Prawo Zipfa
Modelowanie komputerowe
Lista 1b

Łukasz Chmielowski (307713)

2023

1 Treść zadania

Na liście znajdują się dwa zadania:

W zadaniu pierwszym należało napisać własny kod do analizy Zipfa dowolnego tekstu podanego w formie pliku tekstowego.

W zadaniu drugim mieliśmy przeanalizować 3 teksty za pomocą programu z zadania 1 oraz pokazać je na wykresach.

2 Zadanie pierwsze

Kod do analizy Zipfa napisałem w języku Python, ponieważ wydał mi się najwygodniejszy do analizy tekstu. Cały kod znajduje się w pliku dołączonym do sprawozdania (zipf.py).

3 Zadanie drugie

Do analizy Zipfa wykorzystałem 3 teksty:

- Skrypt do filmu "Shrek"
- Książka "Winnie the Pooh"A. A. Milne'a
- Książka "Dracula"Brama Stokera

Po użyciu mojego kodu na tych tekstach, otrzymałem pliki z liczbą wsytąpień każdego ze słów posortowane malejąco wg wystąpień. Dla każdego tekstu utworzyłem wykres za pomocą programu GnuPlot (wykresy narysowane na końcu sprawozdania).

Żeby wykresy można było odczytać w jakiś sensowny sposób, przedstawiłem je w skali log-log. Następnie dopasowałem do wykresów funkcję $f(x)$:

$$f(x) = ar^b \tag{1}$$

Gdzie:
r jest rangą słowa,
a jest współczynnikiem,
b wyjaśnię poniżej.

Cała analiza Zipfa polega właśnie na zrozumieniu czym jest b. W przypadku analizy tekstu, trzeba rozpatrzyć b z lingwistycznego punktu widzenia. Można tą liczbę porównać do *poziomu zaawansowania* tekstu. im wartość tej liczby jest większa, tym więcej jest słów unikatowych, co może świadczyć o większym bogactwie tekstu.

Z tą widzą porównałem wartości b dla każdego z tekstów:

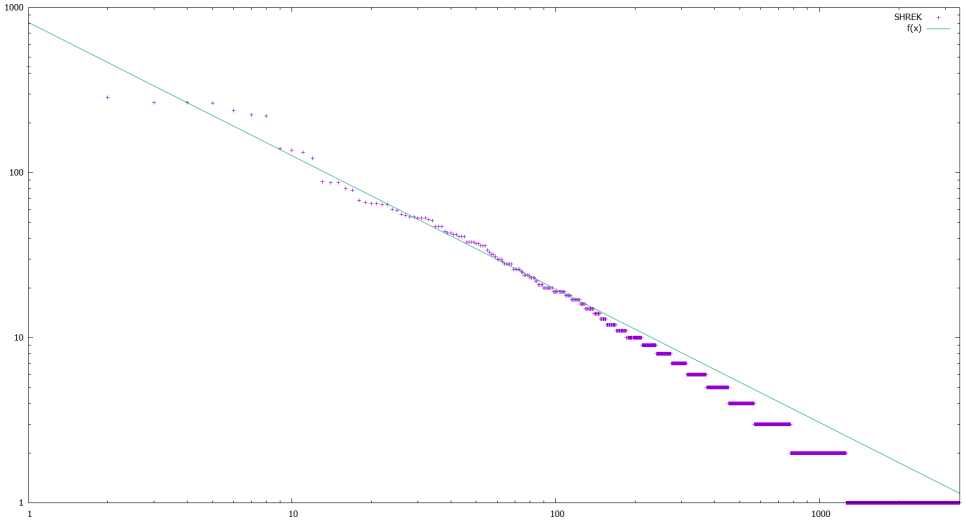
Tekst	b
Shrek	-0.757933
Winnie the Pooh	-0.918092
Dracula	-1.18244

Możemy zauważyć, że wartość b dla *Draculi* jest największy, a dla *Shreka* najmniejszy. Od razu nasuwają się wnioski i wyjaśnienia.

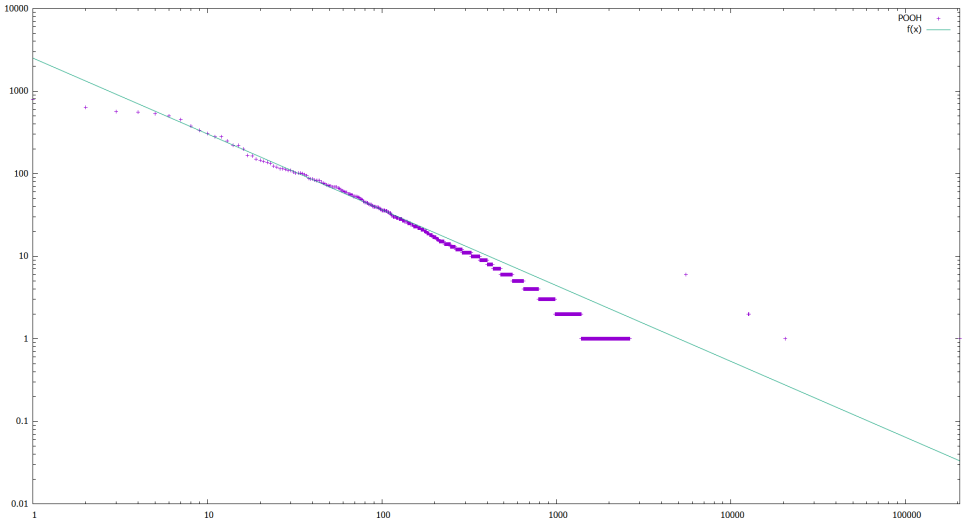
Shrek jest filmem, przez co automatycznie posiada uboższe słownictwo, ponieważ nie zawiera szczegółowych opisów otoczenia i wydarzeń, tylko same wypowiedzi postaci. Dodatkowo jest to film skierowany do dzieci, wobec czego nie będzie miał wielu unikatowych słów, które często są również słowami trudnym lub rzadko używanymi.

Winnie the Pooh jest książką skierowaną do dzieci. Jednak w przeciwieństwie do *Shreka*, tutaj już są wymagane opisy miejsc akcji, chociaż ogólne, żeby czytelnik mógł się lepiej wczuć w opowieść jak i zrozumieć kontekst wypowiedzi postaci.

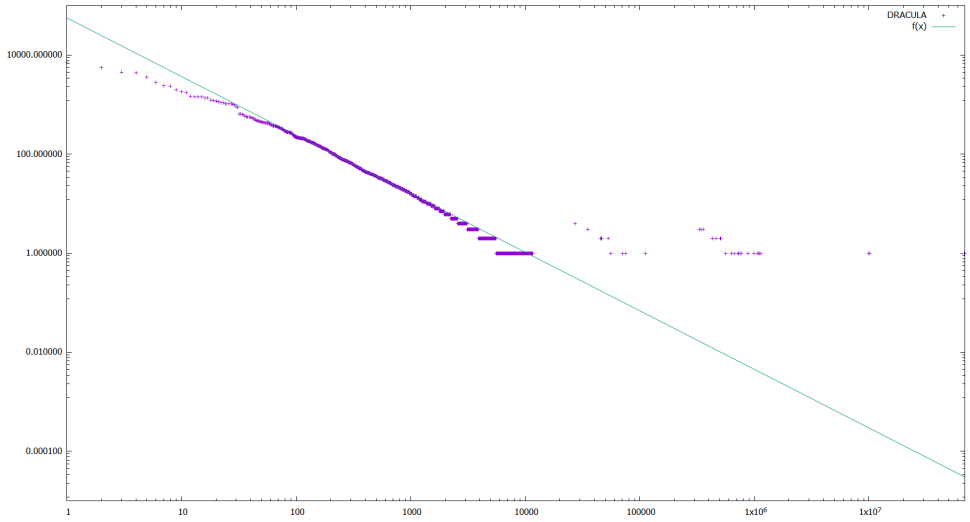
Dracula z kolei jest już literaturą skierowaną do osób dorosłych. Ten tekst będzie już bardziej wymagający przez bardziej szczegółowe opisy miejsc i wydarzeń. Znaczenie również ma sama długość powieści, która jest o wiele większa niż pozostałe 2 teksty. Im dłuższy tekst, tym większa szansa na wystąpienie unikatowego słowa.



Rysunek 1: Prawo potęgowe dla *Shreka*



Rysunek 2: Prawo potęgowe dla *Winni the Pooh*



Rysunek 3: Prawo potęgowe dla *Draculi*