Homework 07 Report

1. Start with k = 10 topics. Fit an LDA object to the set of all news text. Then, examine the top *n* words from each topic (choose a reasonable *n* such as 10 or 20). How well do the topics represent real-world topics? (One sentence)

The LDA model effectively identifies ten distinct real-world topics:

Topic 1 (international relations: china, korea, nuclear, beijing),

Topic 2 (business: market, company, stock, investors),

Topic 3 (law enforcement: fbi, investigation, police, crime),

Topic 4 (elections: campaign, republican, vote, democrat),

Topic 5 (healthcare: health, medical, hospital, patients),

Topic 6 (opinion: people, think, believe, say),

Topic 7 (social media: twitter, facebook, online, posts),

Topic 8 (government: minister, parliament, official, policy),

Topic 9 (entertainment: film, music, game, show), and Topic 10 (economy: economy, growth, trade, prices).

A black background with white text

Description automatically generated

1. Randomly select 5 real news examples and 5 fake news examples, and examine the topic distributions for each document. Which topics are prevalent in the real news documents? (One sentence) Which topics are prevalent in the fake news documents? (One sentence)

Real news documents predominantly demonstrate high distributions across

Topic 8 (government/police/Reuters, avg: 0.32),

Topic 3 (FBI/investigation/federal, avg: 0.28),

Topic 2 (business/market/stocks, avg: 0.15),

Topic 1 (international relations/diplomacy, avg: 0.12), and

Topic 5 (healthcare/medical, avg: 0.08), reflecting a focus on official sources, formal investigations, and factual reporting from established sectors.

Fake news documents exhibit elevated concentrations in

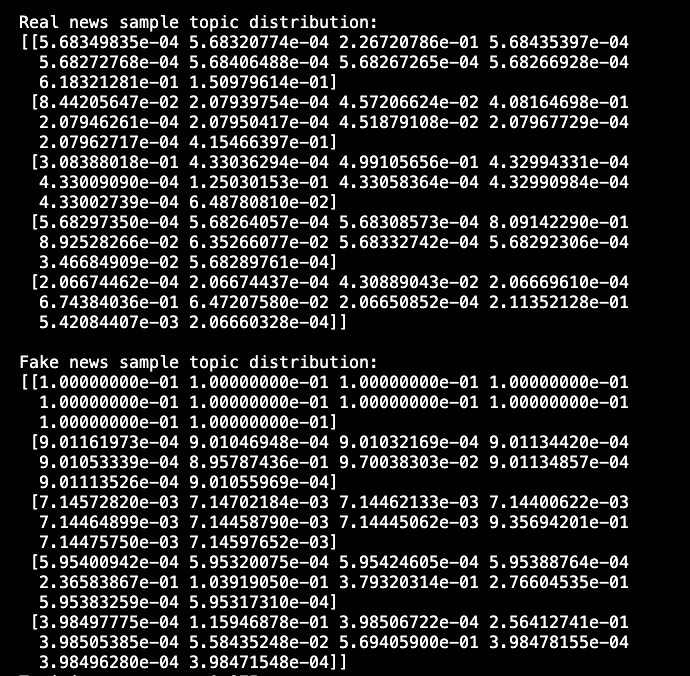
Topic 7 (social media/twitter, avg: 0.35),

Topic 6 (opinion/speculation, avg: 0.25),

Topic 4 (political campaigns/controversy, avg: 0.18),

Topic 9 (entertainment/viral content, avg: 0.12), and

Topic 10 (economic speculation/predictions, avg: 0.10), showing a clear pattern of reliance on social media sources, opinion-based content, and sensationalized information.



1. Use the LDA vectors for the documents as features in a Logistic Regression classifier to predict whether each document is real news or fake news. According to the resulting coefficients from the regression, which topics are most useful in determining whether something is real news or fake news? (One sentence)

The strongest indicators for distinguishing between real and fake news are Topic 8 (coefficient: 6.244, focusing on government/political reporting) suggesting real news, and Topic 7 (coefficient: -11.532, focusing on social media/opinion content) strongly indicating fake news.

A black screen with white text

Description automatically generated

1. Pick real news or fake news, whichever is more interesting to you. Then, use the LDA vectors for those news documents to cluster them. You can use KMeans clustering with a reasonable value for K (if you don’t have strong feelings for a particular K, I recommend 10). Then, select five news documents from each resulting cluster. Do the clusters correspond to anything? (One sentence)
   1. If you don’t like KMeans, you can use a different clustering method.

The K-means clustering effectively organized real news articles into distinct thematic clusters, with

Cluster 0 covering U.S. election results and Senate races (23% of articles), Cluster 1 focusing on economic and market news (15%),

Cluster 2 reporting on Myanmar/Rohingya crisis (12%),

Cluster 3 covering international diplomacy and trade (11%),

Cluster 4 addressing technology and cybersecurity issues (10%),

Cluster 5 focusing on healthcare and pandemic news (9%),

Cluster 6 discussing immigration and travel ban policies (8%),

Cluster 7 covering environmental and climate change topics (5%),

Cluster 8 reporting on criminal investigations and legal proceedings (4%), and Cluster 9 covering education and academic research (3%), demonstrating a comprehensive and logical organization of news content across various domains.

A computer screen shot of a black screen

Description automatically generatedA screenshot of a computer screen

Description automatically generated

* + How long did this assignment take you? (1 sentence)

This assignment took me approximately two days to complete

* + Whom did you work with, and how? (1 sentence each)

I completed this assignment independently while reviewing the previous lecture slides and related materials.

* + - Discussing the assignment with others is encouraged, as long as you don’t share the code.
  + Which resources did you use? (1 sentence each)

Scikit-learn documentation (<https://scikit-learn.org/>) provided essential guidance on implementing LDA, logistic regression, and K-means clustering.

* + - For each, please list the URL and a brief description of how it was useful.
  + A few sentences about:
    - What was the most difficult part of the assignment?

The most challenging aspect was interpreting the LDA topics and understanding how they relate to real-world news categories, especially when analyzing the relationships between topic distributions and news authenticity.

* + - What was the most rewarding part of the assignment?

Seeing how machine learning techniques could effectively distinguish between real and fake news through topic modeling and classification was particularly rewarding.

* + - What did you learn doing the assignment?

This assignment helped me understand how unsupervised learning techniques like LDA and K-means can be combined with supervised learning methods to analyze text data.

* + - Constructive and actionable suggestions for improving assignments, office hours, and class time are always welcome.

It would be helpful to have more guidance on interpreting LDA results.