Understanding Topic Modeling with LDA and Visualizing Results with pyLDAvis

A Practical Guide to Text Analysis

Presenter: [hyunmin han Name]

Date: [24.11.25]

Journal of the Korea Academia-Industrial cooperation Society
Vol. 25, No. 10 pp. 728-736, 2024

https://doi.org/10.5762/KAIS.2024.25.10.728
ISSN 1975-4701 / eISSN 2288-4688

LDA 토픽 모델링 분석을 통한 챗GPT 활용에 대한 탐색적 연구
-챗GPT와 교육을 중심으로-

강태구
건양사이버대학교 디지털마케팅학과

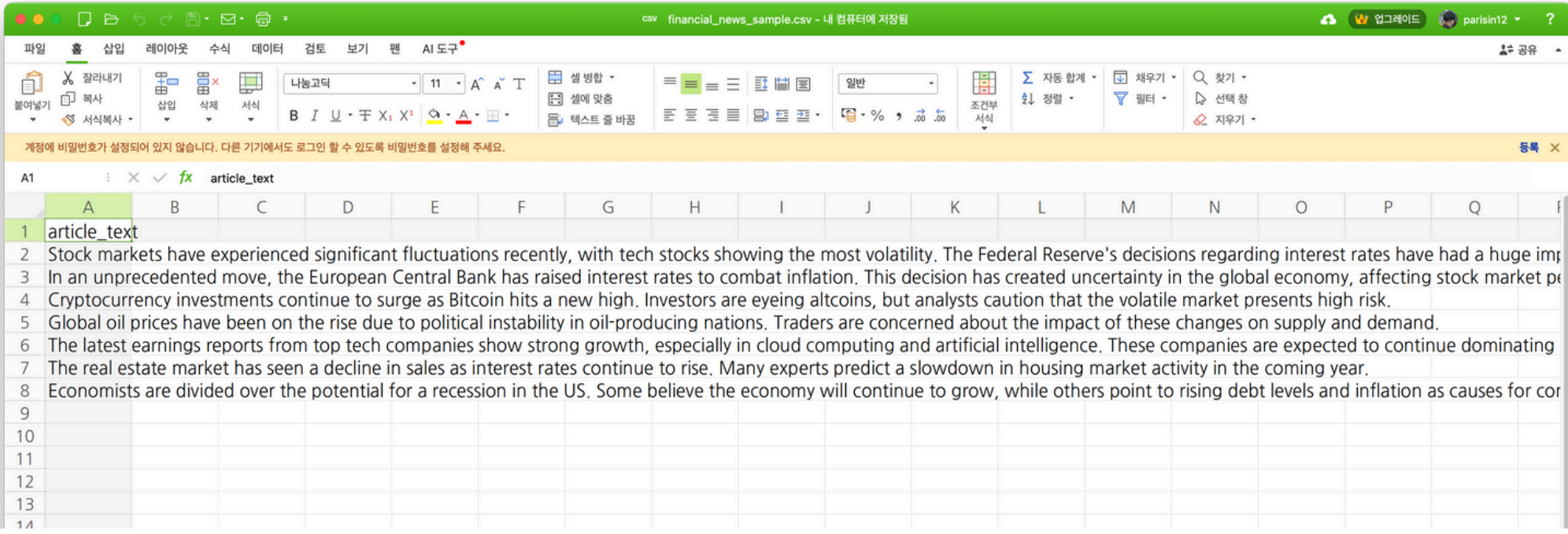
An Exploratory Study on the Utilization of ChatGPT through LDA Topic Modeling Analysis
-Focusing on ChatGPT and education-

Tae-Gu Kang
Dept. of Digital Marketing, Konyang Cyber University

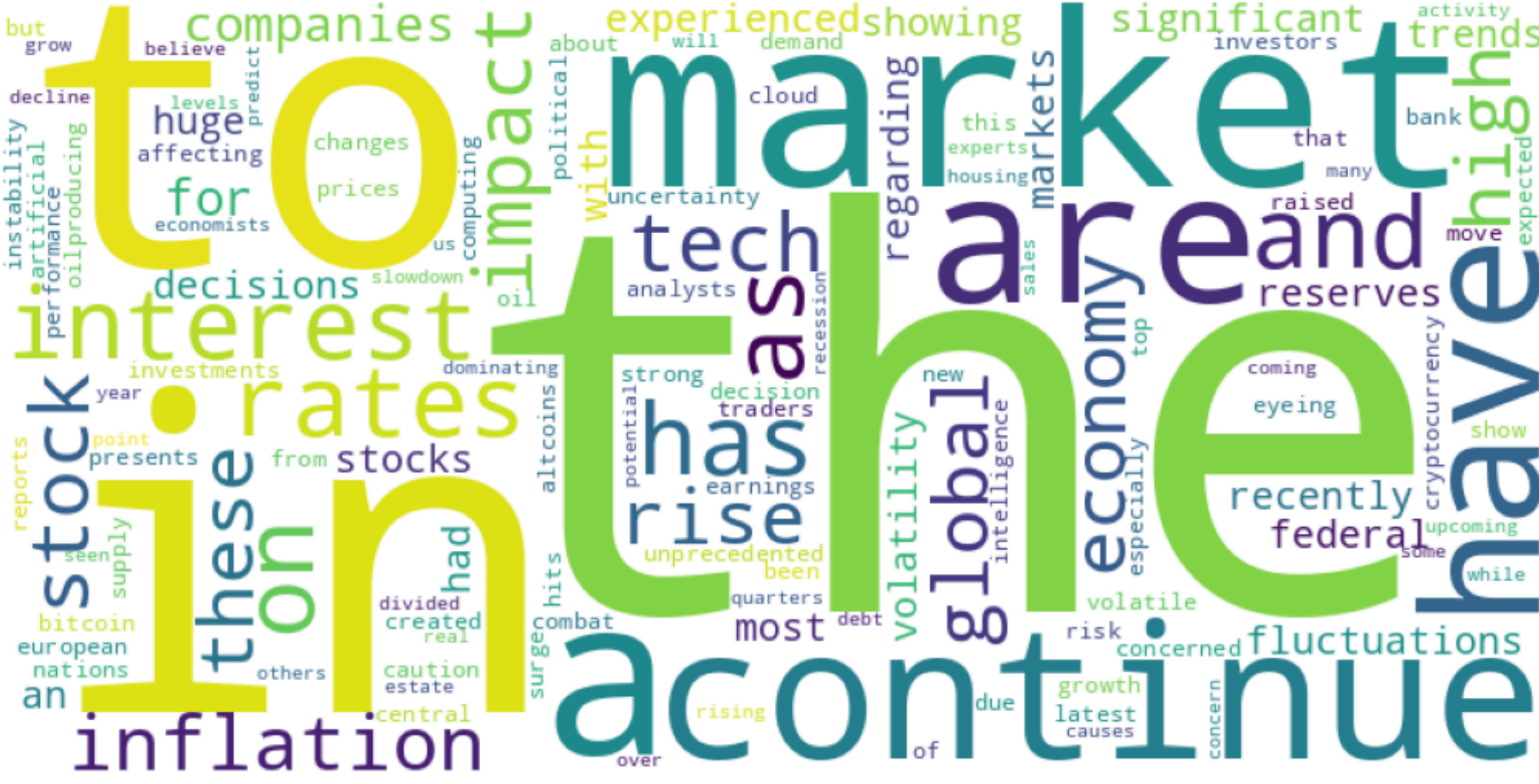
Introduction to Topic Modeling

- - What is Topic Modeling?
- - A type of statistical model used to discover topics that appear in a collection of documents.
- - Purpose: Identify hidden thematic structures in large text corpora.
- - Applications:
- - Text summarization, content recommendation, document clustering, information retrieval.

dataset


financial_news_sample.csv - 내 컴퓨터에 저장됨
업그레이드
parisin12
파일 홈 삽입 레이아웃 수식 데이터 검토 보기 팬 AI 도구
공유
계정에 비밀번호가 설정되어 있지 않습니다. 다른 기기에서도 로그인 할 수 있도록 비밀번호를 설정해 주세요.
등록
A1
article_text
1 article_text
2 Stock markets have experienced significant fluctuations recently, with tech stocks showing the most volatility. The Federal Reserve's decisions regarding interest rates have had a huge imp
3 In an unprecedented move, the European Central Bank has raised interest rates to combat inflation. This decision has created uncertainty in the global economy, affecting stock market pe
4 Cryptocurrency investments continue to surge as Bitcoin hits a new high. Investors are eyeing altcoins, but analysts caution that the volatile market presents high risk.
5 Global oil prices have been on the rise due to political instability in oil-producing nations. Traders are concerned about the impact of these changes on supply and demand.
6 The latest earnings reports from top tech companies show strong growth, especially in cloud computing and artificial intelligence. These companies are expected to continue dominating
7 The real estate market has seen a decline in sales as interest rates continue to rise. Many experts predict a slowdown in housing market activity in the coming year.
8 Economists are divided over the potential for a recession in the US. Some believe the economy will continue to grow, while others point to rising debt levels and inflation as causes for cor

total word crowd




Topic #1
impact
rates
inflation
rise
continue
economy
tech
global
stock
market

What is LDA (Latent Dirichlet Allocation)?

- - LDA is a generative probabilistic model used to discover topics in a collection of documents.
- - Key Ideas:
- - Topics: distributions over words.
- - Documents: mixtures of topics.
- - Words: assigned to topics based on the model.

How LDA Works – Simplified

- 1. Input: A collection of documents (text corpus).
- 2. Process:
- – LDA assigns each word in each document to a topic.
- – Adjusts topic assignments to maximize likelihood.
- 3. Output: Topics that best represent the themes in the documents.

1. Importing Libraries:

content_copy

```
import pandas as pd
import numpy as np
import nltk
from nltk.corpus import stopwords
from sklearn.feature_extraction.text
import TfidfVectorizer
from gensim import corpora
from gensim.models import LdaModel
import pyLDAvis.gensim_models
import matplotlib.pyplot as plt
from wordcloud import WordCloud
from sklearn.decomposition import
LatentDirichletAllocation
```

```
# TF-IDF Vectorization
vectorizer = TfidfVectorizer(max_df=0.95, min_df=2, stop_words='english')
X = vectorizer.fit_transform(data['processed_text'])
```

Example of LDA Topics

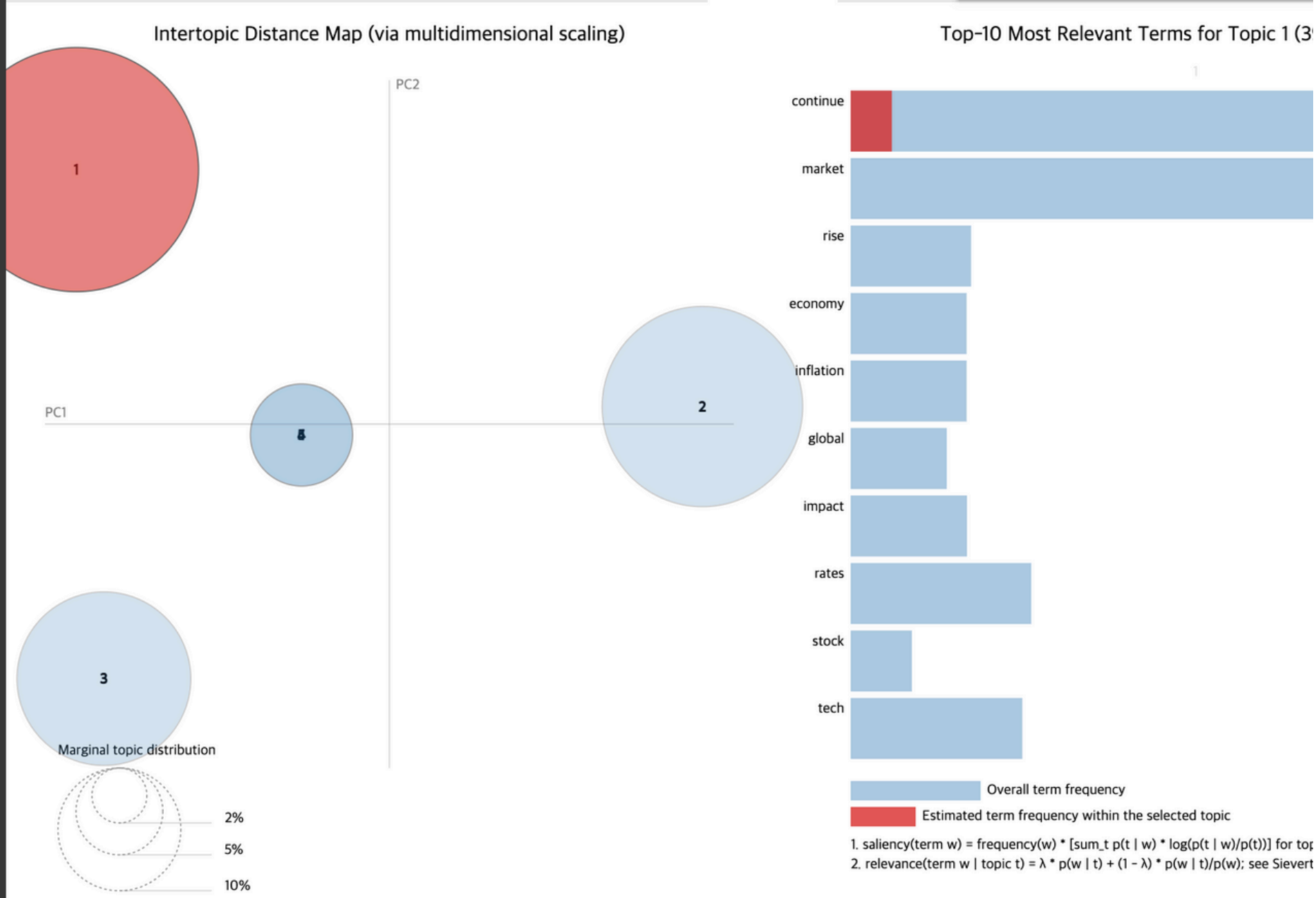
- – Topic 1: Stock Market, Interest Rates, Inflation
- – Topic 2: Cryptocurrency, Bitcoin, Investment
- – Topic 3: Real Estate, Housing Market, Sales
- – Topic 4: Tech Earnings, Artificial Intelligence, Growth

Why LDA is Useful

- – Topic Discovery: Helps identify underlying themes in large datasets.
- – Text Categorization: Improves document classification.
- – Data Exploration: Identifies meaningful patterns in unstructured data.

pyLDAvis: Visualization of LDA Results

- - pyLDAvis provides interactive ways to visualize topic modeling results.
- - Key Features:
- - Interactive visualization of topic distribution and word frequency.
- - Topic-topic distance using dimensionality reduction (t-SNE, MDS).
- - Global and local views of topics and word associations.


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
Marginal topic distribution
2%
5%
10%
Top-10 Most Relevant Terms for Topic 1 (3
continue
market
rise
economy
inflation
global
impact
rates
stock
tech
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for to
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert

Key Components of pyLDAvis

- 1. Topic-Word Distribution: Most frequent words in each topic.
- 2. Topic-Topic Distance: Relationship between topics using 2D map.
- 3. Interactivity: Hover over points to view top words for each topic.

pyLDAvis

```
# Replace the import statement
import pyLDAvis
import pyLDAvis.lda_model
# Replace pyLDAvis.sklearn.prepare with pyLDAvis.lda_model.prepare
# Use 'pcoa' instead of 'tsne' for MDS
vis = pyLDAvis.lda_model.prepare(lda, X, vectorizer, mds='pcoa') # Updated visualization function call

# Display the visualization
pyLDAvis.display(vis) # Assuming you want to display the visualization in a notebook environment
```

How to Use pyLDAvis with LDA

- - Train an LDA model on your text data.
- - Prepare the output from the LDA model and term-document matrix.
- - Use pyLDAvis to visualize the topics and their distributions.

Visualizing Financial Data Topics

- – Use pyLDAvis to explore how topics are distributed across financial articles.
- – Identify key words for each topic (e.g., “stocks,” “tech,” “cryptocurrency”).
- – Explore topic relationships and gain insights into market trends.

Applications of LDA and pyLDAvis in Finance

- – Financial Text Analysis: Discover trends and themes in financial news.
- – Sentiment Analysis: Combine LDA with sentiment analysis to understand market sentiment.
- – Market Prediction: Predict trends by analyzing the emergence of key topics.

Conclusion

- – LDA and pyLDAvis help uncover topics and visualize them in large datasets.
- – Applications in finance include analyzing market trends, understanding sentiments, and predicting market behavior.
- – pyLDAvis enhances topic model interpretability through interactive visualization.

Q&A

- Questions?