

TASS

Projekt 2

Radosław Pietkun, Sebastian Pietras

Spis treści

Zadanie	2
Dane	2
Technologie	3
Uruchomienie	3
Analiza	3
Wyniki	5
Podsumowanie	12

Zadanie

Temat nr 9: Przedstaw ewolucję znaczenia określonego terminu na podstawie analizy jego kolokacji z innymi słowami w starodrukach.

Celem naszego projektu jest przedstawienie ewolucji znaczenia słów na podstawie analizy ich współwystępowania z innymi słowami w starodrukach. Całość zgromadzonych danych została przeanalizowana, a uzyskane wyniki i wnioski zostały zaprezentowane w postaci raportu końcowego.

Dane

Jako źródło danych do analizy wykorzystaliśmy ebooki dostępne w ramach [projektu Gutenberg](#), które można pobrać w formacie tekstowym. Większość książek tam dostępnych pochodzi z XVII, XVIII, XIX i XX wieku. Możemy znaleźć tam wiele różnorodnych tekstów, lecz trudno znaleźć takie, które pochodziłyby sprzed XVI wieku.

Spośród dostępnych tekstów ręcznie wyselekcjonowaliśmy i pobraliśmy niektóre z nich, wraz z zapisaniem informacji o dacie powstania. Książki podzieliliśmy na dwie grupy: stare (z XVI i XVII wieku) i nowe (z XIX i XX wieku). Analiza ewolucji znaczenia słów została przeprowadzona między tymi dwoma grupami. Dokładny spis książek z każdej grupy został przedstawiony poniżej.

Stare książki (XVI i XVII wiek):

- [Ben Jonson - The Alchemist \(1610\)](#)
- [Ben Jonson - Volpone; Or, The Fox \(1606\)](#)
- [Ben Jonson - Every Man in His Humor \(1598\)](#)
- [The Complete Works of William Shakespeare \(1585–1613\)](#)
- [Thomas Kyd - The Spanish Tragedie \(1587\)](#)
- [John Webster - The Duchess of Malfi \(1623\)](#)
- [John Webster - The White Devil \(1612\)](#)

Nowe książki (XIX i XX wiek):

- [James Joyce - Ulysses \(1920\)](#)
- [James Joyce - Dubliners \(1914\)](#)
- [Joseph Conrad - Heart of Darkness \(1899\)](#)
- [Joseph Conrad - Lord Jim \(1899\)](#)
- [Franz Kafka - Metamorphosis \(1915\)](#)
- [Thomas Hardy - Tess of the d'Urbervilles: A Pure Woman \(1891\)](#)
- [John Galsworthy - The Forsyte Saga \(1906\)](#)
- [H. G. Wells - The Time Machine \(1895\)](#)
- [Henry James - The Turn of the Screw \(1898\)](#)

Technologie

Projekt został wykonany z użyciem języka Python. Do analizy tekstu wykorzystaliśmy popularne pakiety do przetwarzania języka naturalnego: [nltk](#) i [spacy](#). Do pracy z sieciami posłużyła nam biblioteka [NetworkX](#). Wizualizacja wyników była możliwa dzięki notatnikom [Jupyter](#) oraz pakietom [matplotlib](#) i [pyvis](#).

Uruchomienie

Aby przeprowadzić analizę ponownie najpierw należy utworzyć wirtualne środowisko z interpreterem języka Python w wersji 3.9. Następnie należy w nim zainstalować potrzebne pakiety, których spis znajduje się w pliku `requirements.txt`. Do tego celu zalecamy wykorzystanie systemu zarządzania środowiskami [conda](#) i wskazanie na plik `environment.yml`, który zawiera konfigurację środowiska, włącznie z poleceniem instalacji pakietów z `requirements.txt`.

Dane do analizy należy umieścić w katalogu `data`. Podkatalog 1 powinien zawierać pliki z jedną grupą tekstów, a podkatalog 2 z drugą grupą. Nazwy plików nie mają znaczenia, wczytane zostaną wszystkie pliki.

Następnie należy uruchomić środowisko Jupyter (zalecamy poprzez polecenie `python -m jupyterlab`), otworzyć plik `analysis.ipynb` i uruchomić analizę. W zależności od posiadanego sprzętu i analizowanych danych, całość może potrwać trochę czasu (w naszym przypadku było to około dwóch godzin). Wymagane jest również ok. 2GB wolnej pamięci RAM.

Analiza

Wstępna analiza danych polegała na przygotowaniu zebranych dokumentów do dalszej pracy. Wykorzystaliśmy w tym celu narzędzia `spacy`. Faza ta obejmowała kilka operacji:

- Wyszczególniliśmy elementy tekstu (słowa), które były właściwymi elementami danego dokumentu (pominęliśmy niepotrzebne adnotacje, nagłówki, informacje o licencjach). Wykorzystaliśmy tutaj informację, że wszystkie książki w ramach projektu Gutenberg zawierają specjalne i jednakowe oznaczenia początku i końca właściwej treści książki.
- Wykluczyliśmy słowa pospolite, które pojawiają się często w języku. Pominęliśmy również nazwy własne, symbole specjalne oraz znaki interpunkcyjne. Do filtracji wykorzystaliśmy atrybuty klasy `Token` takie jak np. `is_stop`, `is_digit`, `is_punct`.
- Ujednoliliśmy też słowa przez zignorowanie wielkości liter oraz różnic w formach gramatycznych (chcieliśmy, by rozważane słowo było znajdowane za każdym razem niezależnie od jej formy gramatycznej). Słowa sprowadziliśmy do ich form podstawowych pomijając odmienne końcówki (atrybut `lemma_` klasy `Token`).

Właściwa analiza danych polegała na zbadaniu właściwości bigramów, czyli par słów. Poszczególne kolokacje oceniliśmy na podstawie dwóch kryteriów:

- 1) Przy pomocy miary Jaccarda wyznaczyliśmy istotność kolokacji dla danego słowa, dzięki czemu wyodrębniliśmy kolokacje charakterystyczne dla tego słowa.

- 2) Na podstawie częstości występowania danej kolokacji w korpusie tekstów ocenialiśmy, czy rzeczywiście jest ona istotna. Chcieliśmy się w ten sposób uchronić przed silnymi kolokacjami pod względem miary Jaccarda, ale występującymi np. tylko raz w całym korpusie. Dlatego wprowadziliśmy pewną wartość progową, poniżej której wszystkie kolokacje były odrzucane w dalszej analizie.

Następnie dla danej grupy tekstów pochodzących z jednej epoki stworzyliśmy graf, którego wierzchołkami są słowa występujące w tych tekstach. Waga krawędzi łączącej wierzchołki odpowiadała sile kolokacji słów reprezentowanych przez wierzchołki (im kolokacja była bardziej charakterystyczna, tym waga tej krawędzi była większa). Utworzone zostały dwa grafy, po jednym dla każdej rozpatrywanej epoki.

Elementem, który postanowiliśmy zmienić względem naszej koncepcji, jest rodzaj grafu. Uznaliśmy, że do badań bardziej odpowiednie będą grafy skierowane, ponieważ kolejność słów w kolokacji (wskazywana przez zwrot krawędzi) jest istotna i nie powinna być dowolna.

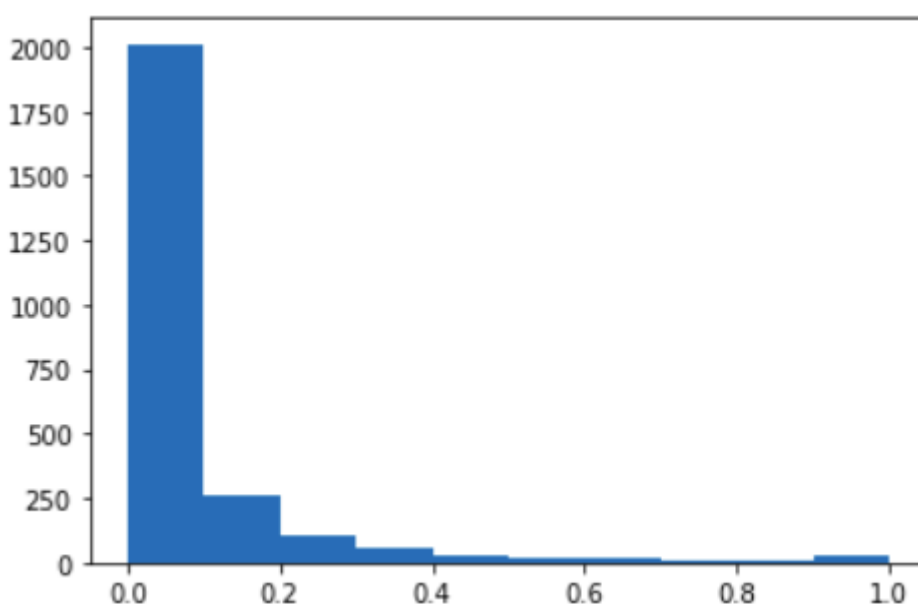
Grafy zostały uzgodnione, by zawierały ten sam zbiór wierzchołków. W pierwszej wersji była to suma zbiorów wierzchołków grafów pierwotnych; tzn. w przypadku słów występujących tylko w jednym grafie, zostały one umieszczone w drugim grafie jako wierzchołki izolowane. W drugiej wersji wyznaczyliśmy część wspólną zbiorów wierzchołków. Chcieliśmy się w ten sposób zabezpieczyć przed sytuacją, w której wiele słów występowałoby tylko w jednym korpusie. Mogłoby to bowiem przyczynić się do uzyskania niemiernie niskich wartości podobieństwa cosinusowego, co mogłoby błędnie świadczyć o dużej zmianie znaczenia słowa. Dzięki takim przekształceniom każde słowo występujące w jednym grafie miało swój odpowiednik w drugim grafie. Operacja ta pozwoliła później na wygodne porównywanie własności węzłów w grafach.

Do analizy ewolucji znaczenia słów wykorzystaliśmy macierze wag. Dla każdego słowa obliczona została wartość podobieństwa wierzchołków reprezentujących to słowo w obu grafach. Każdy wierzchołek reprezentowany był przez wektor wag krawędzi łączących go z innymi wierzchołkami, czyli przez odpowiedni wiersz z macierzy wag. Jako miarę podobieństwa wykorzystaliśmy podobieństwo cosinusowe między tymi wektorami. Im mniejsza wartość podobieństwa między daną parą wierzchołków, tym dane słowo tworzyło istotnie różne kolokacje w różnych epokach. Może to być przesłanką do stwierdzenia, że zmieniło się znaczenie słowa reprezentowanego przez ten wierzchołek.

Wyniki

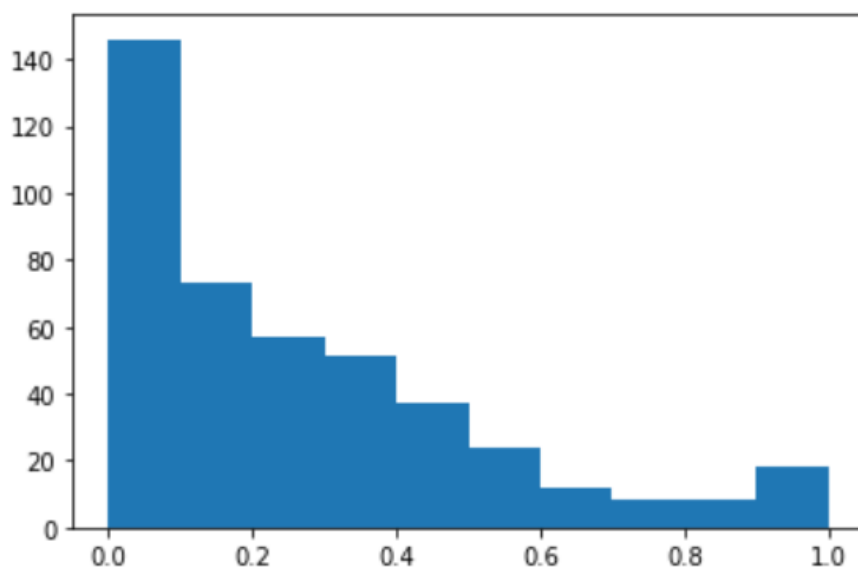
Słownik stworzony na podstawie pierwszego korpusu składa się z 24190 słów, natomiast słownik dla drugiego korpusu z 27124 słów. Z kolei wspólny słownik stworzony na podstawie obu korpusów (ich część wspólna) składa się z 10540 słów.

Pierwsze badania przeprowadziliśmy bez filtracji częstotliwości występowanie kolokacji w korpusie tzn. wystarczyło jednokrotne pojawienie się jakiejś kolokacji, by została ona uwzględniona w dalszej analizie. Na rys. 1 jest przedstawiony histogram rozkładu uzyskanych odległości cosinusowych. Odległości równe zero i jeden zostały pominięte, ponieważ oznaczają one odpowiednio, że dane słowo tworzy całkowicie różne kolokacje w każdej epoce i że dane słowo tworzy dokładnie takie same kolokacje w każdej epoce. Żadna z tych informacji nie ma dużej wartości z punktu widzenia analizy zmienności znaczenia słów. Liczba słów dla których odległość cosinusowa jest z przedziału (0, 1) wynosi 2524. Na podstawie rys. 1 można zauważyć, że zdecydowana większość słów w każdej epoce tworzyła istotnie różne kolokacje. Nie jest to jednak miarodajna obserwacja, którą można by efektywnie wykorzystać w dalszej analizie.



Rys. 1. Histogram rozkładu uzyskanych odległości cosinusowych, wartości równe zero i jeden zostały pominięte

W następnym kroku odfiltrowaliśmy wszystkie kolokacje, które pojawiały się w korpusach mniej niż 3 razy. Otrzymany histogram rozkładu uzyskanych odległości cosinusowych jest przedstawiony na rys. 2. Tym razem liczba słów dla których odległość cosinusowa jest z przedziału (0, 1) zmalała do 422. Uzyskany rozkład jest już bardziej zrównoważony. Na rys. 3 przedstawiona jest lista 30 słów o najmniejszej wartości odległości cosinusowej (z pominięciem odległości zerowej) razem z jej wartością.



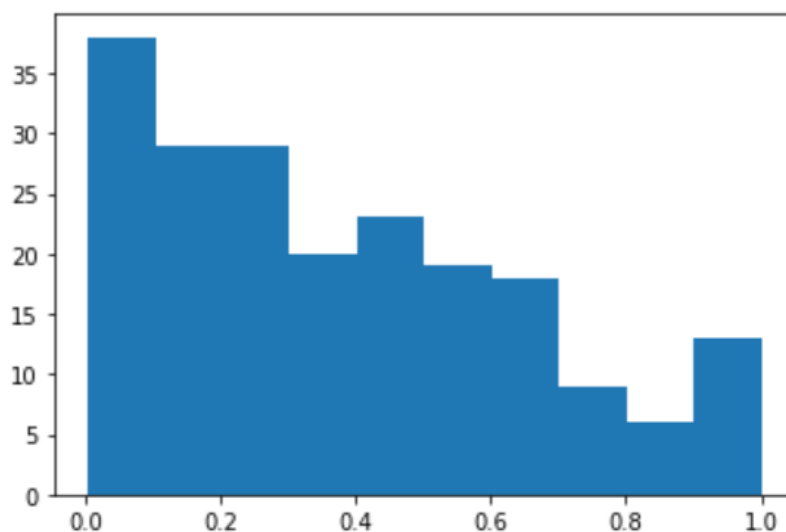
Rys. 2. Histogram uzyskanych odległości cosinusowych po odfiltrowaniu kolokacji występujących w korpusie mniej niż 3 razy, węzły dla odległości równej zero i jeden zostały pominięte

```
[('order', 0.0014833820760245873),
 ('ye', 0.0017049051057104556),
 ('call', 0.0017082107559337626),
 ('move', 0.001723490443222293),
 ('wonder', 0.0018154323161513504),
 ('horse', 0.002185989107451322),
 ('wait', 0.0023837926802108144),
 ('past', 0.0029550210463275975),
 ('dress', 0.0031244202965065495),
 ('clear', 0.0032164160416067685),
 ('burn', 0.0033950432948832137),
 ('tall', 0.0034076971107835975),
 ('glance', 0.0037189639747779554),
 ('fat', 0.004142634660548695),
 ('late', 0.004291354321662616),
 ('person', 0.004440889702350154),
 ('cover', 0.004644905284168357),
 ('upper', 0.0046832472649797905),
 ('doctor', 0.0047613719227488985),
 ('soon', 0.005005947804094808),
 ('ship', 0.005760184764551474),
 ('tongue', 0.005958666772010377),
 ('shine', 0.006828368818131571),
 ('foot', 0.007303205595193304),
 ('suppose', 0.007435690507210641),
 ('serve', 0.007647028510471098),
 ('land', 0.007680060314250044),
 ('dream', 0.008242211765560904),
 ('brow', 0.008259815432215201),
 ('devil', 0.009061358754957278)]
```

Rys. 3. Lista 30 słów o najmniejszej wartości odległości cosinusowej po odfiltrowaniu kolokacji występujących w korpusie mniej niż 3 razy

Następnie odfiltrowaliśmy wszystkie kolokacje, które pojawiały się w korpusach mniej niż 5 razy. Otrzymany histogram uzyskanych odległości cosinusowych jest przedstawiony na rys. 4.

Liczba słów dla których odległość cosinusowa jest z przedziału (0, 1) zmalała do 198. Uzyskany rozkład jest jeszcze bardziej równomierny. Można już zdecydowanie łatwiej wyszczególnić te słowa, które cechowały się małą wartością odległości cosinusowej. Lista 30 pierwszych słów wraz z odpowiadającymi im miarami podobieństwa jest pokazana na rys. 5.



Rys. 4. Histogram rozkładu uzyskanych odległości cosinusowych po odfiltrowaniu kolokacji występujących w korpusie mniej niż 5 razy, węzły dla odległości równej zero i jeden zostały pominięte

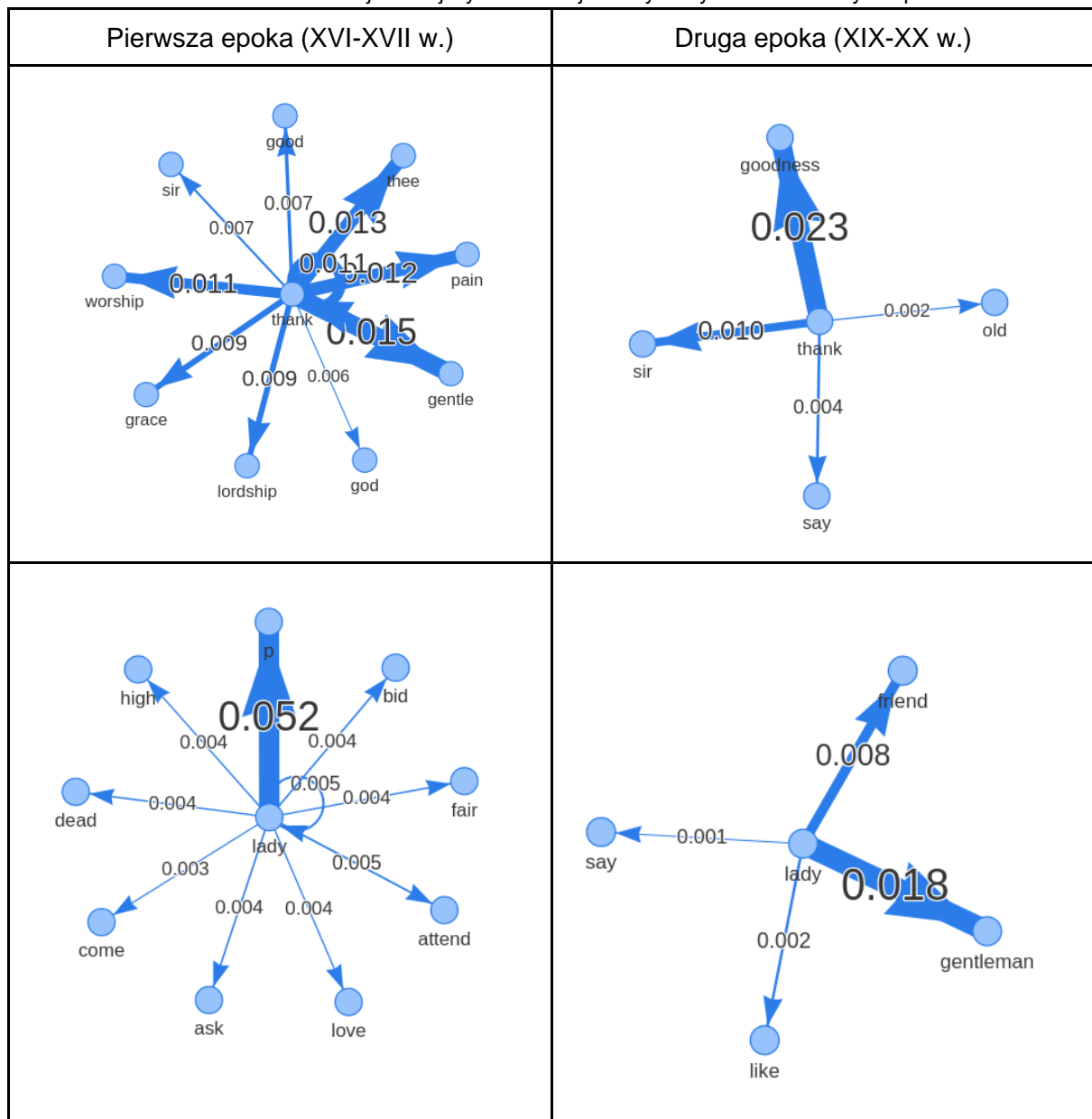
```
[('lady', 0.003484425138998452),
 ('past', 0.004594905252338177),
 ('colour', 0.005649907517195315),
 ('enter', 0.005863003125294315),
 ('glad', 0.011348646708614012),
 ('money', 0.012613839520366774),
 ('answer', 0.015129220970544874),
 ('tall', 0.016110270184075008),
 ('home', 0.01653169052370461),
 ('sound', 0.01786448309723066),
 ('matter', 0.0199904954354292),
 ('soon', 0.02420323308457267),
 ('meet', 0.026260354839165386),
 ('bright', 0.028415621156803048),
 ('spend', 0.028421751381830614),
 ('sit', 0.02968226270419058),
 ('far', 0.03089476951873908),
 ('talk', 0.032519730452764684),
 ('place', 0.05739964283871083),
 ('walk', 0.060625895503122605),
 ('new', 0.06132877024985742),
 ('short', 0.061367232029001864),
 ('near', 0.06358789776015285),
 ('room', 0.06475757035245955),
 ('tear', 0.06613763809510262),
 ('married', 0.0745689219258304),
 ('grow', 0.07738995988889905),
 ('bear', 0.07752737115707808),
 ('lie', 0.07823386445750606),
 ('thank', 0.08034395056187073)]
```

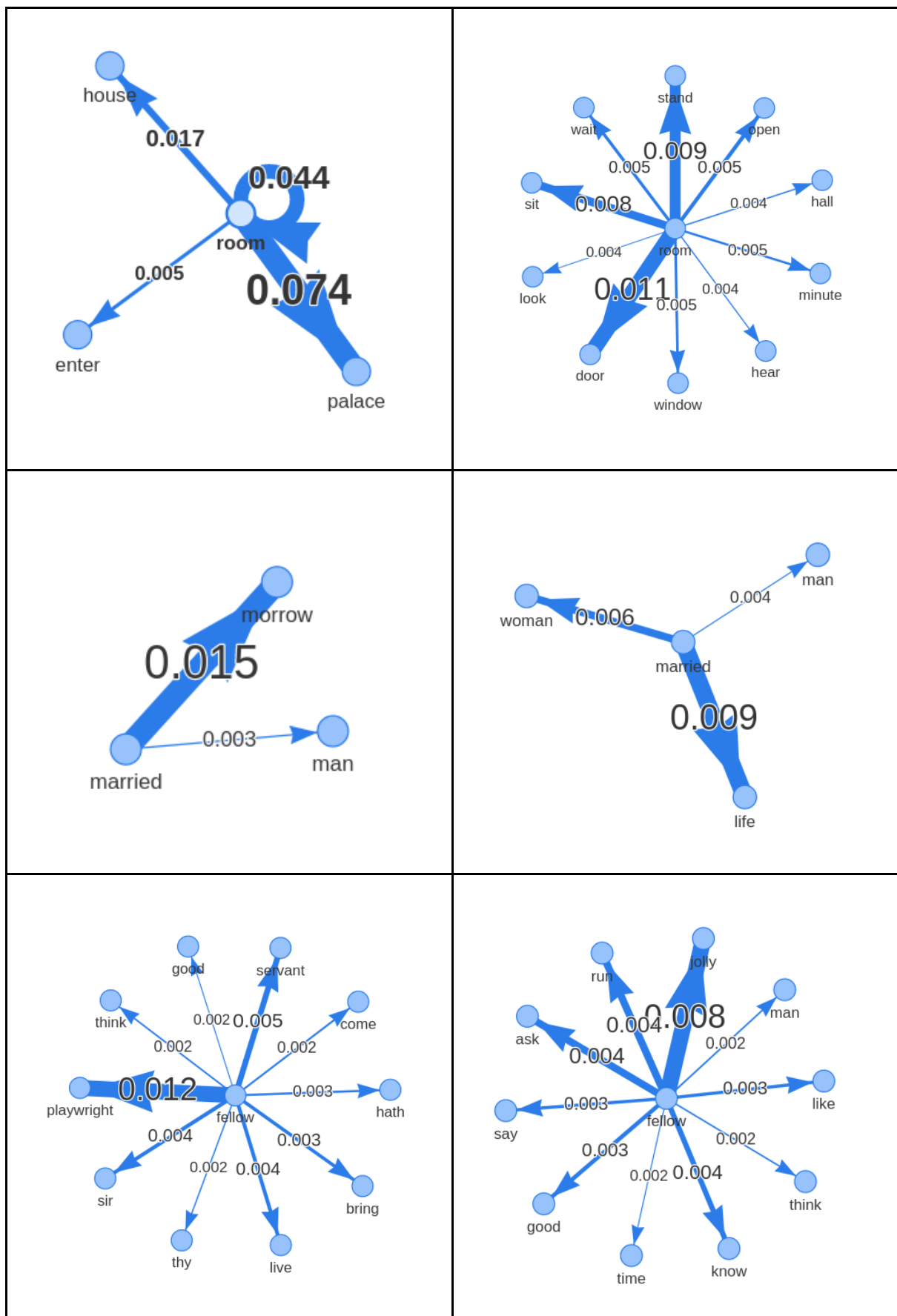
Rys. 5. Lista 30 słów o najmniejszej wartości odległości cosinusowej po odfiltrowaniu kolokacji występujących w korpusie mniej niż 5 razy

W tabeli 1 umieściliśmy porównanie kolokacji tworzonych przez wybrane słowa w obu epokach. Wierzchołki reprezentują tutaj słowa, a krawędzie kolokacje. Waga krawędzi

odpowiada sile kolokacji według miary Jaccarda. Na każdym grafie zostało uwzględnionych maksymalnie 10 połączeń o największej wadze. Analizowane słowa mają wartości odległości cosinusowej z przedziału (0; 0,1).

Tabela 1. Porównanie najważniejszych kolokacji dla wybranych słów w różnych epokach





Na podstawie tabeli 1 można sformułować kilka ciekawych wniosków. Na przykład słowo “thank” w pierwszej epoce często tworzyło kolokacje ze słowami wiążącymi się z przejawem kultu religijnego, takimi jak “God”, “worship”, “grace”. Natomiast w drugiej epoce liczba istotnych kolokacji tworzonych przez słowo “thank” znacznie zmalała. Zdecydowanie wyróżniała się tutaj kolokacja “thank goodness”.

Na przykładzie słowa “lady” można zauważyć, że w drugiej epoce pojawiła się silna kolokacja między słowami “lady” oraz “gentleman”. Powstała ona najprawdopodobniej ze skrócenia pierwotnej popularnej kolokacji “ladies and gentlemen” po pominięciu słowa ‘and’ na etapie wstępnej analizy (ponieważ jest to tzw. stopword).







Silna kolokacja “lady-p”, która została zidentyfikowana dla pierwszej epoki pochodzi z tekstu “Volpone; Or, The Fox” i oznacza jedną z bohaterek dramatu: Lady P. Zaburza to nieco obraz kolokacji tworzonych przez to słowo, ponieważ jest to nazwa własna, która nie została rozpoznana na etapie początkowej analizy. Można by temu zapobiec przez wprowadzenie dodatkowego filtra, który by wymagał, by każde słowo w kolokacji składało się np. z co najmniej 3 znaków.

Zaobserwowaliśmy również kilka słów, które pojawiły się w istotnych kolokacjach w pierwszej epoce, natomiast z czasem całkowicie wypadły z języka i nie występują już one we współczesnym języku (np. słowa “thee”, “thy”).

Analizując przykładowe słowa z tabeli 1 można zauważyć, że zastosowana przez nas miara istotności kolokacji korzystająca ze wskaźnika Jaccarda okazała się dobrym podejściem. Udało się nam bowiem wyszczególnić kolokacje, które rzeczywiście występują w języku i które można uznać za charakterystyczne dla danego słowa.

Dodatkowo w tabeli 2 umieściliśmy porównanie kolokacji tworzonych przez wybrane słowa, dla których odległość cosinusowa wynosi 1. Świadczy to zatem o stałości danego związku w języku. Na każdym grafie zostało uwzględnionych maksymalnie 10 połączeń o największej wadze. Analizując grafy w tabeli 2 można zauważyć, że dane słowo tworzyło w takich sytuacjach kolokacje tylko z jednym innym słowem.

Tabela 2. Porównanie najważniejszych kolokacji dla wybranych słów o wartości odległości cosinusowej równej 1 w różnych epokach

Pierwsza epoka (XVI-XVII w.)	Druga epoka (XIX-XX w.)
 <p>north</p> <p>0.083</p> <p>south</p>	 <p>north</p> <p>0.054</p> <p>south</p>
 <p>lock</p> <p>0.017</p> <p>door</p>	 <p>lock</p> <p>0.013</p> <p>door</p>
 <p>palm</p> <p>0.003</p> <p>hand</p>	 <p>palm</p> <p>0.006</p> <p>hand</p>

Podsumowanie

Na podstawie przeprowadzonych badań nie udało się nam jednoznacznie wskazać słów, których znaczenie by w znaczącym stopniu ewoluowało na przestrzeni epok. Być może uwzględnienie innego lub większego zbioru tekstów przyczyniłoby się do większej jednoznaczności uzyskanych wyników.

Udało nam się natomiast zaobserwować, jak zmieniają się najbardziej charakterystyczne kolokacje z biegiem czasu. W zależności od epoki wiele słów tworzyło silne kolokacje z różnym zestawem wyrazów. Zaobserwowaliśmy również, że niektóre słowa w jednej epoce tworzyły wiele różnych kolokacji, podczas gdy w drugiej epoce liczba takich kolokacji była niewielka. Może to świadczyć o zawężeniu lub rozszerzeniu znaczenia danego słowa w języku, jak również o tym, że niektóre związki wyrazowe wypadły z użycia, a niektóre się dopiero pojawiły.

Przeprowadzone badania skłoniły nas do refleksji, że język jest żywym tworem, który ciągle ewoluuje. Analiza kolokacji wyrazowych może być silnym narzędziem do analizy jego zmienności.