Przetwarzanie języka naturalnego w systemach sztucznej inteligencji - projekt 3

Adam Stajek, Maciej Trzaskacz

1 Wstęp

Nasza praca dotyczy analizy połączeń oraz zależnościami między czasownikami i rzeczownikami w języku angielskim.

2 Sposób wykonania badań

Kod napisany w języku Python, z użyciem bibliotek NLTK oraz Numpy, zastosowany do wykonania zadania można znaleźć na platfomie Github.

2.1 Korpus tekstowy

Do wykonania zadania został wykorzystany korpus *Wikitext* firmy *Salesforce* zawierający najpopularniejsze artykuły na angielskiej Wikipedii. Korzystamy z jego podzbioru zawierającego 50 milionów słów. Zbiór danych pobrać można za pomocą biblioteki datasets, a więcej dotyczących niego szczegółów znajduje się pod linkiem.

3 Preprocessing danych

Przed analizą korpus tekstu należy przygotować tak, by wyekstraktować z niego zbiory najpopularniejszych połączeń czasowników i rzeczowników. Początkowo, posiadaliśmy po prostu listę słów zawierającą 50 milionów elementów (słów, numerów, znaków specjalnych). Tekst należało zlematyzować, a użyty przez nas parser *WordNetLemmatizer* wymagał także przypisania każdemu wyrazowi odpowiedniej części mowy. Następnie, uzyskaliśmy listę 50 najpopularniejszych czasowników w zbiorze danych.

Word	**liczba wystąpień**
be	1254523
have	303680
make	65781
include	62390
use	54574
do	53249
become	51174
take	50655
write	39816
say	39421

Table 1: 10 najpopularniejszych czasowników w zbiorze danych wraz z liczbą wystąpień

Word	**liczba wystąpień**
begin	35516
give	34700
know	32130
play	32126
find	31777
call	30306
go	27409
release	27050
win	26159
lead	25546
receive	25472
follow	25114
leave	24454
come	24204
remain	21561
continue	21171
describe	21150
move	21037
work	20969
hold	20390
appear	19166
return	19157
see	18987
build	18548
produce	17660
create	17350
reach	17269
provide	17041
base	17006
allow	16849
name	16552
serve	16378
run	16256
show	16117
consider	15937
feature	15891
state	15667
start	15513
get	15401
record	15176

Table 2: Pozostałe 40 najpopularniejszych czasowników w zbiorze danych wraz z liczbą wystąpień

Następnie dla każdego czasownika wyznaczono 50 najpopularniejszych rzeczowników występujących po każdym z wybranych czasowników. Poniżej prezentujemy przykładowy wycinek danych:

rzeczownik	**liczba wystąpień**
part	3139
something	518
evidence	353
nothing	351
member	309
today	264
briefly	251
home	249
time	219
John	196

Table 3: 10 rzeczowników najczęściej występujących po słowie **be**

rzeczownik	**liczba wystąpień**
difficulty	361
sex	333
access	310
nothing	248
trouble	241
child	148
something	140
time	133
problem	116
fun	81

Table 4: 10 rzeczowników najczęściej występujących po słowie **have**

rzeczownik	**liczba wystąpień**
landfall	1154
use	517
contact	250
way	202
sense	170
room	146
reference	131
plan	121
appearance	119
fun	114

Table 5: 10 rzeczowników najczęściej występujących po słowie **make**

rzeczownik	liczba wystąpień
John	123
Best	105
work	91
William	80
part	80
Robert	72
member	71
James	69
David	57
song	52

Table 6: 10 rzeczowników najczęściej występujących po słowie **include**

rzeczownik	liczba wystąpień
steam	121
today	116
computer	46
element	40
change	37
drug	35
part	34
material	33
data	31
force	31

Table 7: 10 rzeczowników najczęściej występujących po słowie **use**

4 Analiza

W celu znalezienia związków pomiędzy czasownikami, policzono iloczyny pomiędzy zbiorami rzeczowników występujących po konkretnych czasownikach. Znaleziono kilka ciekawych zależności, które prezentujemy poniżej.

Słowa w iloczynie be i have
something
briefly
nothing
time
child
plan
people
concern
anything

Table 8: Iloczyn zbiorów rzeczowników najczęściej występujących po słowach **be** i **have**

Niestety, większość zbiorów nie niesie za sobą istotnych informacji, zawierając jedynie mało znaczące, generyczne rzeczowniki takie jak "something", "nothing" czy "anything". Przykładem takiej relacji są czasowniki **be** i **have** przedstawione powyżej.

Słowa w iloczynie start i begin
service
flight
play
life
work
September
negotiation
development
December
production
construction
operation
January
school

Table 9: Iloczyn zbiorów rzeczowników najczęściej występujących po słowach **start** i **begin**

Szczególnie szerokie są zbiory podobnych znaczeniowo rzeczowników, w tym przypadku **start** oraz **begin**. Widzimy tutaj zarówno miesiące, ale także długotrwałe procesy jak słowo "construction".

Słowa w iloczynie textbfhave i leave
office
room
everything

Table 10: Iloczyn zbiorów rzeczowników najczęściej występujących po słowach **have** i **leave**

Nieco ciekawiej sprawa wygląda dla czasowników **have** i **leave**. Występują tutaj rzeczowniki określające miejsca, ale takie, który przeciętny człowiek jest w stanie posiadać - office oraz room.

Słowa w iloczynie
home
career
book
history
process
song
music
material

Table 11: Iloczyn zbiorów rzeczowników najczęściej występujących po słowach **write** i **record**

Niektóre rzeczy można zarówno nagrać i napisać - chodzi tutaj głównie o treści muzyczne. W powyższym zbiorze można także zauważyć ciekawą zależność językową - można nagrać historię (kamerą?) oraz ją "napisać", w znaczeniu nieco przenośnym.

Słowa w iloczynie
aid
reinforcement
assistance
support
money
water
information
funding
training

Table 12: Iloczyn zbiorów rzeczowników najczęściej występujących po słowach **receive** i **provide**

Szerokie zbiory iloczynowe można zaobserwować w przypadku czasowników o w pewnym stopniu przeciwnym znaczeniu - na przykład **receive** oraz **provide**.

5 Podsumowanie

W badaniu użyto zbioru danych *Wikitext* by porównać znaczeniowo najpopularniejsze czasowniki poprzez ich otoczenie w zdaniach. Wyciągnięto szereg wniosków, w tym możliwość znalezienia w taki sposób podobnych znaczeniowo czasowników czy rzeczowników, które mogą posiadać wiele znaczeń.