Klasyfikacja wiadomości

Autorki: Aleksandra Samsel i Paulina Zapotoczna

Agenda

Etap 2

1. Przygotowanie do FE
2. Feature engineering
3. Standaryzacja + PCA
4. Elbow method & Silhouette score

Etap 3

1. Cel biznesowy
2. Dane
3. Wybór modeli
4. Wnioski

ETAP 2

Przygotowanie do feature engineering

Usunięcie stopwords

Usunięcie słów takich jak:
a, an, the, and, or, but, if, in, on, at, by, with, about, against, between, through, during, before, after, above, below, to, from, up, down, in, out, off, over, under, again, further, then, once, here, ...

Lematyzacja

Sprowadzanie odmienionych form wyrazów do ich podstawowej lub kanonicznej formy, np.
Cats -> Cat
Running -> Run
Went -> Go
Better -> Good

Usunięcie innych słów

Usunięcie innych słów, które naszym zdaniem nie miały wpływu na model

TF-IDF = TF x IDF

$$TF(t) = \frac{\text{Liczba wystąpień słowa t w dokumencie}}{\text{Łączna liczba słów w dokumencie}}$$

$$IDF(t) = \log\left(\frac{\text{Liczba wierszy}}{\text{Łączna wierszy zawierających słowo t}}\right)$$

year	time	new	look	struggle	home
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.157814	0.053022	0.086811	0.057373
0.026633	0.033223	0.000000	0.042141	0.000000	0.000000
0.090908	0.028351	0.000000	0.071922	0.058878	0.000000
0.000000	0.000000	0.000000	0.026361	0.000000	0.000000
0.000000	0.000000	0.029219	0.019634	0.064292	0.000000
0.082638	0.025772	0.048648	0.000000	0.000000	0.035372
0.040042	0.024976	0.070718	0.000000	0.051868	0.034279
0.069289	0.043217	0.000000	0.000000	0.000000	0.000000
0.028161	0.000000	0.016578	0.000000	0.000000	0.024107

Bardzo rzadka macierz - 5000 kolumn (większość wartości to 0)

Sentyment

Ocena sentymentu z użyciem narzędzia TextBlob. Sentyment to liczba z zakresu [-1.0, 1.0]

min(sentyment) = **-0.329**, max(sentyment) = **0.515**, mean(sentyment) = **0.081**

Best sentiment:
Bookmakers back Aviator for Oscar
The Aviator has been tipped by UK bookmakers as the favourite to win the best film award at this year's Oscars. Ray star Jamie Foxx is clear favourite in the best actor category while Million Dollar Baby's Hilary Swank is tipped to win the ...

Worst sentiment:
FA charges Liverpool and Millwall
Liverpool and Millwall have been charged by the Football Association over crowd trouble during their Carling Cup match on 26 October. Millwall, who lost the match 3-0, have also been charged over alleged racist behaviour by ...

Word Embeddings

Word2Vec

Technika uczenia maszynowego opracowana przez **Google**. Korzysta z dwóch głównych modeli:

- **Continuous Bag of Words** - przewiduje słowo na podstawie kontekstu (sąsiednich słów)
- **Skip-gram** - przewiduje kontekstowe słowa na podstawie danego rozpatrywanego słowa

Zdanie pierwsze: “**Ola jest blondynką**”
Zdanie drugie: “**Paulina ma rude włosy**”

Zdanie pierwsze: v1, v2, v3
Zdanie drugie: u1, u2, u3, u4

Zdanie pierwsze: v = mean(v1, v2, v3)
Zdanie drugie: u = mean(u1, u2, u3, u4)

Wektory odpowiadające dwóm słowom o podobnym znaczeniu są sobie bliskie w przestrzeni wektorowej

Word Embeddings

GloVe

Minimalizuje różnicę między iloczynem skalarów wektorów słów a logarytmem ich częstotliwości współwystępowania

Technika uczenia maszynowego opracowana przez **Stanford University**. Metoda ta łączy w sobie podejście statystyczne z techniką embeddingów. Metoda ta jest usprawnieniem Word2Vec.

Zdanie pierwsze: “**Ja lubię jeździć rowerem.**”

Zdanie drugie: “**Ja kocham jeździć autem**”

	ja	lubię	kocham	jeździć	rowerem	autem
ja	0	2	1	1	1	1
lubię	2	0	0	1	1	0
kocham	1	0	0	1	0	1
jeździć	1	1	1	0	1	1
rowerem	1	1	0	1	0	0
autem	1	0	1	1	0	0

Bigramy i Trigramy

Pojedyncze słowo nie zawsze niesie ze sobą “pełną” inforamcję, dlatego postanowiłyśmy wykorzystać również analizę bi- i trigramów.

Nasza tabela zawiera 1000 wierszy, każdy z nich reprezentuje bi- bądź trigram. Wartości to liczba występujących kawałków tekstu w każdym wierszu.

seven year	survey found	earlier year	three month	say party
0	0	0	0	0
0	0	0	0	0
0	0	1	1	0
2	1	0	1	0
0	0	0	0	0
0	2	0	0	0
0	0	0	0	0
0	0	0	0	2
0	0	1	0	0
1	0	0	0	1

Co dalej?

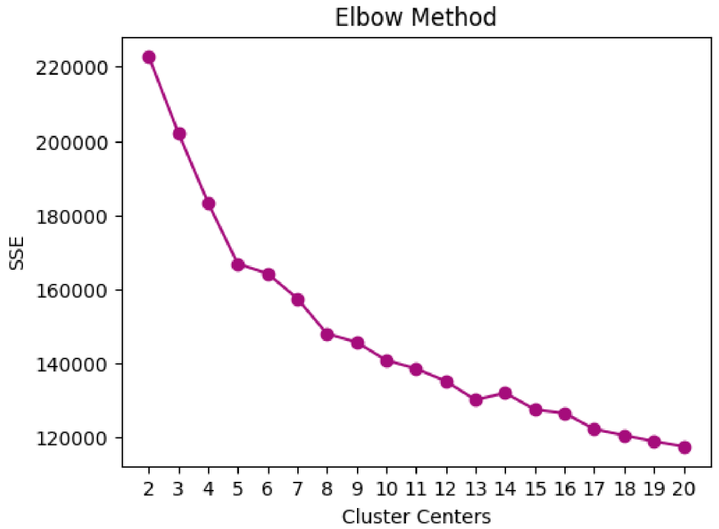
Tabele, które powstały:

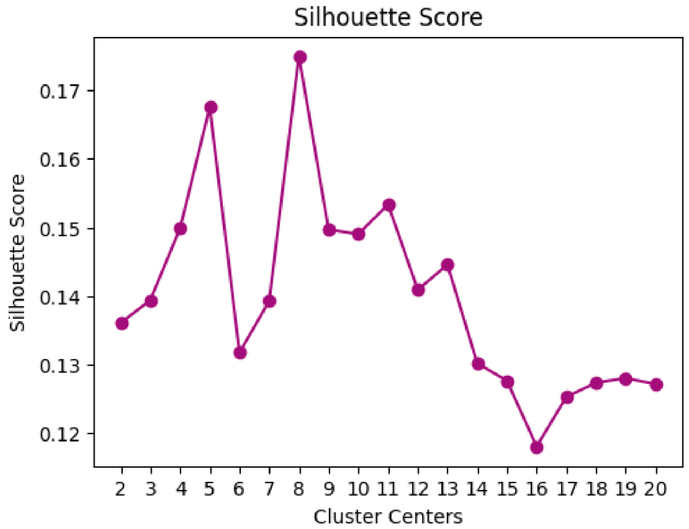
- td_idf + word2vec + glove + n-gram + sentyment
- td-idf
- word2vec
- glove
- n-gram

5 tabel, które posłużyły nam do stworzenia modeli

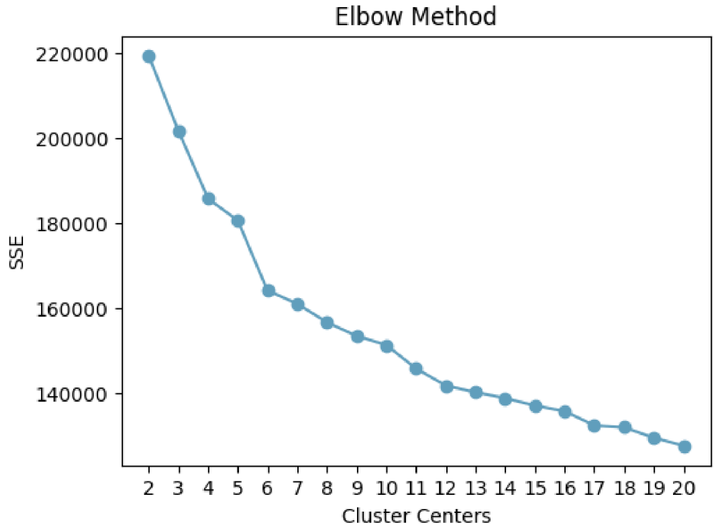
Elbow method & Silhouette score

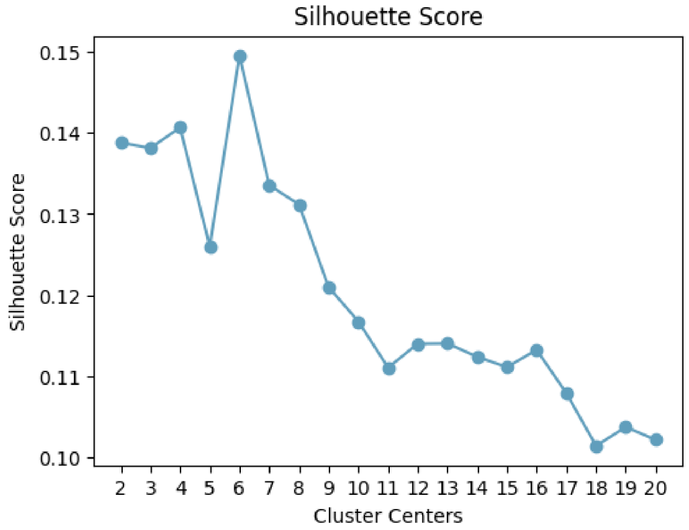
GloVe





Word2Vec





ETAP 3

Cel biznesowy

Stworzymy produkt, skierowany do redakcji czasopism, który z łatwością będzie klasteryzował dane na zbiory o tej samej tematyce

Dane

Blair blasts Tory spending plans.
Tony Blair has launched an attack on Conservative spending plans, saying they are a "ludicrous improbability". The prime minister has told a Labour Party gathering that the Tory policies would cause economic failure...

Greek pair attend drugs hearing.
Greek sprinters Kostas Kenteris and Katerina Thanou have appeared before an independent tribunal which will decide if their bans should stand. They were given provisional suspensions by athletics' ruling ...

Games enter the classroom.
Video games could soon be transplanted from their natural habitat to the more academic atmosphere of the classroom. With violent titles continuing to top the charts, gaming and learning ...

Da Vinci film to star Tom Hanks.
Actor Tom Hanks and director Ron Howard are reuniting for The Da Vinci Code, an adaptation of the international best-selling novel by Dan Brown. Distributor Sony Pictures said production ...

China keeps tight rein on credit.
China's efforts to stop the economy from overheating by clamping down on credit will continue into 2005, state media report. The curbs were introduced earlier this year to ward off the risk that rapid expansion ...

Przetestowane algorytmy

DBSCAN

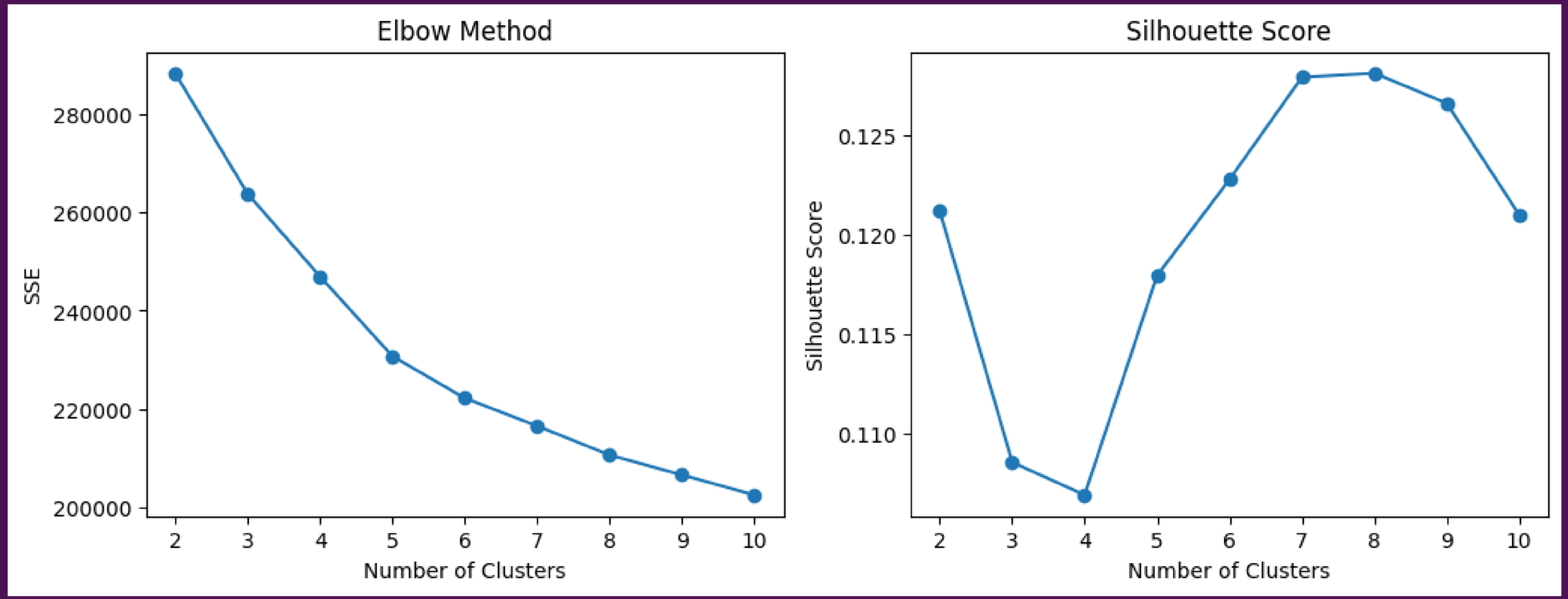
KMeans

Hierarchical Agglomerative Clustering

z różnymi sposobami łączenia danych w klastry

DIANA

Optymalizacja liczby klastrów



Najlepsza liczba: 6, 7 lub 8

Dane

Kategoria 0

Ailing EuroDisney vows turnaround
EuroDisney, the European home of Mickey Mouse and friends, has said it will sell 253m euros (£175m; $328m) of new shares as it looks to avoid insolvency.
The sale is the last part of a plan to restructure 2.4bn euros-worth of debts. Despite struggling since it was opened...

Kategoria 1

Double eviction from Big Brother
Model Caprice and Holby City actor Jeremy Edwards have both left the Celebrity Big Brother house in a surprise double eviction on Friday.
Caprice, who left in the scheduled fourth eviction having gained just 5% of the public vote, afterwards said...

Kategoria 2

'Jamelia's return to the top
R&B star Jamelia had three Brit nominations to go with her triple triumph at last year's Mobo awards.
The Birmingham-born singer, full name Jamelia Davis, was signed to a record label at the age of 15 and released her first single So High at 18. She released four number ones from her 2000 album Drama...

Kategoria 3

Musicians 'upbeat' about the net
Musicians are embracing the internet as a way of reaching new fans and selling more music, a survey has found.
The study by US researchers, Pew Internet, suggests musicians do not agree with the tactics adopted by the music industry against file-sharing. While most considered file-sharing as illegal...

Kategoria 4

Butler strikes gold in Spain
Britain's Kathy Butler continued her impressive year with victory in Sunday's 25th Cross Internacional de Venta de Banos in Spain.
The Scot, who led GB to World Cross Country bronze earlier this year, moved away from the field with Ines...

Kategoria 5

Mexicans tracking unhappy Juninho
Mexican outfit Red Sharks Veracruz hope to sign Juninho if the Brazilian decides to leave Celtic frustrated at his lack of first-team action.
Their president, Gustavo Parente Sanchez, says Juninho "does not wish to remain in Scottish football anymore"...

Kategoria 6

Federer breezes into semi-finals
Roger Federer reached the last four of the Qatar Open with an easy 6-1 6-2 win over seventh seed Feliciano Lopez.
The Swiss world number one reeled off a series winners to outclass the Spaniard and set up a semi-final match...

Dane

Kategoria 0	Business and Market Trends
Kategoria 1	Politics and Entertainment
Kategoria 2	Film
Kategoria 3	Technology and Music
Kategoria 4	Sports
Kategoria 5	Government and Media
Kategoria 6	Business

Dziękujemy za uwagę!