NLP Project - FCIS ‘23

Martyna Kuśmierz,
Wiktoria Koniecko

Zbiór danych

Zbiór danych zawiera około 20000 artykułów.
Typowy artykuł składa się z informacji na wstępie takich jak temat, grupa dyskusyjna.
Następnie jest część główna artykułu.
Na końcu czasem występuje podpis czy też inspirujący cytat niezwiązany z tematem.

```
Xref: cantaloupe.srv.cs.cmu.edu alt.atheism:51060
Path: cantaloupe.srv.cs.cmu.edu!crabapple.srv.cs.c
etsys!ibmpcug!mantis!mathew
From: mathew <mathew@mantis.co.uk>
Newsgroups: alt.atheism,alt.atheism.moderated,news
Subject: Alt.Atheism FAQ: Introduction to Atheism
Summary: Please read this file before posting to a
Keywords: FAQ, atheism
Message-ID: <19930405122245@mantis.co.uk>
Date: Mon, 5 Apr 1993 12:22:45 GMT
Expires: Thu, 6 May 1993 12:22:45 GMT
Followup-To: alt.atheism
Distribution: world
Organization: Mantis Consultants, Cambridge. UK.
Approved: news-answers-request@mit.edu
Supersedes: <19930308134439@mantis.co.uk>
Lines: 646
```

Cel biznesowy

Naszym celem jest klasteryzacja artykułów, która może być wykorzystana np. w systemie rekomendacji podobnych treści do przeczytania.

W naszym przypadku istotnym założeniem jest to, że do klasteryzacji wykorzystujemy tylko główny tekst artykułu, to znaczy, że nie wykorzystujemy informacji zawartych we wstępie.

Preprocessing

W ramach prerocessingu wykonałyśmy następujące kroki:

1. Formatowanie tekstu w celu usunięcia znaków nowej linii.
2. Oddzielenie informacji na wstępie od głównej treści artykułu.
3. Usunięcie pustych artykułów.
4. Sprawdzenie w jakich językach są artykuły (tylko pojedyncze artykuły były w innych językach niż angielski, jednak często wynikało to z błędnej klasyfikacji w powodu długości tekstu lub cytatu w obcym języku.

Preprocessing

5. Usunięcie inspirujących cytatów w zakończeniu oraz wszystkich niestandardowych znaków, słów i cyfr.

6. Usunięcie stopwords, interpunkcji, spacji i lematyzacja słów.

7. Sprawdzenie najczęściej występujących słów w artykułach.

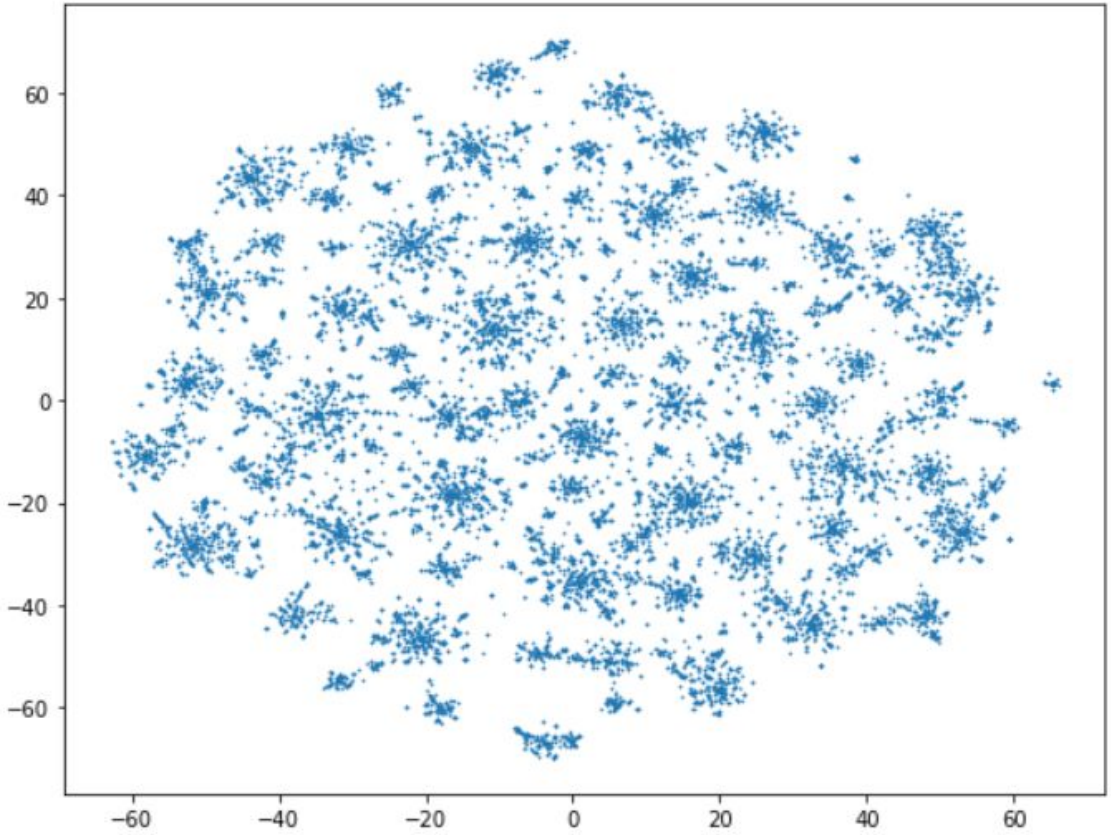
```
[('know', 3.06866705135603),
 ('like', 3.2659160696008187),
 ('think', 3.6119538148064296),
 ('good', 4.463905987688864),
 ('time', 4.487763713080168),
 ('people', 4.537542662116041),
 ('use', 5.0873724489795915),
 ('want', 5.321547698465643),
 ('way', 5.443193449334698),
 ('work', 5.48059086224665),
 ('look', 5.572476423332169),
 ('say', 5.572476423332169),
 ('thing', 5.706008583690987),
 ('come', 5.774158523344191),
 ('need', 5.950764640059679),
 ('go', 5.96411214953271),
 ('find', 5.973043803818794),
 ('get', 6.0249244712990935),
 ('try', 6.043181818181818),
```

Preprocessing

8. Wektoryzacja słów.
9. Tfidf model.
10. NMF.
11. Normalizacja wektorów.

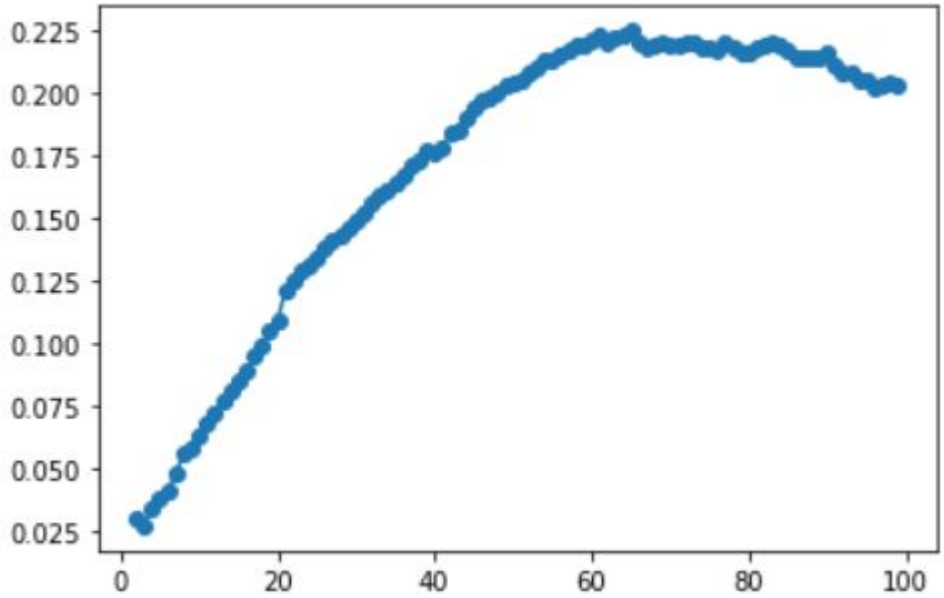
T-SNE

Wizualizacja dokumentów w postaci punktów za pomocą T-SNE.



Silhouette score

Optymalna liczba klastrów została wybrana przy pomocą silhouette score. W naszym przypadku widzimy, że liczba w przedziale 45-70 wydaje się być optymalna.



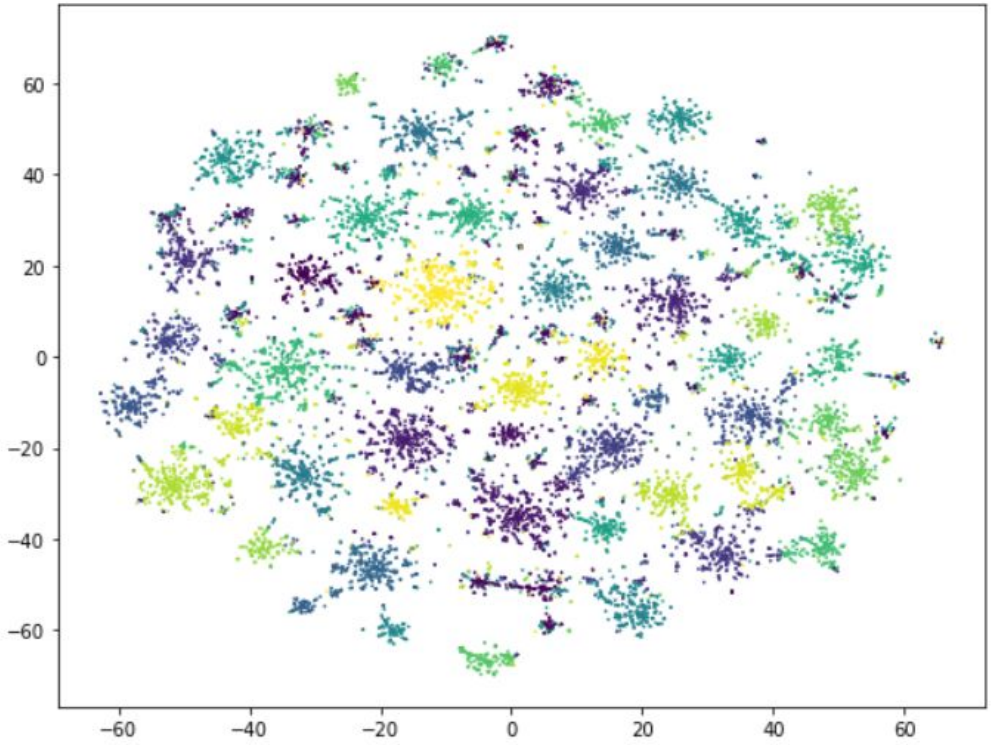
Liczba klastrów

Naszym celem było pogrupowanie artykułów w bardziej szczegółowe grupy teamatyczne. Przykładowo artykuł dotyczący Windowsa i Linuxa można umieścić w jednej grupie dotyczącej systemów operacyjnych, czy też informatyki, jednak zdecydowałyśmy się na mniej ogólny podział.

Dla naszego przypadku wybrałyśmy liczbę 50 klastrów, która nie jest jeszcze za duża.

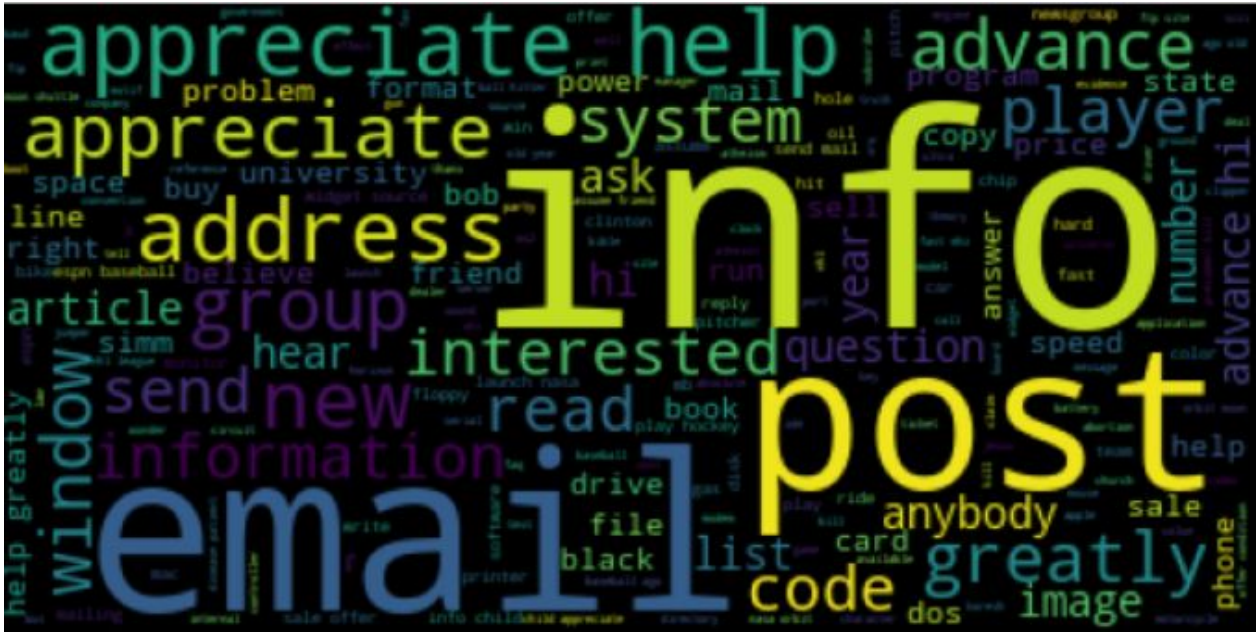
KMeans

Wyniki algorytmu dla 50 klastrów.



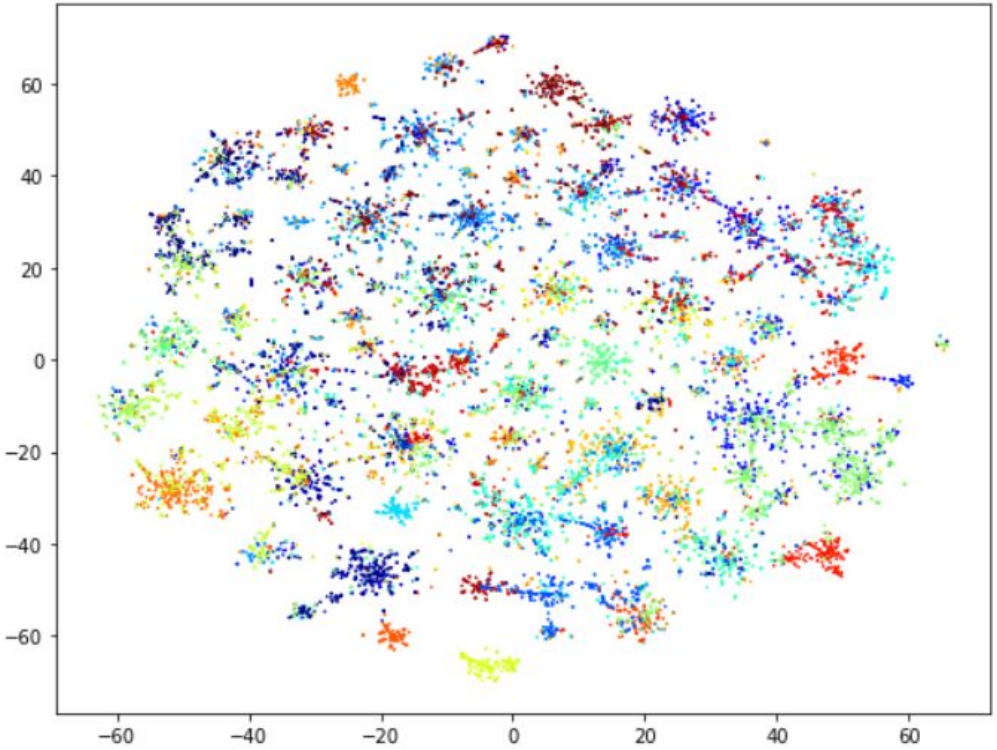
KMeans

Chmura słów pokazująca jakie słowa miały znaczenie w przykładowo wybranych klastrze.



GaussianMixture

Niestety niezależnie od parametrów wyniki przy użyciu tego algorytmu były obarczone dużym szumem.

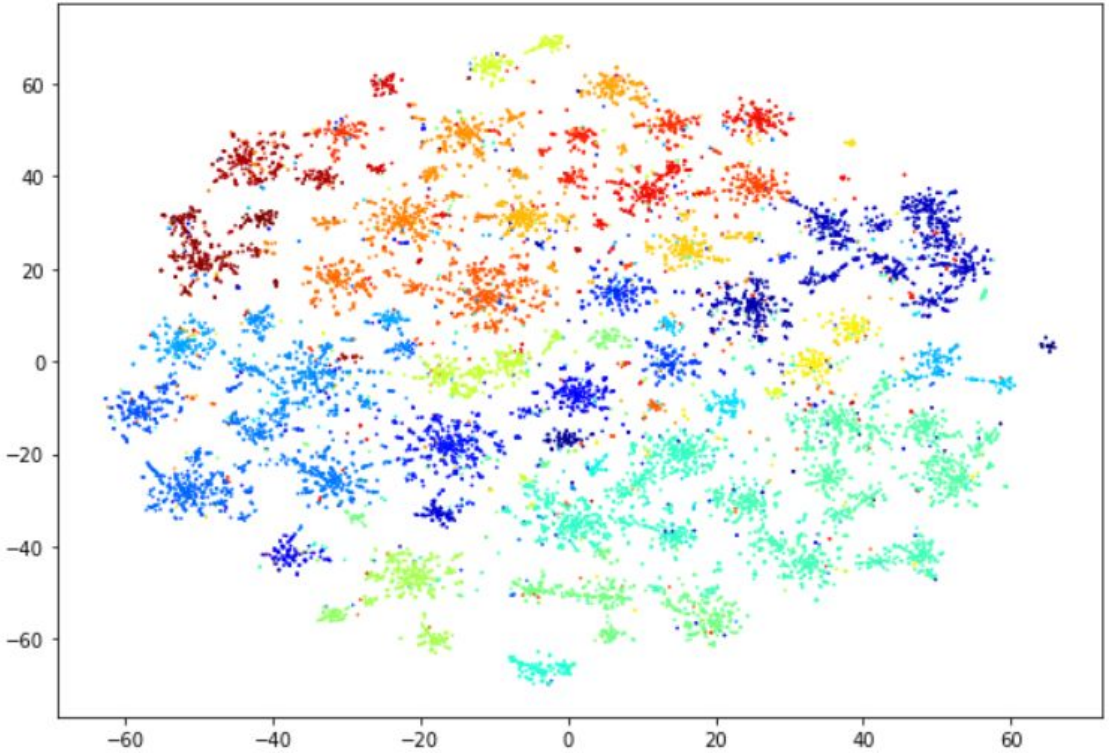


DBSCAN

W przypadku tego algorytmu, prawie wszystkie dokumenty były przyporządkowywane praktycznie do jednego klastra, więc ostatecznie nie był brany pod uwagę.

Hierarchical clustering

Wyniki tego algorytmu okazały się dla nas najlepsze.



Hierarchical clustering

Chmura słów pokazująca jakie słowa miały znaczenie w przykładowo wybranych klastrze.

Wyniki walidacji

Niestety grupa, która była odpowiedzialna za walidację naszego projektu, nie dostarczyła nam wyników.