

POLITECHNIKA WROCŁAWSKA
WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI



Sieci złożone

Sprawozdanie z laboratorium

AUTOR

Jędrzej Jamnicki

nr albumu: **254290**

kierunek: **Inżynieria systemów**
specjalizacja: **Inżynieria danych**

28 stycznia 2022

Streszczenie

Celem badań jest analiza słownictwa użytego w artykułach z wielojęzycznej encyklopedii internetowej - *Wikipedia* w oparciu o zbiór 4,604 artykułów spośród 15 kategorii.

1 Wstęp

Dane zostały pobrane z <https://snap.stanford.edu/data/wikispeedia.html> w formie plików w formacie *txt* wraz z przypisanymi im kategoriami takimi jak matematyka, sztuka czy historia.

Artykuły w "surowej" formie nie mogą zostać poddane analizie. Muszą zostać poddane wstępemu przetwarzaniu czyli swojego rodzaju oczyszczenia tekstu z m.in. takich szumów jak znaki specjalne (czasami są pożądane) Następnym etapem jest wyznaczenie unikalnego zbioru słów w formie podstawowej (hasłowej) zwanych dalej *tokenami* oraz określenie ich część mowy dla każdej z danych kategorii.

Do przetwarzania zbioru danych użyto języka programowania **Python**, wizualizacje przedstawiono przy użyciu biblioteki **altair**.

2 Analiza

Artykuły poddano wstępemu przetwarzaniu, kolejno:

- usunięto znaki specjalne oraz cyfry
- usunięto zduplikowane białe znaki
- usunięto słowa o długości mniejszej niż 3 znaki
- usunięto słowa wchodzące w skład tzw. stop listy
- sprowadzono litery do jednolitego formatu (małych liter)
- sprowadzono słowa do formy podstawowej (lematyzacja)

Do wyeliminowania niechcianych znaków w tekście użyto wyrażeń regularnych. Do lematyzacji oraz usunięcia słów ze stop listy wykorzystano bibliotekę **nltk** (korpus WordNet)

Zliczono wystąpienia tokenów dla każdej kategorii. Z wykorzystaniem metody **nltk.pos_tag** określono część mowy (*Part-of-Speech*) dla każdego z nich. Na podstawie ilości wystąpień oraz sumarycznej ilości unikalnych tokenów we wszystkich artykułach dla poszczególnej kategorii wyznaczono częstotliwości występowania.

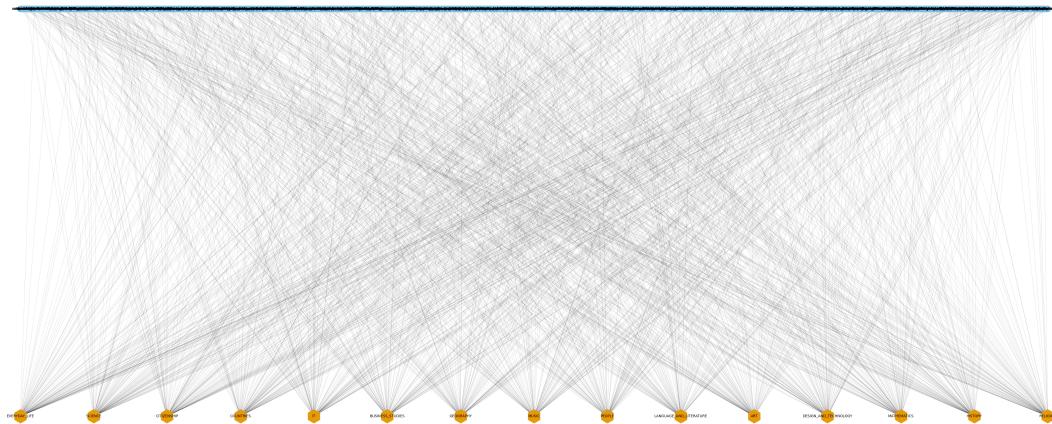
Otrzymane wartości zapisano do ramki danych biblioteki **Pandas** prezentującej się następująco:

Category	Token	Part-of-Speech	Occurrences	Frequency
History	klan	Noun	188	0.001222
History	hmas	Adverb	131	0.000851
History	titanic	Adjective	108	0.000702
History	seacole	Verb	98	0.000637
History	yugoslavia	Adverb	86	0.000559
History	colditz	Noun	77	0.000500
:	:	:	:	:
Countries	celtic	Noun	8	0.000101
Countries	galicia	Adjective	8	0.000101
Countries	zimbabwean	Noun	5	0.000063
Countries	lankans	Verb	5	0.000063

1322306 rows × 5 columns

Tabela 1: Dane każdego z tokenów.

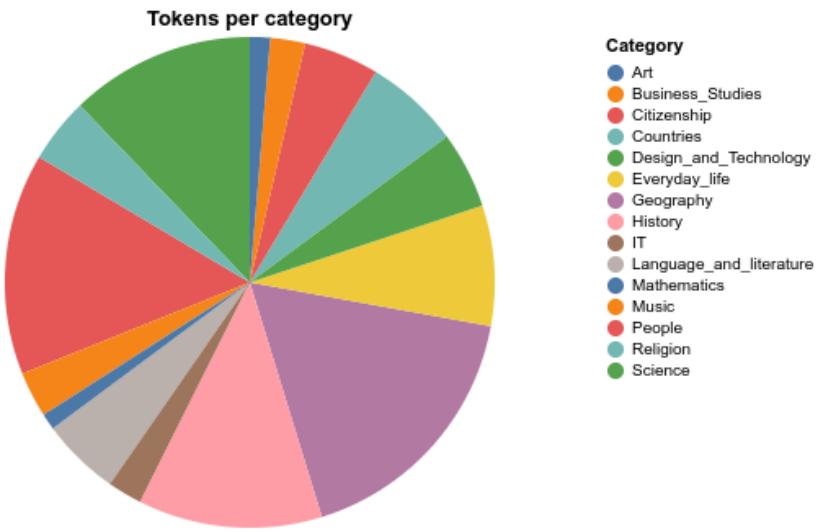
Związek tokenów z kategoriami przedstawiono również w postaci grafu dwudzielnego, w którym zbiór górnych wierzchołków stanowią tokeny a dolnych - kategorie.



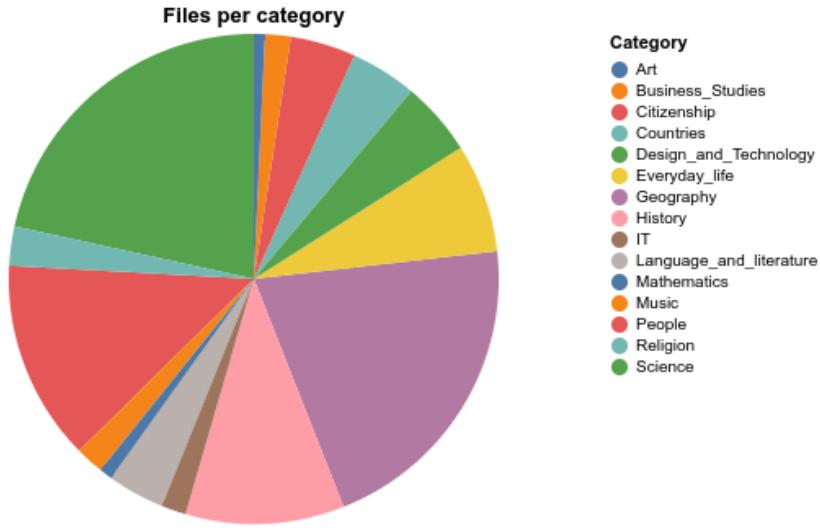
Rysunek 1: Graf dwudzielny reprezentujący zbiór danych.

Category	Tokens	Files
Geography	220024	1084
People	184725	689
Science	154367	1122
History	153855	545
Everyday_life	100361	374
Countries	79565	229
Language_and_literature	65750	196
Design_and_Technology	63841	254
Citizenship	62802	224
Religion	54071	134
Music	38039	97
Business_Studies	29105	88
IT	28624	85
Art	16890	38
Mathematics	13486	45

Tabela 2: Ilość unikalnych tokenów oraz plików dla kategorii.



Rysunek 2: Ilość unikalnych tokenów dla każdej z kategorii.



Rysunek 3: Ilość artykułów dla każdej z kategorii.

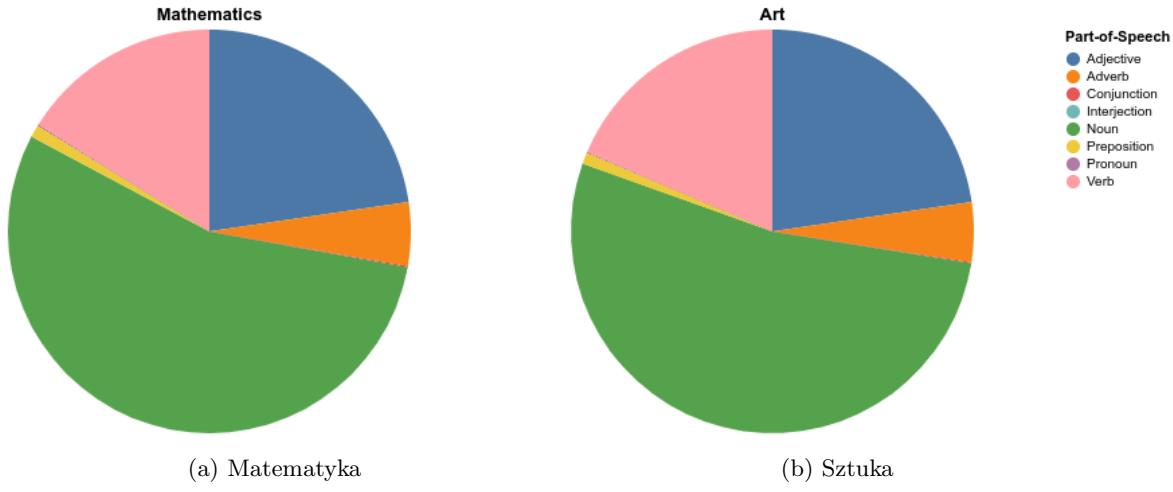
Tab. 2 oraz rys. 2, 3 pokazują, że każda z kategorii reprezentuje nieproporcjonalną część zbioru danych, co nie napawa optymizmem przed kolejnymi analizami.

Na podstawie wartości z tab. 1 wyznaczono procentowy udział części mowy dla każdej kategorii.

Category	Noun	Adjective	...	Adverb
Geography	0.5965	0.2371	...	0.0411
People	0.5716	0.2335	...	0.0472
Science	0.5803	0.2304	...	0.0479
History	0.5627	0.2445	...	0.0458
Everyday_life	0.5817	0.2240	...	0.0464
Countries	0.5961	0.2405	...	0.0401
Language_and_literature	0.5881	0.2340	...	0.0481
Design_and_Technology	0.5792	0.2217	...	0.0471
Citizenship	0.5555	0.2473	...	0.0439
Religion	0.5685	0.2439	...	0.0469
Music	0.5713	0.2312	...	0.0522
Business_Studies	0.5633	0.2339	...	0.0469
IT	0.5591	0.2260	...	0.0523
Art	0.5571	0.2385	...	0.0504
Mathematics	0.5708	0.2354	...	0.0529

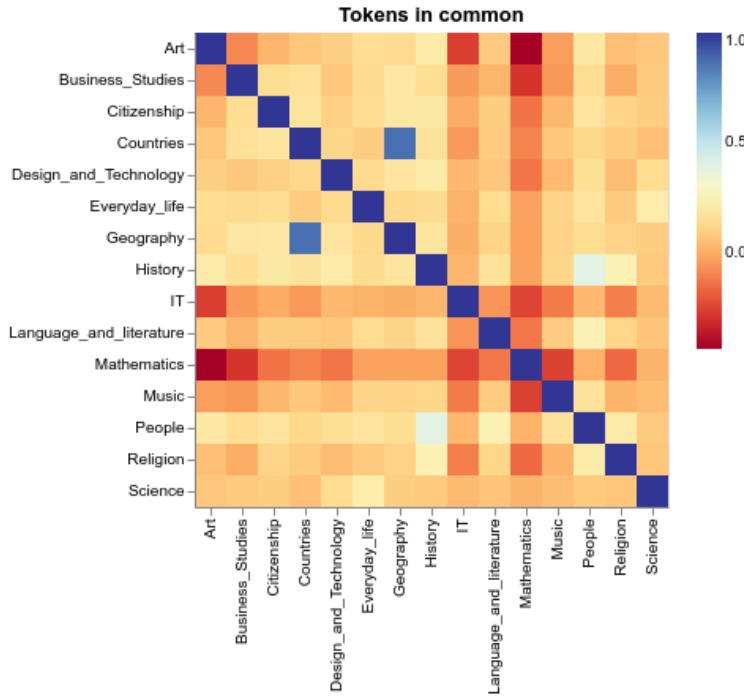
Tabela 3: Procentowy udział części mowy według kategorii.

Diagramy kołowe stworzone dla każdej kategorii pokazały, że nie odbiegają od siebie nawzajem pod względem procentowego udziału części mowy w artykułach. Diagramy dla dwóch kategorii przedstawiono poniżej:



Rysunek 4: Procentowy udział części mowy dla kategorii.

Na zakończenie zbadano ilość wspólnych tokenów między kategoriami. W tym celu stworzono macierz, w której każdy element odpowiada procentowej wartości tokenów wspólnych. Do wizualizacji wykorzystano diagram korelacji przedstawiony na rys. 5 poniżej.



Rysunek 5: Procentowa wartość wspólnych tokenów między kategoriami.

3 Wnioski

Tab. 3 oraz rys. 4a, 4b jasno pokazują, że części mowy są ściśle między sobą powiązane i nie odbiegają procentowym udziałem w artykułach między kategoriami.

Na rys. 5 zauważono silną korelację między kategoriami:

- Geografia - Kraje (aż 86.62%)
- Ludzie - Historia (aż 39.41%)
- IT – Sztuka (jedynie 28.36%)
- Matematyka - Sztuka (jedynie 19.90%)

Części wspólne lub przeciwnostawne wydają się być logiczne. Ponadto Historia wykazuje się wysoką średnią wartością wspólnych tokenów między wszystkimi innymi kategoriami, bo aż 38.31%. W końcu historia to “wszystko w przeszłości”...