# Zadanie 1 - Analiza została przeprowadzona dla języka włoskiego.

**Część 1**

Na potrzeby analizy wykorzystano gotowy korpus języka włoskiego, liczący 1 600 000 słów (dostępny pod adresem: https://clarin.eurac.edu/repository/xmlui/handle/20.500.12124/3).
Zastosowany plik zawierał dane przefiltrowane, pozbawione symboli oraz cyfr.

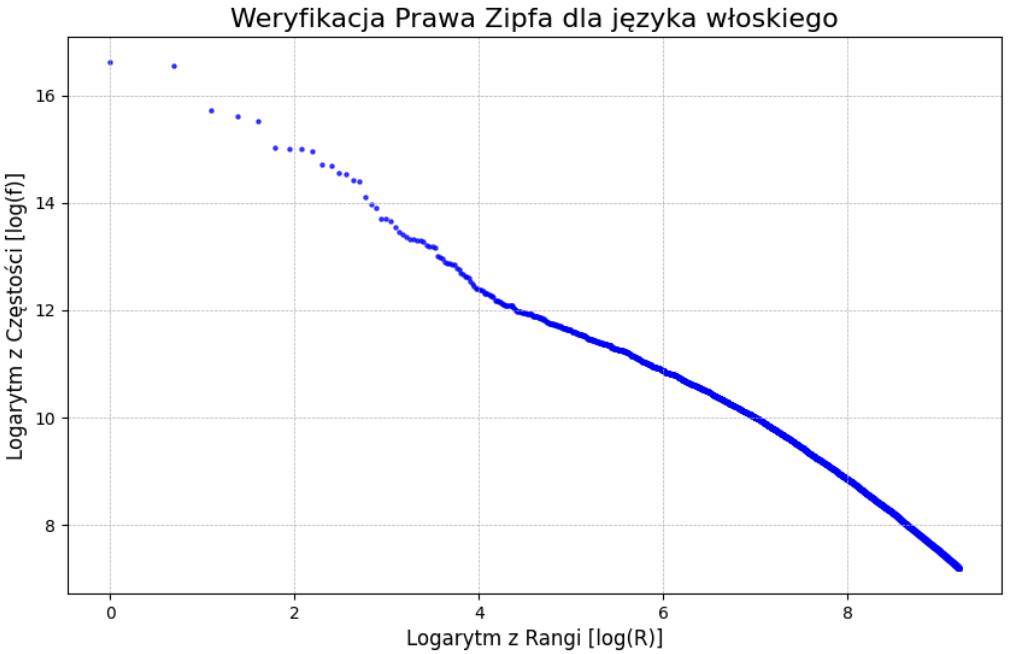
| | slowo | ranga | czestotliwosc |
|---|---|---|---|
| 1 | di | 1 | 16520744 |
| 2 | il | 2 | 15591226 |
| 3 | in | 3 | 6667272 |
| 4 | essere | 4 | 5972216 |
| 5 | e | 5 | 5550860 |
| 6 | che | 6 | 3354704 |
| 7 | al | 7 | 3296016 |
| 8 | da | 8 | 3254354 |
| 9 | a | 9 | 3167841 |
| 10 | un | 10 | 2442451 |
| 11 | per | 11 | 2426164 |
| 12 | si | 12 | 2089995 |
| 13 | una | 13 | 2074314 |
| 14 | con | 14 | 1829124 |
| 15 | avere | 15 | 1790433 |
| 16 | non | 16 | 1352772 |
| 17 | suo | 17 | 1169009 |
| 18 | su | 18 | 1082946 |
| 19 | questo | 19 | 894228 |
| 20 | come | 20 | 891248 |
| 21 | più | 21 | 847093 |
| 22 | anche | 22 | 757341 |
| 23 | ma | 23 | 697158 |
| 24 | venire | 24 | 667545 |
| 25 | fare | 25 | 635602 |
| 26 | ad | 26 | 616510 |
| 27 | quello | 27 | 611760 |
| 28 | potere | 28 | 600722 |

W ramach opracowania przygotowano skrypt, który przetwarzał dane korpusowe, generował wykres w skali logarytmiczno-logarytmicznej oraz obliczał współczynniki nachylenia i determinacji.

Uzyskane wyniki jednoznacznie potwierdziły zgodność języka włoskiego z Prawem Zipfa. Na wykresie widoczna jest niemal idealna zależność liniowa, co wskazuje na typowy rozkład frekwencyjny charakterystyczny dla języków naturalnych.
Dodatkowo wynik obliczeń numerycznych wzmacnia tę obserwację - współczynnik nachylenia (s) wynosi -1.120, czyli wartość bardzo zbliżoną do

teoretycznej -1, natomiast współczynnik determinacji $R^2 = 0.9852$ wskazuje na doskonałe dopasowanie modelu do danych empirycznych.

```
Współczynnik nachylenia (s): -1.120
Współczynnik determinacji R²: 0.9852
```

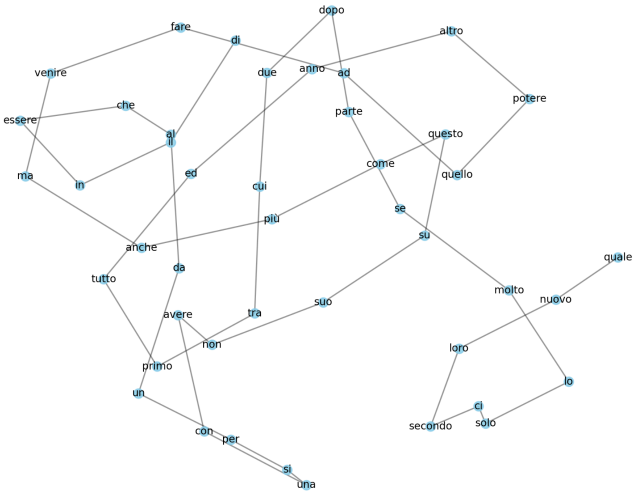



**Część 2**

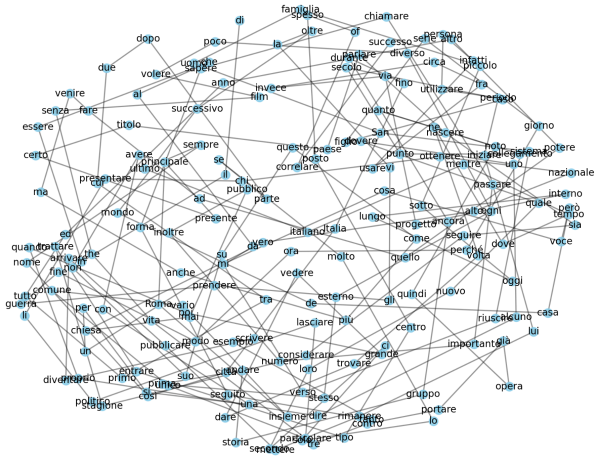
W drugim etapie analizy zbudowano graf współwystępowania słów w języku włoskim.
Wykorzystano bibliotekę NetworkX, w której węzły odpowiadają poszczególnym słowom, a krawędzie ich współwystępowaniu w korpusie.

Na podstawie utworzonego grafu obliczono miary centralności, pozwalające wskazać słowa najczęściej występujące w różnych kontekstach, stanowiące rdzeń języka włoskiego.

Graf współwystępowania słów w języku włoskim dla 50 słów




Graf współwystępowania słów w języku włoskim dla 200 słów



**Część 3**

Każde słowo w korpusie posiada przypisaną częstość występowania, określającą liczbę jego pojawień się w zbiorze tekstów.

Po zsumowaniu wszystkich częstości uzyskano łączną liczbę wystąpień słów w całym korpusie. Następnie obliczano sumę skumulowaną częstości, rozpoczynając od słów najczęściej występujących. W momencie, gdy suma przekroczyła 90% całkowitej liczby wystąpień, uznano, że zbiór obejmuje słownictwo pozwalające na zrozumienie 90% tekstów w języku włoskim.

W tym celu opracowano skrypt, który automatycznie sumował częstości występowania kolejnych słów aż do osiągnięcia progu 90%.

Na podstawie uzyskanych wyników ustalono, że do zrozumienia 90% tekstu w języku włoskim wystarczająca jest znajomość około 11 629 słów.

```
Liczba słów potrzebnych do pokrycia 90% języka: 11629
```

**Zadanie z gwiazdką**
Tłumaczenie 50 rzeczowników ze 180 pierwszych słów korpusu języka:

| Nr | Słowo (IT) | Znaczenie (PL) |
|---|---|---|
| 19 | questo | ten (rzeczownikowo: „to”) |
| 29 | altro | inny (rzeczownikowo: „inny [człowiek/rzecz]”) |
| 30 | anno | rok |
| 33 | tutto | wszystko |
| 38 | dopo | po (rzeczownikowo: „później”, rzadko) |
| 40 | parte | część, strona |
| 48 | nuovo | nowy (rzeczownikowo: „nowy [człowiek/rzecz]”) |
| 61 | volta | raz, okazja |
| 64 | tempo | czas |
| 66 | italiano | Włoch / włoski (rzeczownik: Włoch) |
| 67 | ultimo | ostatni (rzeczownikowo: „ostatni [człowiek]”) |
| 69 | nome | imię, nazwa |
| 71 | città | miasto |
| 72 | Italia | Włochy |
| 76 | gruppo | grupa |
| 77 | storia | historia |
| 80 | stesso | ten sam (rzeczownikowo: „ten sam [człowiek/rzecz]”) |
| | | |
| 84 | poco | mało (rzeczownikowo: „niewiele”) |
| 86 | film | film |
| 89 | vita | życie |
| 94 | opera | dzieło, opera |
| 99 | punto | punkt |
| 100 | secolo | wiek (np. XX wiek) |
| 101 | caso | przypadek |
| 102 | figlio | syn |
| 103 | serie | seria |
| 104 | giorno | dzień |

| 107 | modo | sposób
| 108 | fine | koniec
| 109 | guerra | wojna
| 114 | mondo | świat
| 118 | casa | dom
| 120 | voce | głos
| 121 | cosa | rzecz
| 123 | verso | kierunek, zwrot
| 125 | sistema | system
| 126 | San | święty (rzeczownikowo: Święty [np. San Marco])
| 127 | uomo | mężczyzna, człowiek
| 128 | famiglia | rodzina
| 129 | piccolo | mały (rzeczownikowo: „maluch”)
| 133 | interno | wnętrze
| 134 | progetto | projekt
| 135 | via | ulica, droga
| 141 | persona | osoba
| 143 | periodo | okres, czas
| 155 | forma | forma, kształt
| 159 | numero | liczba, numer
| 162 | pubblico | publiczność
| 164 | chiesa | kościół
| 166 | alto | wysoki (rzeczownikowo: „wysokość”)
| 167 | nazionale | narodowy (rzeczownikowo: „reprezentacja narodowa”) |
| 172 | politico | polityk
| 174 | particolare | szczegół
| 175 | unico | jedyny (rzeczownikowo: „jedyny”)
| 176 | comune | gmina, wspólnota
| 177 | vario | różny (rzeczownikowo: „różność”)