



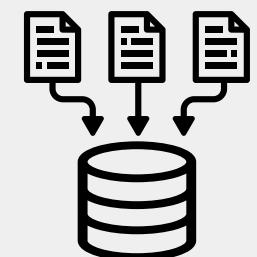
# **OPINEO REVIEWS CLASSIFIER**

**MICHAŁ BUDZYŃSKI**

# AGENDA

- 
- 
- 01** OPIS PROBLEMU
  - 02** CEL ANALIZY I ZASTOSOWANE NARZĘDZIA
  - 03** WSTĘPNA EKSPLORACJA DANYCH
  - 04** PRZYGOTOWANIE DANYCH DO MODELOWANIA
  - 05** MODELE KLASYFIKACJI TEMATYCZNEJ (LDA, LSI, HDP)
  - 06** MOCNE I SŁABE STRONY MODELU
  - 07** POTENCJALNE KIERUNKI ROZWOJU
  - 08** PODSUMOWANIE I WNIOSKI

# OPIS PROBLEMU

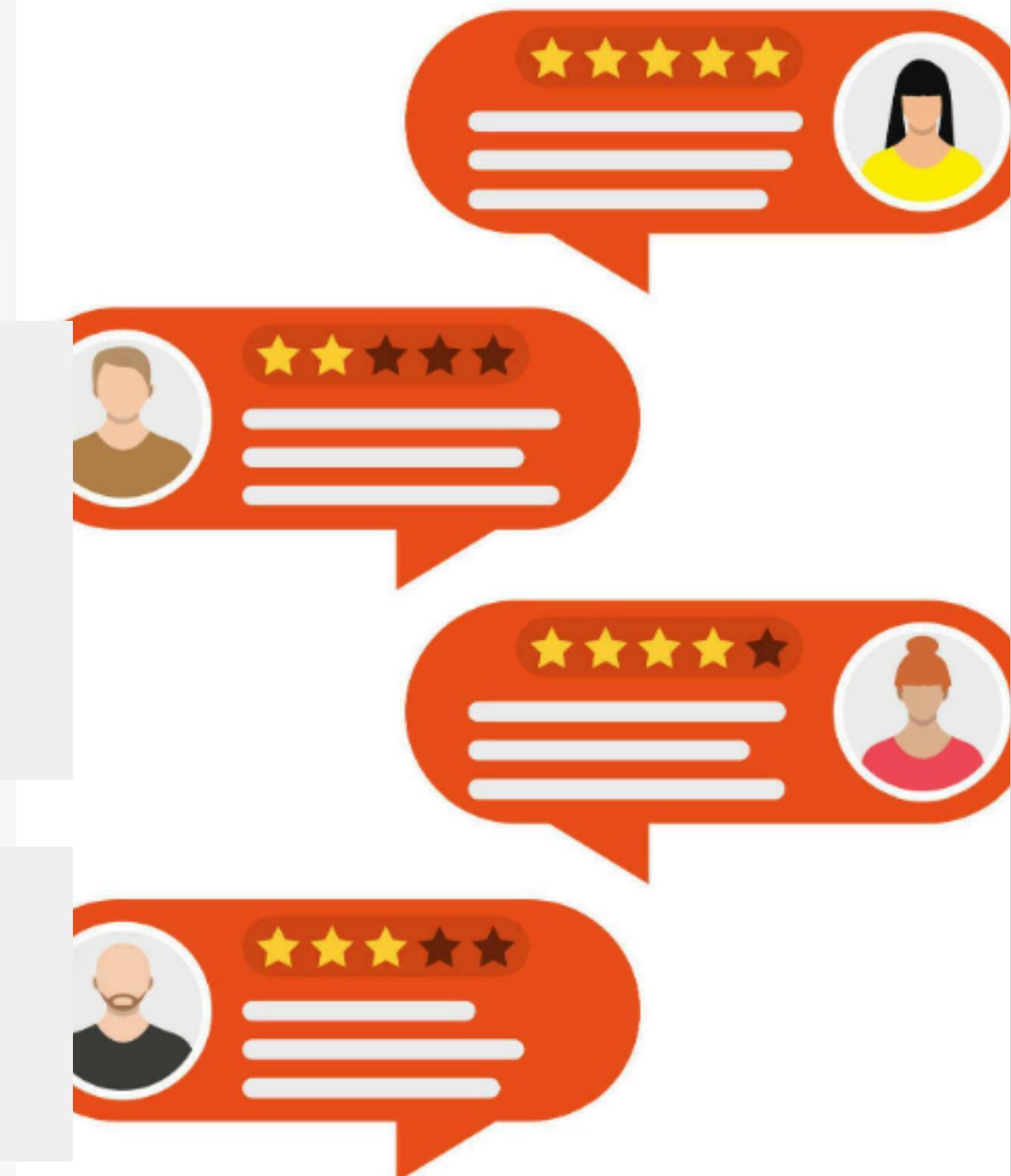


Baza danych zawiera informacje o treści recenzji wydanych przez użytkowników oraz ich ocenie firmy, gdzie:

- ocena 1 – oznacza ocenę pozytywną
- ocena 0 – oznacza ocenę neutralną
- ocena -1 – oznacza ocenę negatywną



Opinie i oceny użytkowników pochodzą z serwisów Opineo, Twitter oraz YouTube. Opinie wydane są w języku polskim.



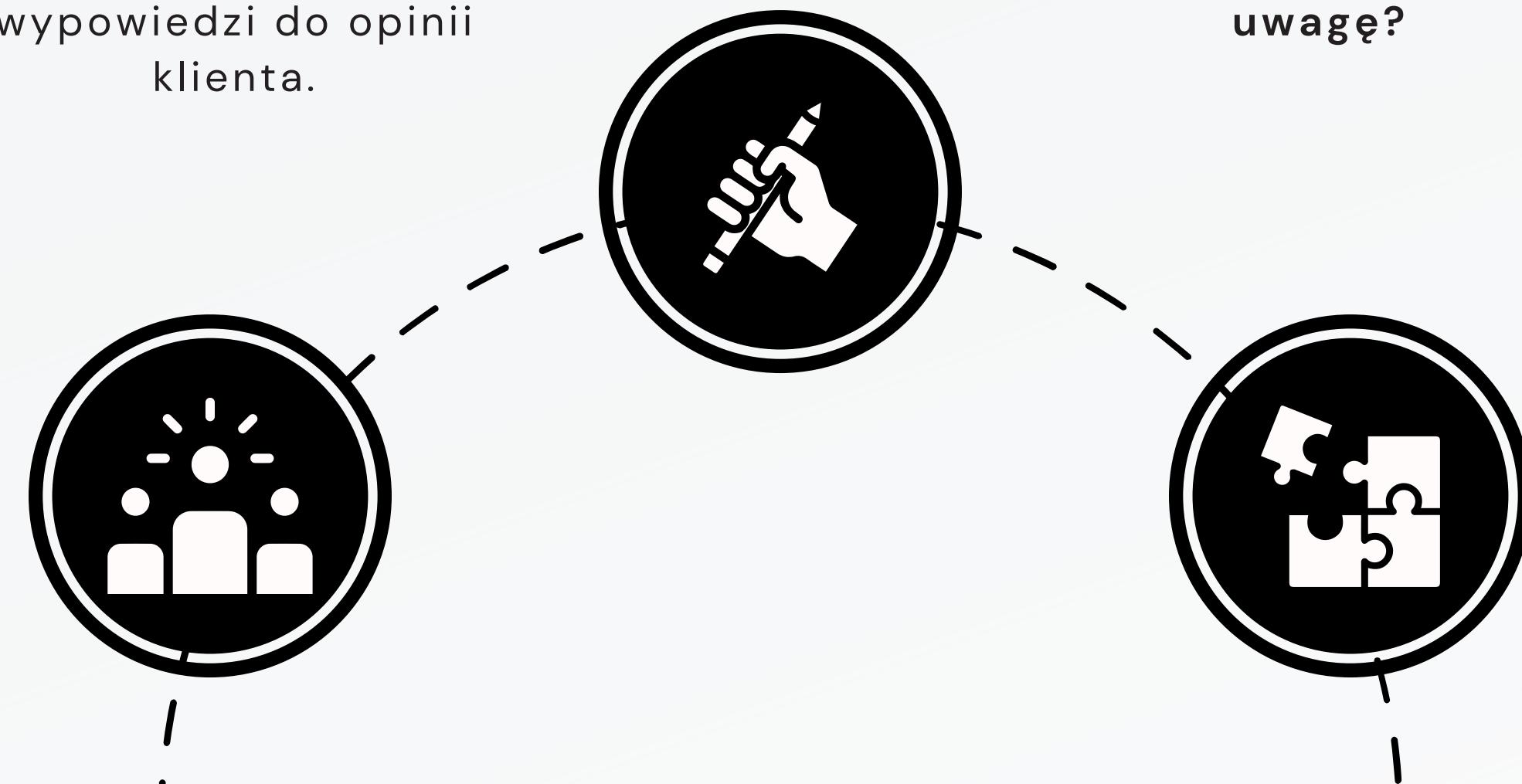
# CEL ANALIZY I ZASTOSOWANE NARZĘDZIA

## Cel 1

Budowa modelu klasyfikacyjnego, który będzie przypisywał temat wypowiedzi do opinii klienta.

## Cel 2

Wyodrębnienie elementów istotnych z perspektywy klienta.  
**Na co klienci zwracają uwagę?**



# CEL ANALIZY I ZASTOSOWANE NARZĘDZIA



- podstawowa analiza statystyczna
- analiza "n-gramów"
- tokenizacja
- usuwanie "stopwords"
- stemming, lemmatization
- bag of words

WSTĘPNA ANALIZA I  
PRZYGOTOWANIE  
DANYCH



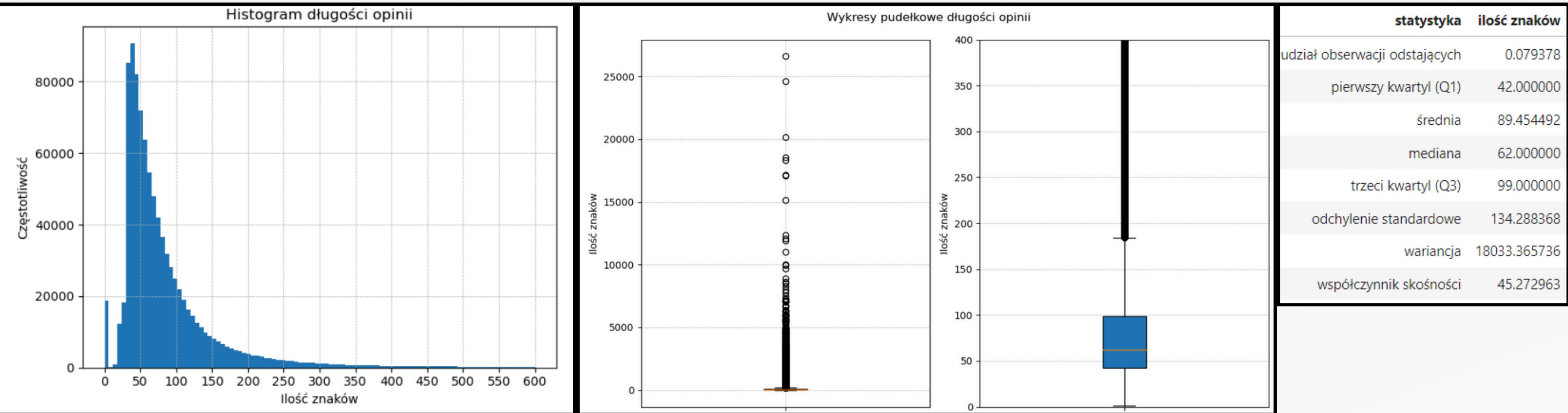
- Latent Dirichlet Allocation (**LDA**)
- Hierarchical Dirichlet Process (**HDP**)
- Latent Semantic Indexing (**LSI**)
- współczynnik spójności (**coherence score**)

MODELOWANIE  
TEMATYCZNE

# WSTĘPNA EKSPLORACJA DANYCH

Struktura opinii		
Charakter opinii	Ilość opinii	Udział
pozytywna	734 250	78%
neutralna	18 547	2%
negatywna	183 391	20%
<b>SUMA</b>	<b>936 188</b>	

# WSTĘPNA EKSPLORACJA DANYCH

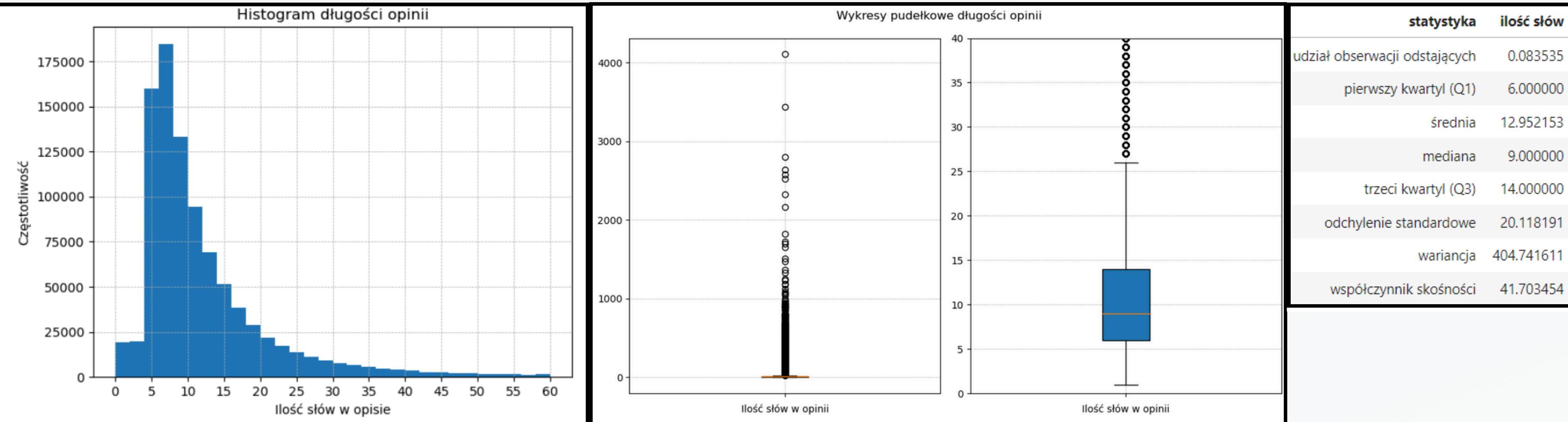


## Wnioski:

Widoczna asymetria prawostronna, co wskazuje na tendencje klientów do formowania relatywnie krótkich wypowiedzi, po przekroczeniu 50 znaków w wypowiedzi obowiązuje reguła: **im dłuższa wypowiedź tym rzadziej ona występuje.**

Z kolei wysoka wariancja informuje o **dużym zróżnicowaniu pod względem długości wypowiedzi**. Trzeba pamiętać jednak, że długość wypowiedzi wyrażono w znakach, zatem w rzeczywistości różnice te mogą nie być tak widoczne.

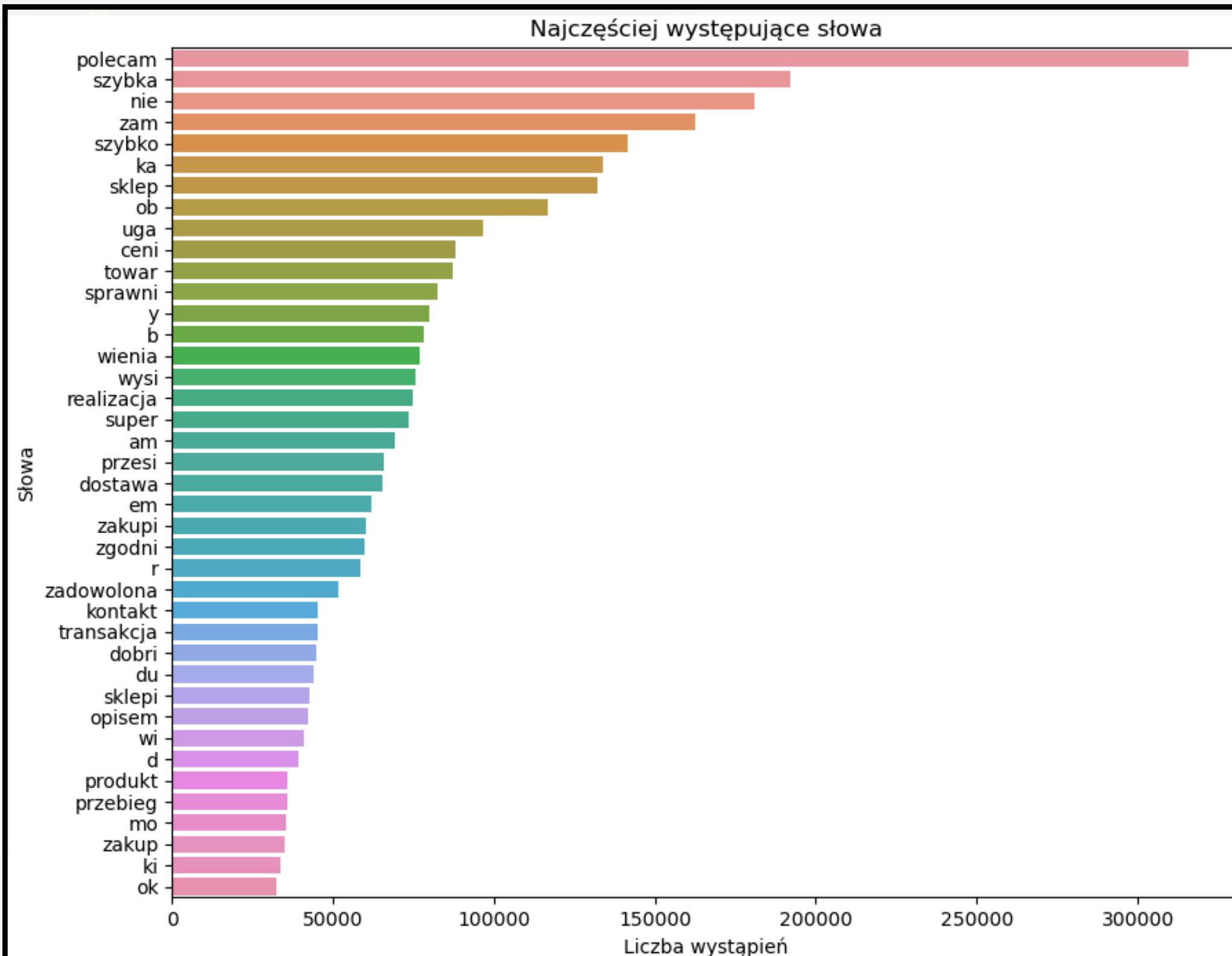
# WSTĘPNA EKSPLORACJA DANYCH



## Wnioski:

Dane zagregowane do słów są znacznie bardziej interpretowalne. Potwierdza się prawostronna asymetria rozkładu długości opinii – **najwięcej konsumentów używa od 5 do 10 słów** w swojej opinii, potem z każdym kolejnym słowem ilość opinii zmniejsza się. **Połowa klientów używa od 6 do 14 słów** w swojej wypowiedzi. Mimo, że wykresy pudełkowe wskazują na dużą ilość obserwacji odstających to ich udział w całej populacji jest relatywnie niewielki – nie przekracza 10%.

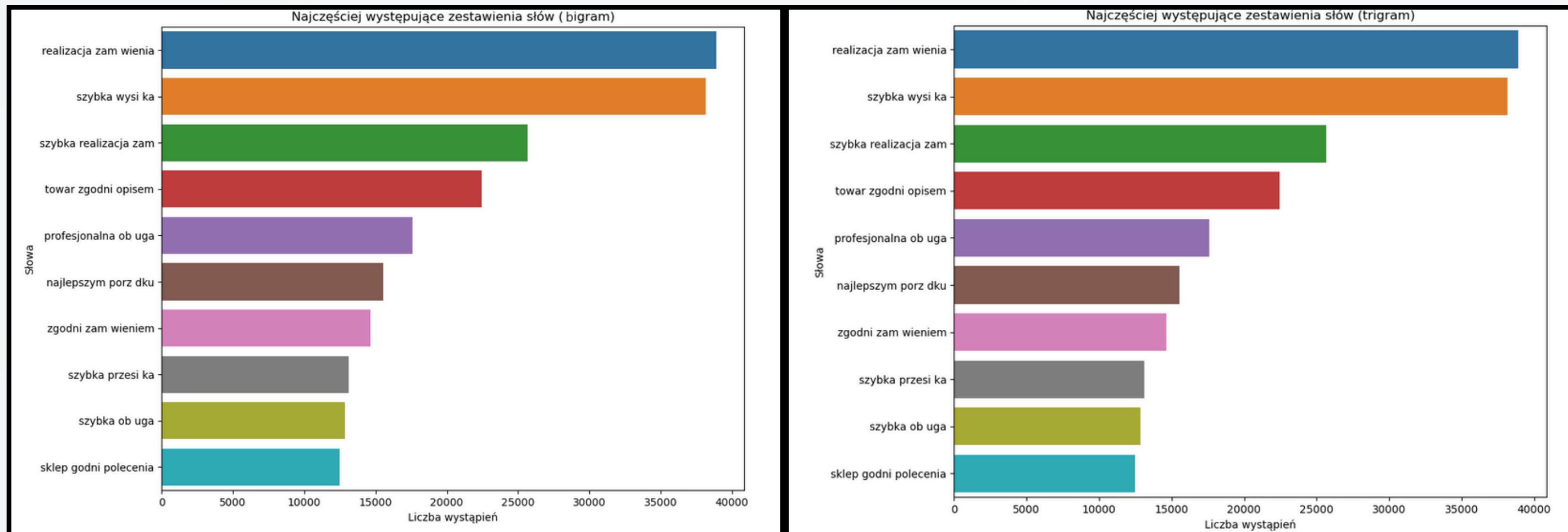
# WSTĘPNA EKSPLORACJA DANYCH



## Wnioski:

- Analiza została przeprowadzona po wykonaniu procesu stemmingu, zatem niektóre słowa naturalnie mogą nie być zrozumiałe.
- Zdecydowanie **przeważają słowa o pozytywnym znaczeniu**: polecam, szybko/szybka, sprawni, super etc.
- Wydaje się, że dla klientów najistotniejszymi kwestiami są
  - szbkość realizacji dostawy**
  - kontakt z obsługą**
  - oferowany asortyment**
  - kwestie związane z transakcją**

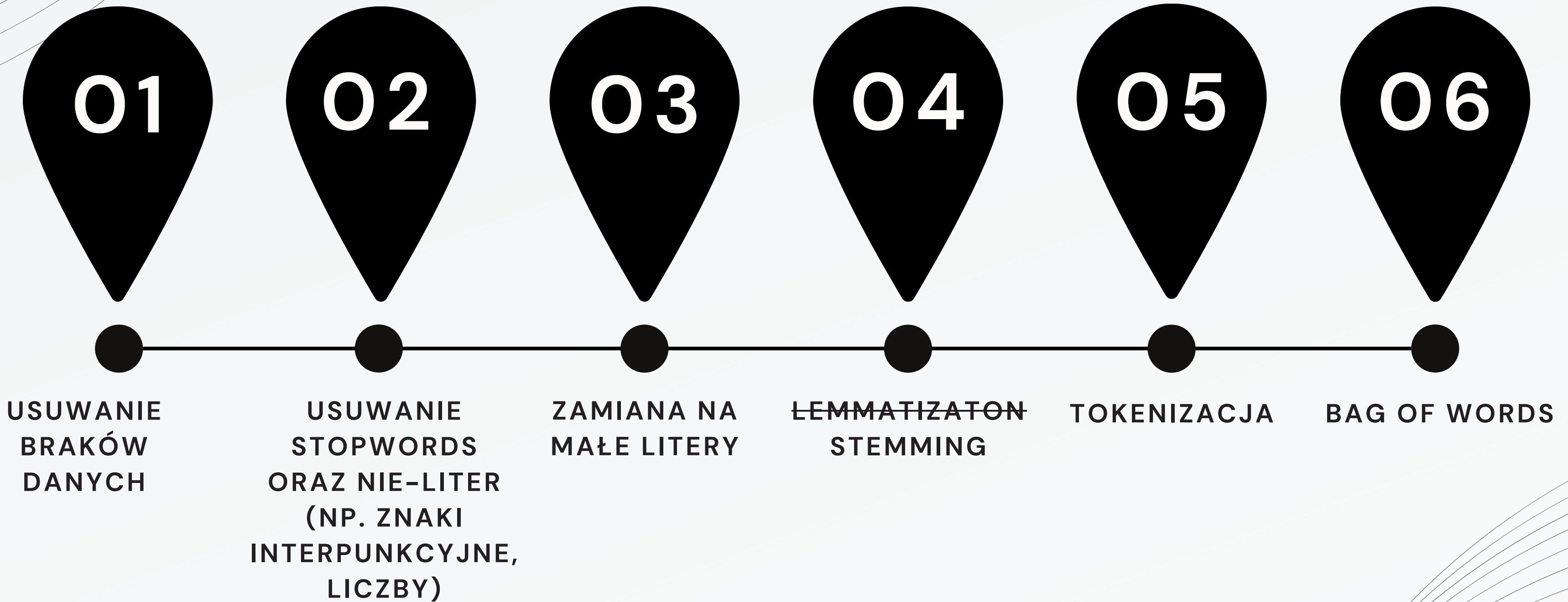
# WSTĘPNA EKSPLORACJA DANYCH



## Wnioski:

Wykresy bigramów oraz trigramów potwierdzają wnioski wysnute na poprzednim slajdzie i je uszczegóławiają. W kontekście szybkości realizacji dostawy klienci szczególnie cenią sobie **szylkę i szybką realizację zamówienia**. W kontekście obsługi klienci wskazują na **szylkość i profesjonalizm obsługi**. Natomiast w odniesieniu do samego asortymentu, klienci dostrzegają **zgodność zamówionego towaru z jego opisem**. Ponownie, **brak wyrażeń o negatywnym charakterze**.

# PRZYGOTOWANIE DANYCH DO MODELOWANIA

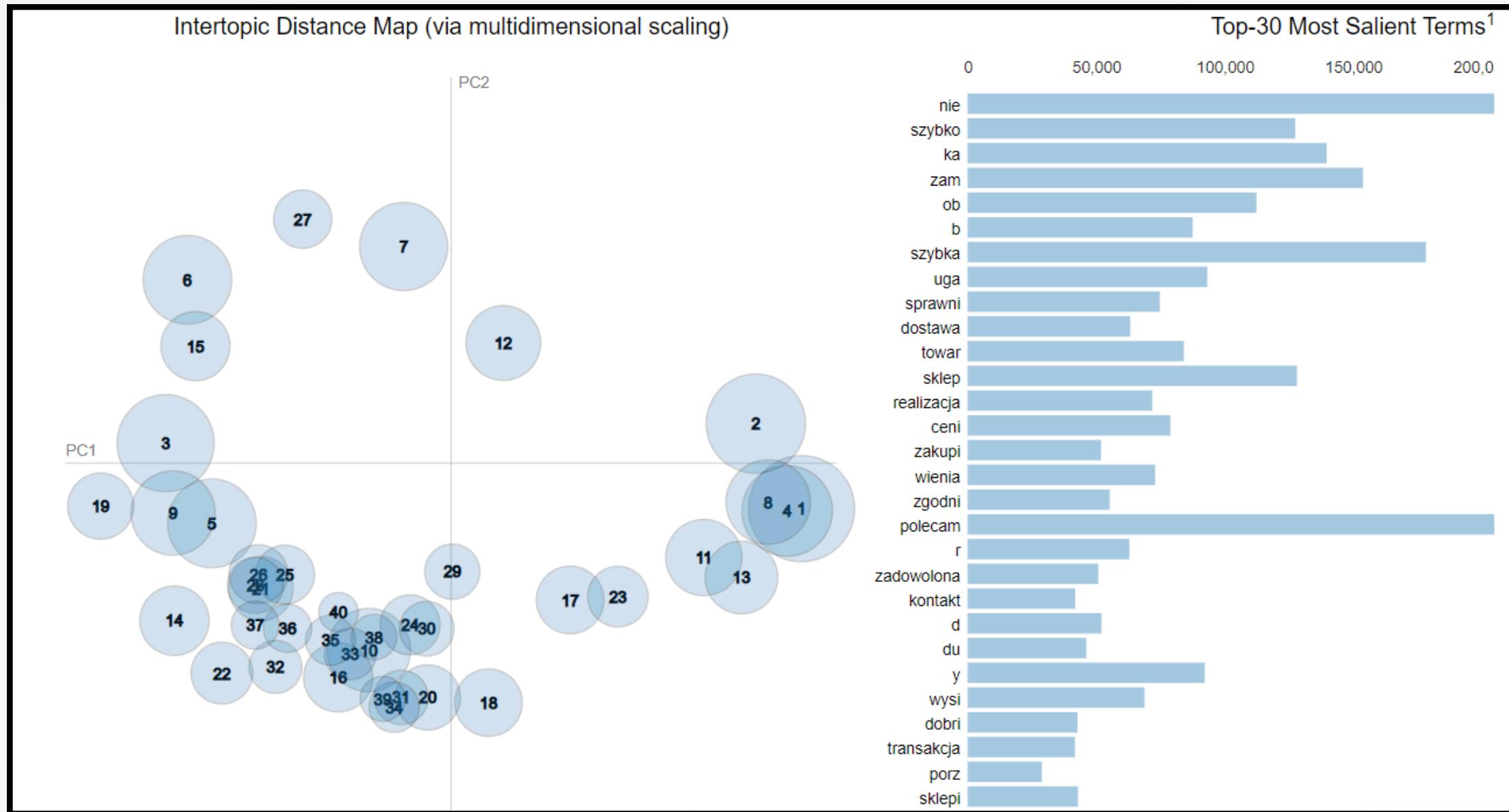


# **MODELE KLASYFIKACJI TEMATYCZNEJ**

- Latent Dirichlet Allocation
- Latent Semantic Indexing
- Hierarchical Dirichlet Process

# MODELE KLASYFIKACJI TEMATYCZNEJ

## MODEL LDA



Początkowo oszacowano model LDA dla 40 tematów z uwagi na duży zbiór danych. Powyższy wykres ukazuje graficzną reprezentację wyników modelu. Z wykresu można odczytać, że występuje wiele tematów, które są relatywnie nieistotne i nieróżnicowane między sobą. Dodatkowo, wyliczono miarę koherencji (coherence score), która również przyjęła relatywnie niską wartość 0.49 co potwierdza konieczność redukcji liczby tematów.

# KOD PYTHON

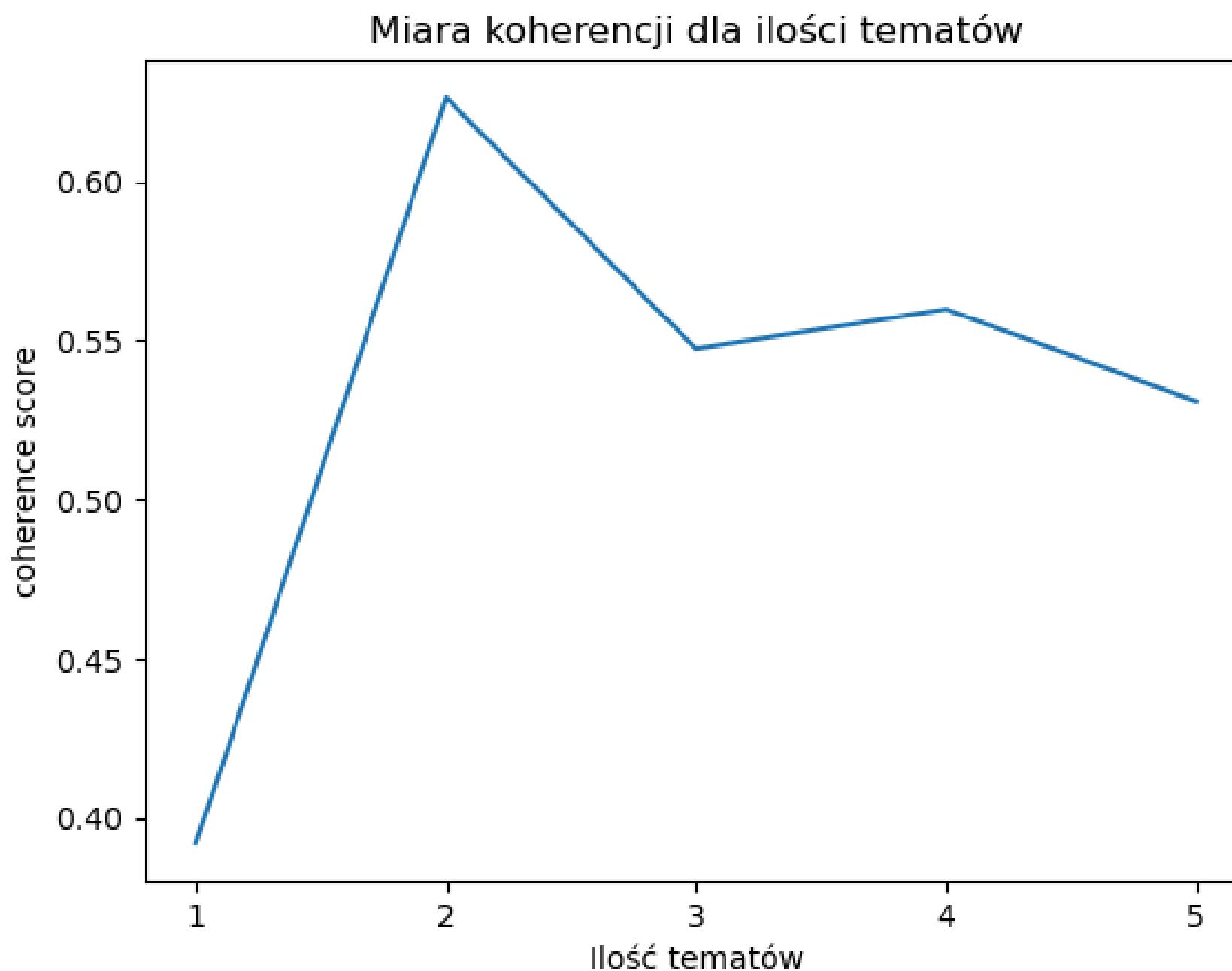
```
# LDA model training
lda_model = gensim.models.LdaMulticore(bow_corpus,
                                         num_topics = 40,
                                         id2word = dic,
                                         random_state=42,
                                         chunksize = 20000,
                                         passes = 3,
                                         workers = 2)

# model visualization with pyLDAvis
pyLDAvis.enable_notebook()
vis = pyLDAvis.gensim_models.prepare(lda_model, bow_corpus, dic)
vis

# calculating model coherence score
coherence_model_lda = CoherenceModel(model=lda_model,
                                       texts=tokenized_corpus, dictionary=dic, coherence='c_v')
coherence_score_lda = coherence_model_lda.get_coherence()
print(f"Coherence Score: {coherence_score_lda}")
```

# MODELE KLASYFIKACJI TEMATYCZNEJ

MODEL LDA - FINE-TUNING



## Sposób analizy:

W celu redukcji ilości tematów trenowano modele LDA zmieniając każdorazowo ilość tematów w modelu. Na tej podstawie wyliczono miary koherencji dla danej ilości tematów, które zaprezentowane są na wykresie obok. Warto nadmienić, że z racji **długiego procesu komplikowania** kodu służącego do trenowania modeli LDA, oszacowano **jedynie kilka pierwszych ilości tematów**. W celu kompleksowej analizy powinno rozszerzyć się ten zakres.

## Wnioski:

Do dalszej analizy postanowiono wybrać **model z 2 oraz 4 tematami**. Mimo, że model z 2 tematami posiada najwyższy współczynnik koherencji to z biznesowego punktu widzenia podejrzewa się, że może on być nieużyteczny.

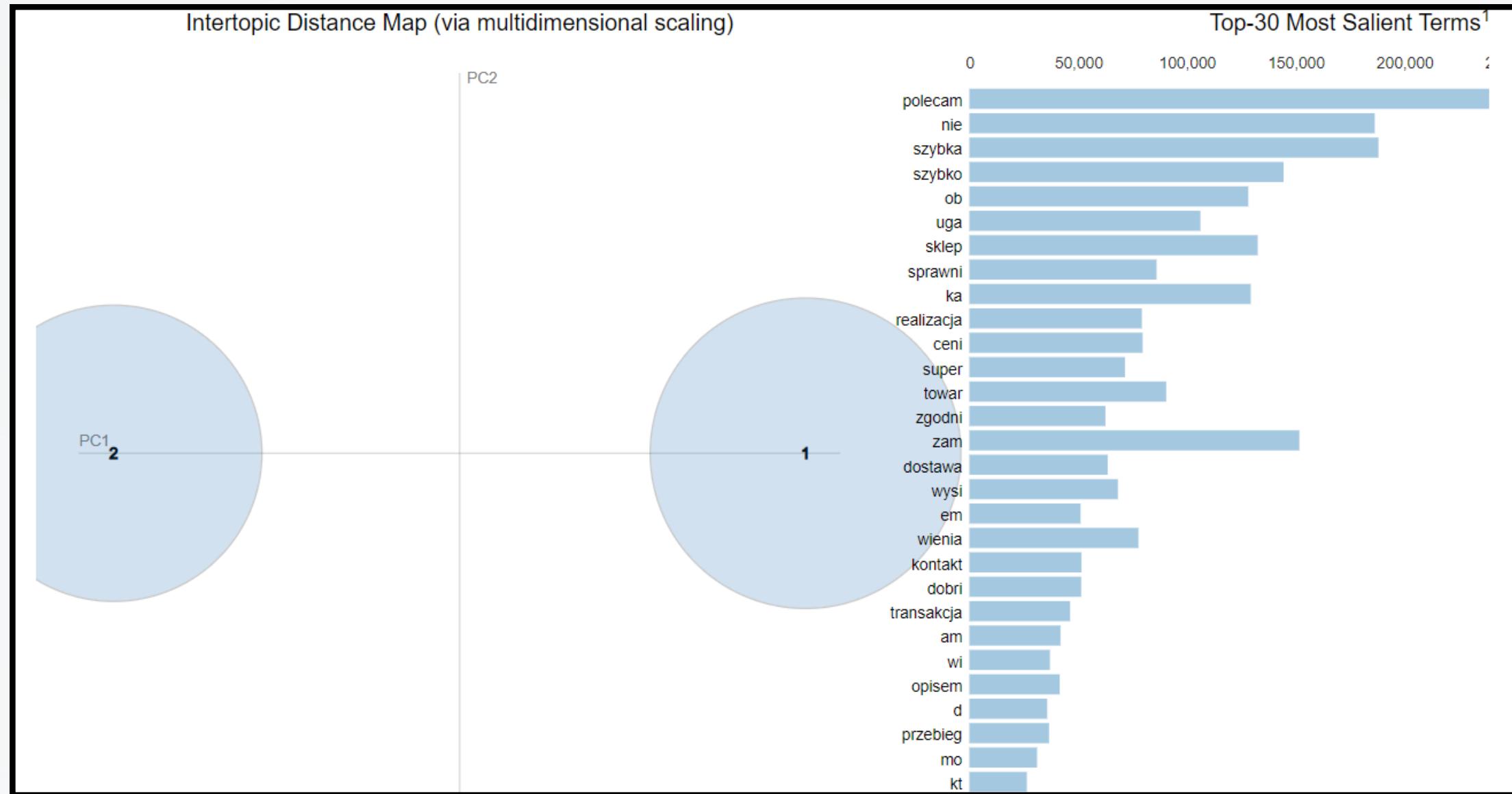
# KOD PYTHON

```
# calculating coherence score for LDA models with number of topics from 1 to 5
number_of_topics = []
coherence = []
for x in tqdm(range(1,6)):
    number_of_topics.append(x)
    lda_model = gensim.models.LdaMulticore(bow_corpus,
                                             num_topics = x,
                                             id2word = dic,
                                             random_state=42,
                                             chunksize = 20000,
                                             passes = 3,
                                             workers = 2)
    coherence_model_lda = CoherenceModel(model=lda_model,
                                           texts=tokenized_corpus, dictionary=dic, coherence='c_v')
    coherence_score_lda = coherence_model_lda.get_coherence()
    coherence.append(coherence_score_lda)

# plotting
plt.plot(number_of_topics,coherence)
plt.xlabel('Ilość tematów')
plt.ylabel('coherence score')
plt.title('Miara koherencji dla ilości tematów')
plt.xticks([1,2,3,4,5])
plt.show()
```

# MODELE KLASYFIKACJI TEMATYCZNEJ

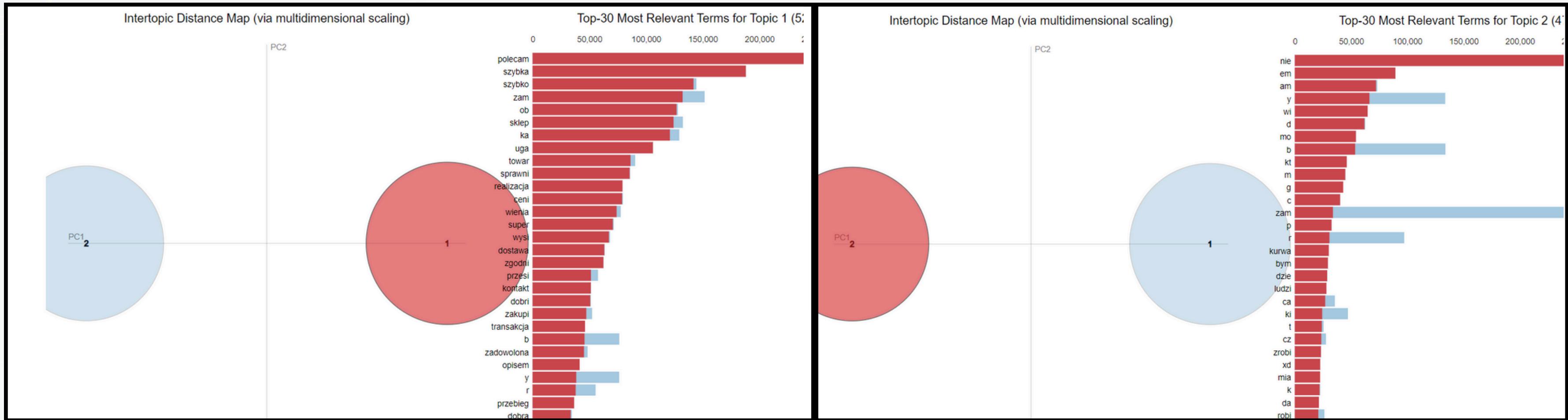
## MODEL LDA - 2 TEMaty



Oba tematy są istotne i zróżnicowane między sobą. Jednak, jak zostało wcześnie wspomniane, wątpliwość budzi bardzo ograniczone ilość tematów.

# MODELE KLASYFIKACJI TEMATYCZNEJ

## MODEL LDA - 2 TEMaty



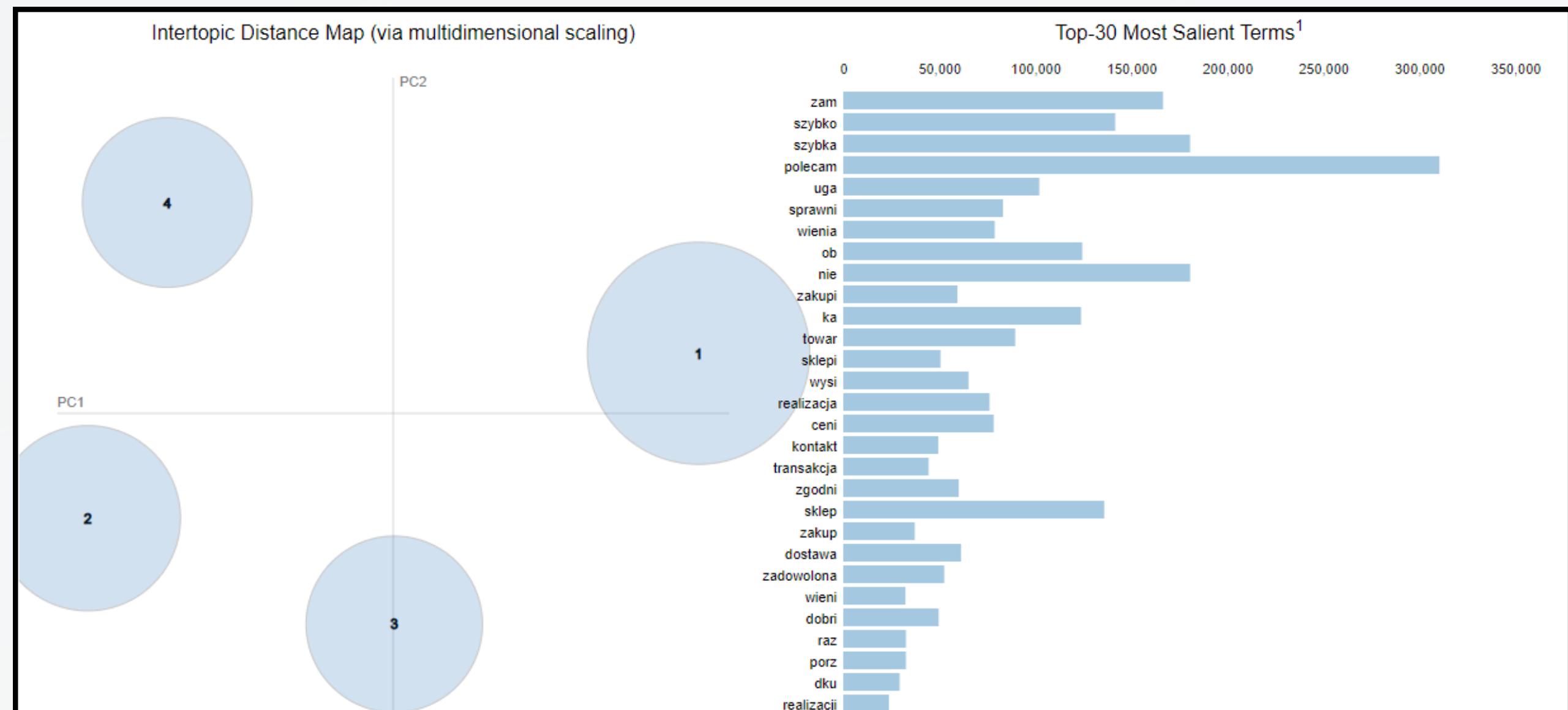
Po prawej stronie wykresów kolorem czerwonym oznaczone zostały histogramy dla wybranego tematu. Natomiast kolorem niebieskim - histogramy alternatywnego tematu.

**Tematy są odpowiednio odseparowane i tylko nieznacznie się zazębają.** Jednak **temat 2** okazuje się być zbiorem elementów roboczo nazwanych jako "śmietnikowe". Znajdują się w nim słowa, która po procesie stemmingu wydają się tracić swoją interpretowalność (nie można odnaleźć ich dokładnego znaczenia). Z kolei temat 1 zawiera pozostałe wyrazy, których znaczenie co do zasady można łatwo "odgadnąć".

Taki wynik uznano za niesatyfakcyjający, zatem w dalszej części przystąpiono do analizy modelu LDA z 4 tematami.

# MODELE KLASYFIKACJI TEMATYCZNEJ

## MODEL LDA - 4 TEMaty



Wszystkie wyodrębnione tematy są istotne i zróżnicowane między sobą. Model posiada również drugi najwyższy współczynnik koherencji spośród analizowanych modeli LDA. Dodatkowo, analiza składu poszczególnych tematów nie wykazała żadnych nieprawidłowości. Finalnie spośród wszystkich modeli klasy LDA wybrano model LDA z 4 tematami.

# MODELE KLASYFIKACJI TEMATYCZNEJ

## PORÓWNANIE MODELI

### Porównanie modeli klasyfikacyjnych

Model	Miara koherencji (coherence score)
LDA (4 tematy)	0.56
LSI (4 tematy)	0.62
HDP	0.39

#### Wnioski:

Model LSI wykazuje nieco wyższy współczynnik koherencji niż model LDA. Najniższy współczynnik koherencji okazuje się osiągać model HDP. Finalnie do wykonywania dalszych predykcji wykorzystano **model LSI z 4 tematami**.

\*Model LSI tworzono w oparciu o 4 tematy. Ponownie z uwagi na wydłużony czas komplikacji tym razem nie zdecydowano się na oddzielny dobór ilości tematów dla modelu LSI – zastosowano optymalną liczbę tematów dla modelu LDA.

# KOD PYTHON

```
# LSI model training
lsi_model = LsiModel(corpus=bow_corpus, id2word=dic, num_topics=4)
# model performance
coherence_model_lsi = CoherenceModel(model=lsi_model,
texts=tokenized_corpus, dictionary=dic, coherence='c_v')
coherence_score_lsi = coherence_model_lsi.get_coherence()
print(f"Coherence Score: {coherence_score_lsi}")

# HDP model training
hdp_model = HdpModel(corpus=bow_corpus, id2word=dic)
# model performance
coherence_model_hdp = CoherenceModel(model=hdp_model,
texts=tokenized_corpus, dictionary=dic, coherence='c_v')
coherence_score_hdp = coherence_model_hdp.get_coherence()
print(f"Coherence Score: {coherence_score_hdp}")
```

# MOCNE I SŁABE STRONY MODELU

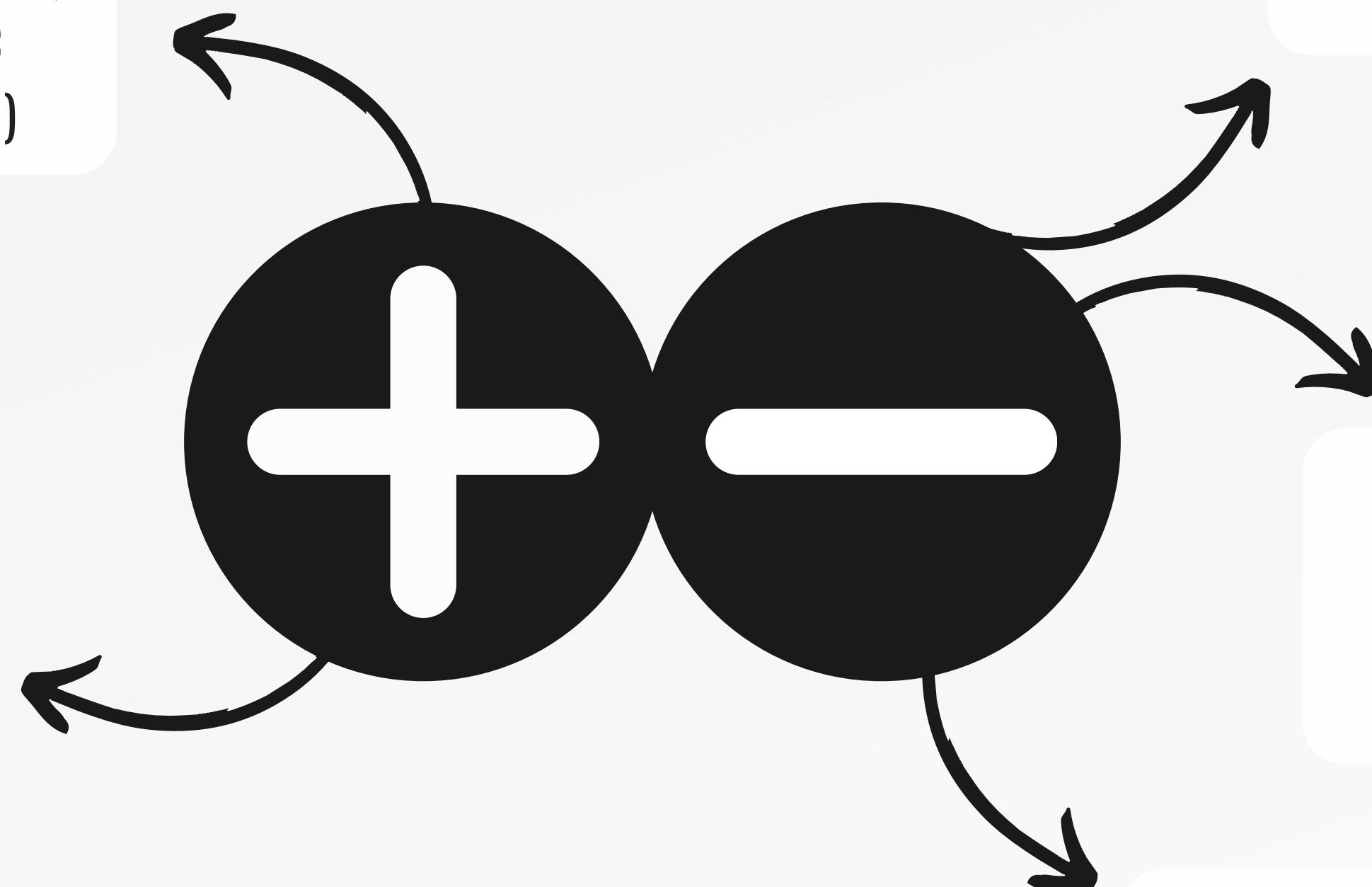
DOBRA JAKOŚĆ MODELU,  
WYSOKA SPÓJNOŚĆ  
(COHERENCE SCORE)

RELATYWNIĘ SZYBKA  
IMPLEMENTACJA DO  
NOWYCH DANYCH

PROBLEM GŁOSNYCH  
SŁÓW (OVERFITTING TO  
HIGH-FREQUENCY  
WORDS)

NISKA  
INTERPRETOWALNOŚĆ  
ZBIORU SŁÓW W  
TEMACIE

STATYCZNOŚĆ MODELU



# POTENCJALNE KIERUNKI ROZWOJU

1

LEMMATIZATION ZAMIAST  
STEMMING'U

2

INDYWIDUALNE DOPASOWANIE  
ILOŚCI TEMATÓW DLA MODELU

3

ZWIĘKSZENIE UDZIAŁU OPINII  
NEUTRALNYCH ORAZ  
NEGatywnych PRóBIE DO  
BUDOWY MODELI

4

MINIMALIZACJA PROBLEMU  
OVERFITTINGU MODELU DO SŁÓW  
GŁOŚNYCH

DODATKOWE USUNIĘCIE  
STOPWORDS

WAŻENIE CZĘSTOŚCIĄ  
TERMÓW (TF-IDF)

# PODSUMOWANIE I WNIOSKI

- Cel projektu został spełniony
- Wykazano, że konsumenti najbardziej cenią sobie w działaniu firmy elementy takie jak:
  - szybka realizacja zamówienia
  - profesjonalizm i szybkość obsługi
  - zgodność zamówionego towaru z jego opisem
- W wyniku analizy nie wyodrębniono elementów działania firmy, które należałyby poprawić.
- W toku analizy zbudowano łącznie 8 modeli służących do klasyfikacji tematycznej (6 modeli LDA, 1 model LSI, 1 model HDP), spośród których najbardziej efektywny okazał się model LSI
- Oszacowania modelu okazały się być nieinterpretowalne, prawdopodobnie z powodu problemu overfittingu modelu do słów głośnych oraz restrykcyjnego stemmingu



**DZIĘKUJE ZA  
UWAGĘ**

Michał Budzyński

