Przetwarzanie języka naturalnego w systemach sztucznej inteligencji

Projekt:

Prawo Zipfa

AGH Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej

Autorzy
Michał Burda - michaburda@student.agh.edu.pl
Radosław Barszczak - rbarszczak@student.agh.edu.pl

1. Co to jest prawo Zipfa?

Prawo Zipfa to matematyczne prawo opisujące zjawisko występujące w różnych dziedzinach, w szczególności w językoznawstwie, socjologii, ekonomii i informatyce. Zostało ono sformułowane przez amerykańskiego lingwistę George'a Zipfa w 1935 roku na podstawie analizy częstotliwości występowania słów w tekstach językowych.

Zasada Prawa Zipfa

Prawo Zipfa stwierdza, że częstotliwość występowania słowa jest odwrotnie proporcjonalna do jego miejsca (rangi) na liście najczęściej używanych słów. Innymi słowy, drugie najczęściej używane słowo pojawia się dwa razy rzadziej niż pierwsze, trzecie słowo trzy razy rzadziej niż pierwsze, i tak dalej.

Wzór prawa Zipfa:

$$r \times f = \mathrm{constans},$$

gdzie r jest to ranga wyrazu w tekście lub grupie tekstów, a f częstotliwość jego występowania[2].

2. Liczba wszystkich słów z artykułów

Łączna liczba słów: 24436

3. Liczba unikalnych słów w tekście:

Łączna liczba unikalnych słów: 8501

4. Tabela Zipfa dla pierwszych 30 wyrazów

Nr	Słowo	Ranga	Częstotliwość	Ranga * Częst.
1	je	1	1180	1180
2	i	1	918	918
3	u	2	577	1154
4	se	2	550	1100
5	su	2	488	976
6	na	3	463	1389
7	da	3	404	1212
8	a	5	250	1250
9	za	6	208	1248
10	od	7	170	1190
11	s	8	147	1176
12	nije	9	128	1152
13	koji	9	125	1125
14	te	10	115	1150
15	to	11	112	1232
16	će	11	106	1166
17	bi	12	100	1200
18	mu	12	99	1188
19	kako	12	98	1176
20	što	12	96	1152
21	ga	13	93	1209
22	do	13	91	1183
23	ne	13	90	1170
24	bambi	14	86	1204
25	kao	14	84	1176
26	kada	15	81	1215
27	sve	16	75	1200
28	ali	16	73	1168
29	bio	16	72	1152
30	iz	17	70	1190

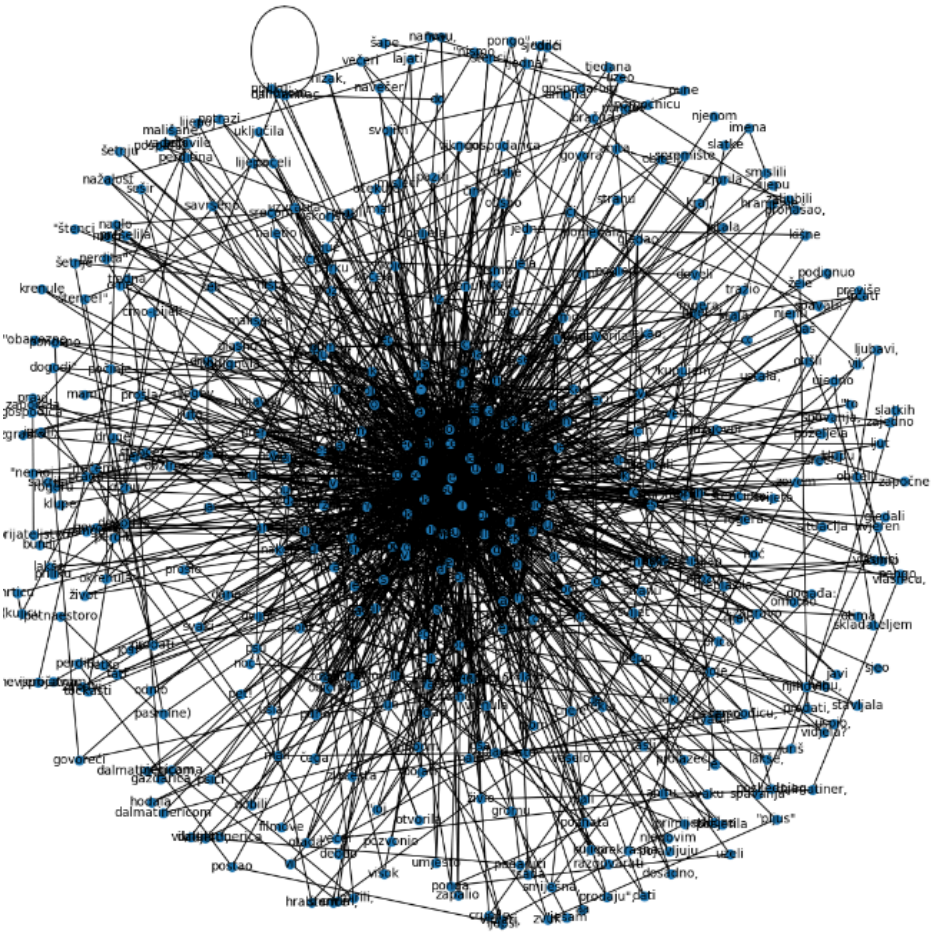
5. Ile słów należy znać, aby rozumieć tekst?

a) 10% - Aby rozumieć 10% tekstu, trzeba znać 3 unikalne słowa, co pozwala zrozumieć 2675 z 24436 słów.
b) 20% - Aby rozumieć 20% tekstu, trzeba znać 9 unikalnych słów, co pozwala zrozumieć 5038 z 24436 słów.

c) 30% - Aby rozumieć 30% tekstu, trzeba znać 33 unikalne słowa, co pozwala zrozumieć 7333 z 24436 słów.
d) 40% - Aby rozumieć 40% tekstu, trzeba znać 114 unikalnych słów, co pozwala zrozumieć 9789 z 24436 słów.
e) 50% - Aby rozumieć 50% tekstu, trzeba znać 350 unikalnych słów, co pozwala zrozumieć 12220 z 24436 słów.
f) 60% - Aby rozumieć 60% tekstu, trzeba znać 891 unikalnych słów, co pozwala zrozumieć 14664 z 24436 słów.
g) 70% - Aby rozumieć 70% tekstu, trzeba znać 1923 unikalnych słów, co pozwala zrozumieć 17107 z 24436 słów.
h) 80% - Aby rozumieć 80% tekstu, trzeba znać 3614 unikalnych słów, co pozwala zrozumieć 19549 z 24436 słów.
i) 90% - Aby rozumieć 90% tekstu, trzeba znać 6058 unikalnych słów, co pozwala zrozumieć 21993 z 24436 słów.
j) 100% - Aby rozumieć 100% tekstu, trzeba znać 8501 unikalnych słów, co pozwala zrozumieć 24436 z 24436 słów.

6. Graf połączeń wyrazowych

Graf dla pierwszych 500 słów z plików tekstowych:



Źródła:

Artykuły:

1. Kulinarny https://www.24sata.hr/lifestyle/penicilin-juha-povrce-i-meso-nemojte-vadite-nego-ga-usitnite-1007499?24sata_ref=frontpage-home
2. Sport https://gol.dnevnik.hr/clanak/rubrika/nogomet/modriceva-tajna-dugovjecnosti-jede-rib u-jaja-i-piletinu-a-tri-puta-pocasti-se-omiljenom-delicijom---871740.html?itm_source=TopNews&itm_medium=Dnevnik&itm_campaign=Naslovnica&_gl=1*1j3w9z8*_gcl_au*MTc2MzA3ODY5MC4xNzI3ODEwODcx
3. Wiadomości 1 https://dnevnik.hr/vijesti/hrvatska/hac-je-odlucio-i-u-ostatku-godine-zadrzati-cijene-cestarina-koje-su-trebale-biti-samo-tijekom-ljetne-sezone---871762.html?itm_source=TopNews2&itm_medium=Dnevnik&itm_campaign=Naslovnica
4. Wikipedia Pies https://hr.wikipedia.org/wiki/Doma%C4%87i_pas
5. Wikipedia Komputer https://hr.wikipedia.org/wiki/Ra%C4%8Dunalo
6. Wikipedia Samolot https://hr.wikipedia.org/wiki/Avion
7. Bajka1 https://www.bajke.hr/101-dalmatinac/
8. Bajka2 https://www.bajke.hr/crvenkapica-grimm/#google_vignette
9. Wiadomości 2 https://www.index.hr/vijesti
10. Wiadomości 3 https://n1info.hr/
11. Wiadomości 4 https://www.vecernji.hr/najcitanije-vijesti/2024-5-25?s=1&s=199
12. Bajka3 https://www.bajke.hr/bambi/
13. Wikipedia Tenis https://hr.wikipedia.org/wiki/Tenis
14. Wikipedia Gra https://hr.wikipedia.org/wiki/Igra
15. W pustyni i w puszczy https://www.lektire.hr/kroz-pustinju-i-prasumu/