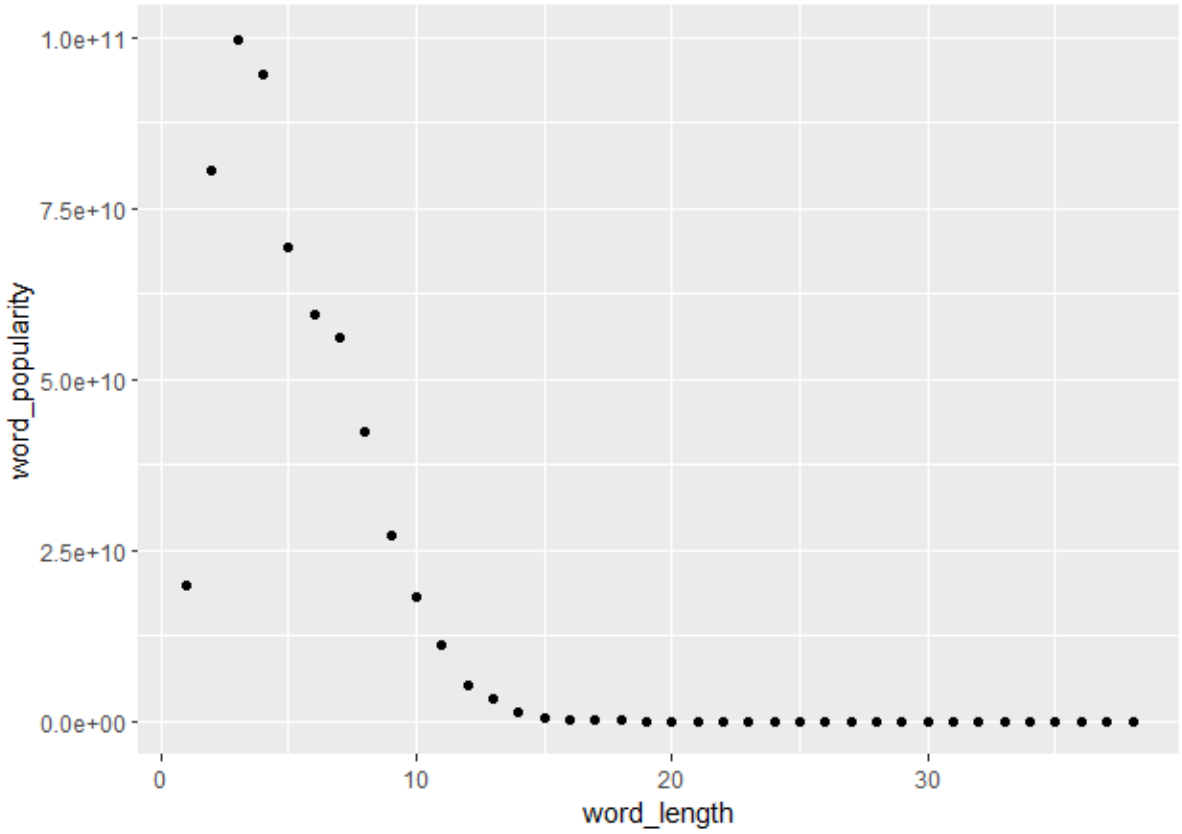
## Zależności

Następnie przeprowadzone zostają dwa identyczne rozumowania dla dwóch zależności między danymi. Zależność między długością słowa a ich liczbą wśród 333333 rozpatrywanych oraz zależność między długością słowa a liczbą wystąpień słów o takiej długości. Oto następujące kroki

- Wyświetlenie wykresu zależności.
- Obliczenie średniej ważonej z długości słów z wagami będącymi odpowiednio liczbą słów i liczbą wystąpień słowa o danej długości.
- Obliczenie wariancji przy pomocy funkcji var_with_mean oraz obliczenie estymatora odchylenia standardowego
- Obliczenie przedziałów ufności dla wartości oczekiwanej i dla odchylenia standardowego przy pomocy odpowiednio rozkładu T Studenta i rozkładu chi kwadrat.

## Zależność między długością słowa a częstością wystąpień słów o tej długości

- Wartość średnia 5.052814
- Estymowana wariancja 337.8623
- Estymator odchylenia standardowego 18.38103
- Przedział ufności odchylenia standardowego (14.98539, 23.78041)
- Długość przedziału ufności odchylenia standardowego 8.795019
- Przedział ufności wartości oczekiwanej (-0.9888803, 11.09451)
- Długość przedziału ufności wartości oczekiwanej 12.08339

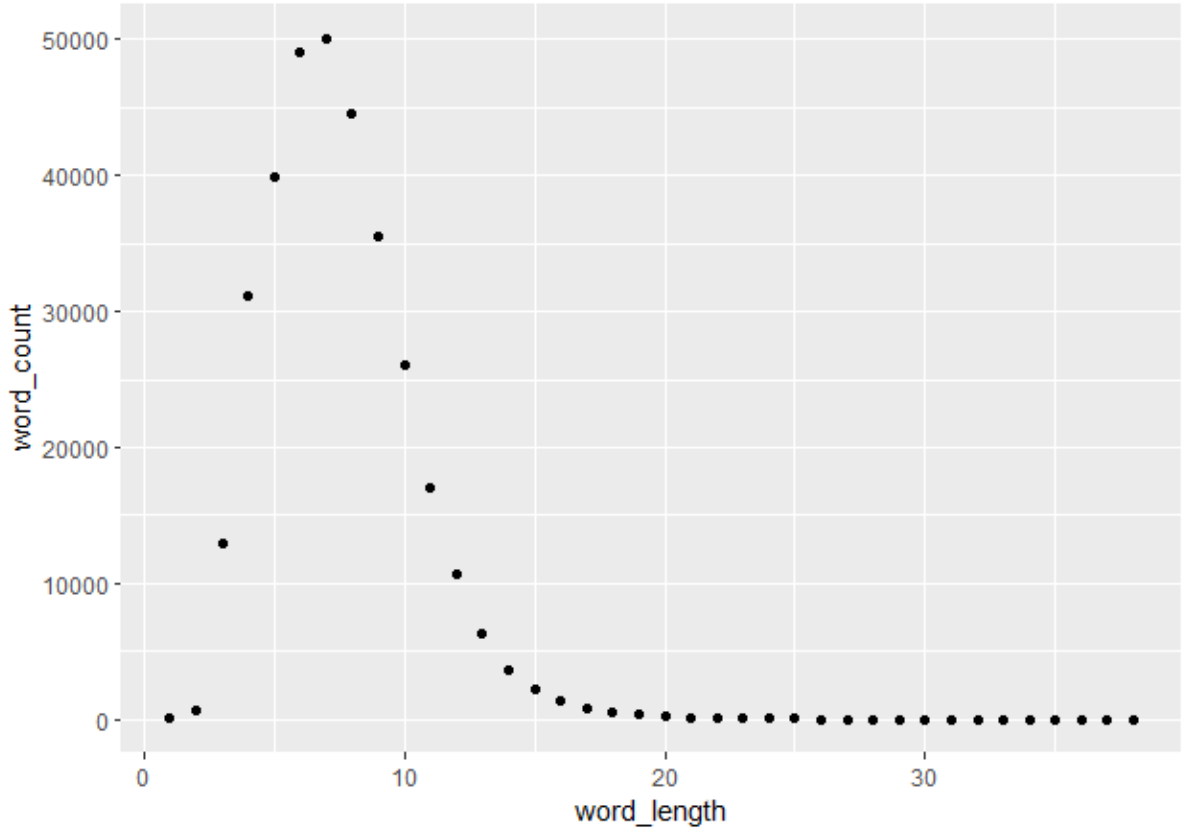



Wnioski

Otrzymane estymatory znajdują się w przedziałach ufności tych wartości, które estymują. To bardzo dobry znak, jednak otrzymana wartość średnia różni się bardzo od wartości najczęściej występującej w próbie, czyli 3 która mieści się w przedziale ufności, jednak bliżej jego brzegu niż obliczona wartość średnia.

## Zależność między długością słowa a liczbą słów o tej długości wśród wybranych 333333 w próbie

- Wartość średnia 7.470334
- Estymowana wariancja 272.124
- Estymator odchylenia standardowego 16.49618
- Przedział ufności odchylenia standardowego (13.44874, 21.34189)
- Długość przedziału ufności odchylenia standardowego 7.89315
- Przedział ufności wartości oczekiwanej (2.048174, 12.89249)
- Długość przedziału ufności wartości oczekiwanej 10.84432




Wnioski

Otrzymane estymatory znajdują się w przedziałach ufności tych wartości, które estymują. Wartość średnia jest bardzo bliska najczęściej występującej długości słowa, czyli 7. W dodatku znajduje się ona blisko środka przedziału ufności wartości oczekiwanej, więc jest jej dobrym estymatorem.

## Porównanie  i wnioski końcowe

Druga próba przyniosła dużo lepsze rezultaty. Nie tylko różnica między najczęstszą wartością a estymatorem wartości oczekiwanej była mniejsza, ale również przedziały ufności były mniejsze, więc dokładniejsze. Estymator wariancji także był mniejszy w drugiej próbie a krańce jego przedziału ufności były mniejsze niż te odpowiadające im w pierwszej próbie. Można wyciągnąć wniosek, że branie pod uwagę liczby słów danej długości w grupie pewnych najczęściej wykorzystywanych może lepiej służyć przyszłemu modelowaniu statystyk związanych z językiem angielskim. Bardziej przypomina on rozkład normalny.

Taki kształt wykresów może wynikać z tego, iż liczba słów o mniejszej długości jest ograniczona. W języku angielskim jest tylko 26 liter, które mogą tworzyć słowa długości 1. Za to dłuższe słowa nie są często używane, a ponadto w języku naturalnym jest ich faktycznie mniej niż krótkich. Od pewnej długości słowa te przestają mieć sens i stają się pozbawionymi od informacji dziwnoneologizmami oraz pseudoonomatopejami.

Najciekawszy wniosek to chyba jednak ten, że w pierwszej próbie dominujące słowa(te o większym prawdopodobieństwie) były krótsze niż w drugiej próbie. Próba druga nie bierze pod uwagę jak często wykorzystywane jest każde z 333333 słów, a jedynie to, ŻE są wykorzystywane częściej niż wszystkie pozostałe, które nie były przetwarzane. Oznacza to, że użytkownicy internetu – miejsca, z którego pochodzą dane optują w stronę użycia krótszych słów niż bardziej skomplikowanych i wyrafinowanych niż by na to wskazywał bardziej naturalny rozkład.
Aż nasuwa się na myśl hipoteza, że źródła, w których pierwszy rozkład jest podobny do drugiego – bardziej naturalnego –są bardziej cywilizowane i/lub inteligentniejsze niż te, w których odchyla się od niego w stronę krótszych słów. Nie pozbawioną sensu jest myśl, że osoby bardziej wykształcone i poważne będą posługiwały się bogatszym słownictwem niż te o mniejszym wykształceniu. Jest to jednak temat niestety na inny projekt i inne badanie, które z chęcią kiedyś przeprowadzę