Przetwarzanie języka naturalnego w systemach sztucznej inteligencji - projekt 1

Adam Stajek, Maciej Trzaskacz

1 Wstęp

Nasza praca dotyczy analizy języka portugalskiego opartej na zastosowaniu prawa Zipfa dla korpusu językowego.

2 Sposób wykonania badań

Kod napisany w języku Python zastosowany do wykonania zadania można znaleźć na platfomie Github.

2.1 Korpus tekstowy

Do wykonania zadania użyta została portugalska wersja zbioru dqnych SQUAD v1.1 (The Stanford Question Answering Dataset). Jest to zbiór danych zawierający pytania i odpowiedzi na nie. Nasz korpus stworzyliśmy poprzez stworzenie pliku tekstowego z połączenia wszystkich odpowiedzi w datasecie. Jest to bardzo różnorodny zbiór danych, gdyż pytania są zadawane na temat bardzo szerokiej gamy artykułów z Wikipedii. Ostatecznie nasz zbiór danych zawierał prawie 2 miliony 300 tysięcy słów. Oryginalna wersja korpusu znajduje się na platformie Github.

2.2 Preprocessing tekstu

Po połączeniu wszystkich odpowiedzi w jeden plik, przyszedł czas na przygotowanie korpusu do analizy. Polegało ono na usunięciu z tekstu wszystkich znaków, które nie były literami oraz zamienieniu wszystkich wielkich liter małymi. Następnie stokenizowano tekst jako metodę podziału ustalając whitespace.

3 Wyniki badań

3.1 Prawo Zipfa

Prawo Zipfa mówi, że "gdy na podstawie ich korpusów językowych ustali się wykaz wyrazów ułożonych w malejącym porządku częstotliwości ich występowania, to ranga (numer porządkowy) wyrazu jest odwrotnie proporcjonalna do częstotliwości, zatem iloczyn częstotliwości i rangi powinien być wielkością stałą". Rzeczywiście, można dostrzeć tę zależność tabeli 1,1. Dużym outlierem są pierwsze dwie wartości, ale potem stabilizują się one. Można więc uznać że używany korpus języka portugalskiego powinien być stosunkowo przystępny dla przeciętnego czytelnika.

Words	Counts	Rank	Zipf Score
de	134765	1	134765
a	80240	2	160480
e	71414	3	214242
o	65293	4	261172
em	47792	5	238960
do	38473	6	230838
da	37659	7	263613
que	35217	8	281736
um	25873	9	232857
para	25527	10	255270
os	25291	11	278201
uma	22764	12	273168
no	19248	13	250224
com	18501	14	259014
como	18071	15	271065
na	16731	16	267696
é	16163	17	274771
as	15303	18	275454
por	14204	19	269876
se	13268	20	265360

Table 1: Wartości związane z prawem Zipfa dla najczęściej występujących słów w języku portugalskim

3.2 Liczba słów, a procent tekstu

Na podstawie powyższej tabeli wyliczyliśmy także liczbę słów, jakie należy umieć, by móc przeczytać konkretny procent tekstu. Oczywiście nie znaczy to, że cokolwiek by się zrozumiało, ponieważ te słowa to głównie spójniki, przedimki i inne formy wyrazowe bez istotnego znaczenia semantycznego.

Percent of Text	Number of Words
10%	3
20%	7
30%	17
40%	51
50%	202
80%	3965

Table 2: Liczba słów potrzebna do przeczytania konkretnego procenta tekstu

3.3 Najczęściej spotykane sąsiednie pary słów

Ostatnią częścią badania było znalezienie najczęściej występujących obok siebie par form wyrazowych w języku portugalskim. Oto wyniki, które niestety też nie mówią za dużo o semantyce języka portugalskiego.

Word 1	Word 2	Count
e	a	4588
e	o	4118
para	o	3430
para	a	3255
de	um	3045
com	a	3003
que	a	2911
com	o	2906
em	de	2879
que	o	2794
de	uma	2452
de	a	2140
no	entanto	2139
como	o	1973
que	os	1892
em	um	1889
estados	unidos	1843
em	uma	1783
como	a	1710
do	século	1669

Table 3: Najczęściej sąsiadujące pary słów

4 Podsumowanie

W powyższym dokumencie wykonano analizę języka portugalskiego za pomocą własności wynikających z Prawa Zipfa. Wywnioskowano między innymi, że należy znać ledwie ponad 200 słów, by potrafić przeczytać 50 procent portugalskiego tekstu. Zaobserwowano także wyniki zgodne z tezą zawartą w prawie Zipfa.