NLP Analysis Report

By Abhishikth

INTRODUCTION

This report documents the Week-4 NLP assignment, demonstrating how Natural Language Processing (NLP) techniques can be applied to analyse a small collection of text documents. The analysis focuses on extracting key terms, uncovering word relationships, and identifying hidden topics within the corpus.

OBJECTIVES

The goal is to apply and explain core NLP methods which are Text preprocessing, TF-IDF, Word2Vec embeddings, and topic modelling (LDA)—to show how they work together to analyse unstructured text data and uncover insights.

Methodology

1. Text Preprocessing:
    - Convert all text to lowercase
    - Remove punctuation and stopwords
    - Tokenize each document into words
2. TF-IDF (Term Frequency–Inverse Document Frequency):
    - Evaluate how important each word is within a document relative to the entire corpus
    - Identify words that uniquely characterize each document
3. Word2Vec Embeddings:
    - Train a model to learn dense vector representations of words

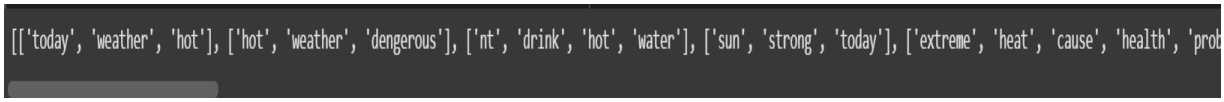
- Capture semantic relationships (similar words are close in vector space)

4. Topic Modeling (LDA):
   - Discover latent topics by representing documents as mixtures of topics and topics as mixtures of words
   - Assign dominant topics to each document

Results & Observations

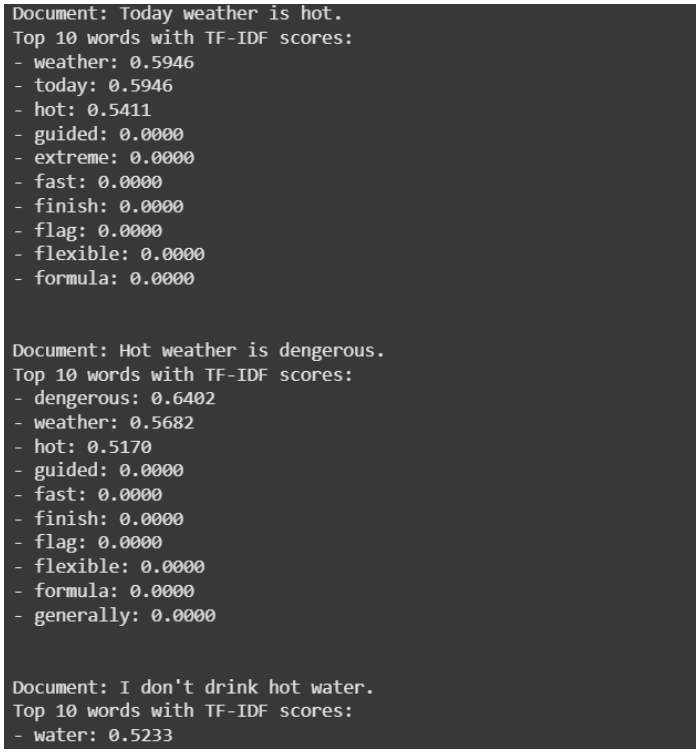
Here are some screenshots of the results and visualizations

1. **Text processing:**



- Lowercased text, removed punctuation and stopwords, and tokenized documents.
- Produced cleaned token lists for each document.

2. TF – IDF:



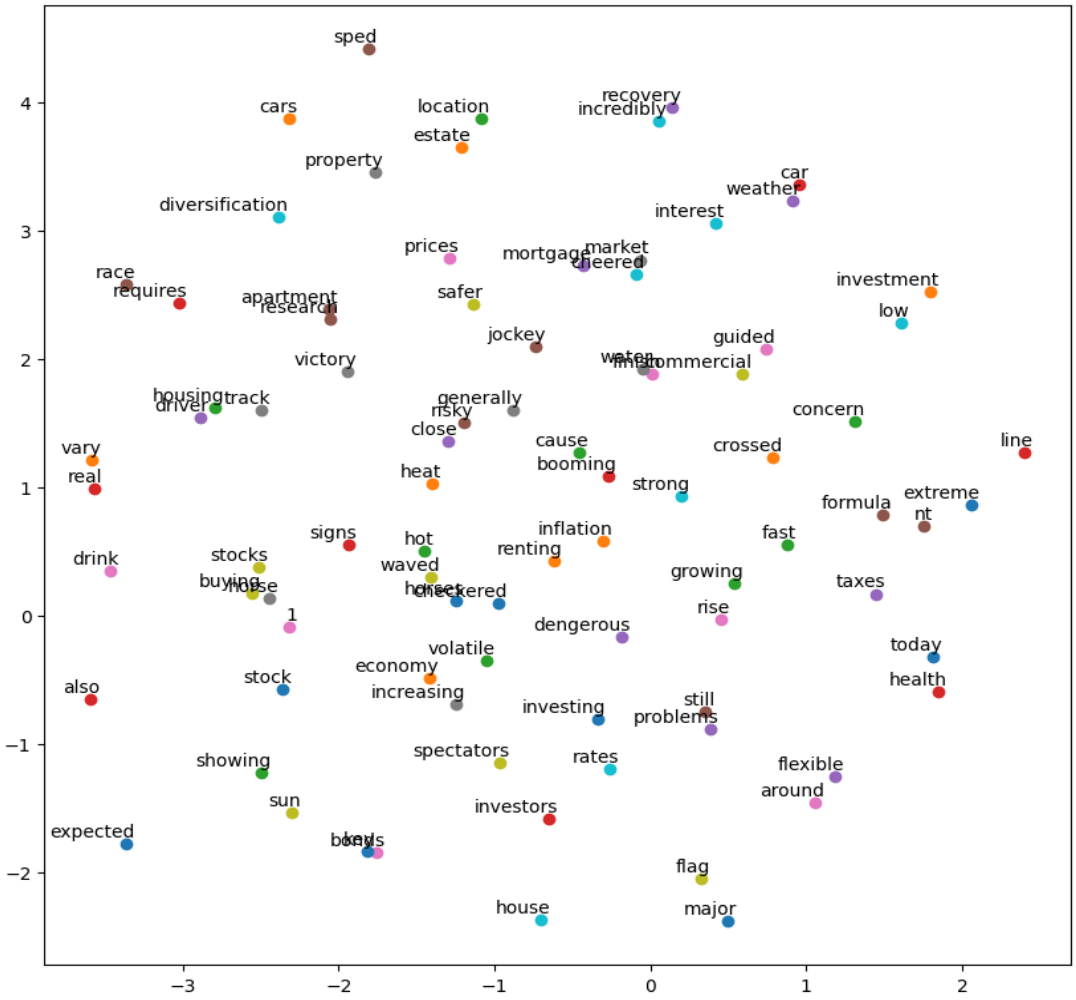
- Highlighted key terms unique to each document.
- Example: "weather", "hot" in weather docs; "race", "driver" in racing docs.
- **Weather-related documents** emphasized terms like weather, hot, sun, heat.
- **Racing-related documents** emphasized race, driver, track, finish, spectators.
- **Finance/real estate documents** emphasized stocks, investing, bonds, real estate, mortgage, rates.
- TF-IDF scores of **0.0** indicate that certain words were not significant compared to the entire corpus.

3. Word2Vec:

```
Words similar to 'weather':
- recovery: 0.2699
- car: 0.2529
- guided: 0.2008
- sun: 0.1957
- checkered: 0.1753
- hot: 0.1702
- fast: 0.1503
- investment: 0.1497
- housing: 0.1477
- water: 0.1452


Words similar to 'hot':
- waved: 0.3698
- heat: 0.2120
- concern: 0.2018
- stocks: 0.1992
- recovery: 0.1888
- renting: 0.1727
- incredibly: 0.1713
- weather: 0.1702
- spectators: 0.1528
- horses: 0.1485


Words similar to 'race':
- requires: 0.3079
- waved: 0.2796
- safer: 0.2344
- investment: 0.1906
- property: 0.1787
```



- t-SNE model used to learn the word embeddings to capture word relationships.
- Similar words (e.g., "race" ~ "track") identified despite small dataset.
- t-SNE visualization showed rough word clustering.

4. Topic modelling:

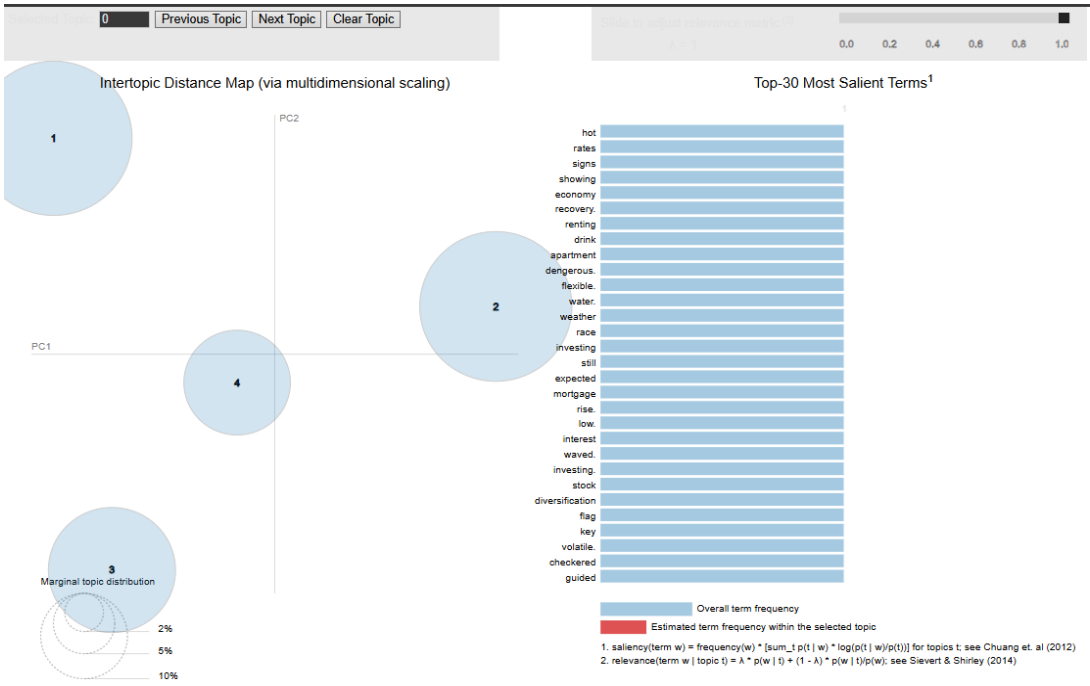
```
Top words for each topic:
Topic 1:
0.040*"investing" + 0.022*"spectators" + 0.022*"crossed" + 0.022*"finish" + 0.022*"horses" + 0.022*"cheered" + 0.022*"line." + 0.022*"cause" + 0.022*"1" + 0.022*"cars"

Topic 2:
0.041*"estate" + 0.041*"real" + 0.041*"race" + 0.023*"around" + 0.023*"driver" + 0.023*"track." + 0.023*"sped" + 0.023*"car" + 0.023*"also" + 0.023*"commercial"

Topic 3:
0.052*"rates" + 0.029*"market" + 0.029*"still" + 0.029*"expected" + 0.029*"mortgage" + 0.029*"rise." + 0.029*"low." + 0.029*"interest" + 0.029*"waved." + 0.029*"investing."

Topic 4:
0.063*"hot" + 0.035*"weather" + 0.035*"signs" + 0.035*"showing" + 0.035*"economy" + 0.035*"recovery." + 0.035*"renting" + 0.035*"drink" + 0.035*"apartment" + 0.035*"dengerous."

Dominant topic for each document:
Document: "Today weather is hot." -> Dominant Topic: 2 (Score: 0.8097)
Document: "Hot weather is dengerous." -> Dominant Topic: 4 (Score: 0.8115)
Document: "I don't drink hot water." -> Dominant Topic: 4 (Score: 0.8124)
Document: "The sun is very strong today." -> Dominant Topic: 1 (Score: 0.8122)
Document: "Extreme heat can cause health problems." -> Dominant Topic: 1 (Score: 0.8748)
Document: "Formula 1 cars are incredibly fast." -> Dominant Topic: 1 (Score: 0.8748)
Document: "The race car driver sped around the track." -> Dominant Topic: 2 (Score: 0.8927)
Document: "Spectators cheered as the horses crossed the finish line." -> Dominant Topic: 1 (Score: 0.8927)
Document: "It was a close race to the finish." -> Dominant Topic: 2 (Score: 0.8123)
Document: "The jockey guided the horse to victory." -> Dominant Topic: 3 (Score: 0.8498)
Document: "The checkered flag waved." -> Dominant Topic: 3 (Score: 0.8123)
Document: "The stock market is volatile." -> Dominant Topic: 3 (Score: 0.8112)
Document: "Investing in stocks can be risky." -> Dominant Topic: 1 (Score: 0.8123)
Document: "Bonds are generally safer than stocks." -> Dominant Topic: 2 (Score: 0.8498)
Document: "Diversification is key in investing." -> Dominant Topic: 3 (Score: 0.8123)
Document: "The economy is showing signs of recovery." -> Dominant Topic: 4 (Score: 0.8499)
Document: "Interest rates are expected to rise." -> Dominant Topic: 3 (Score: 0.8499)
Document: "Inflation is a concern for investors." -> Dominant Topic: 2 (Score: 0.8122)
Document: "Real estate prices are increasing." -> Dominant Topic: 2 (Score: 0.8491)
Document: "Buying a house is a major investment." -> Dominant Topic: 1 (Score: 0.8498)
Document: "The housing market is booming." -> Dominant Topic: 1 (Score: 0.8102)
Document: "Mortgage rates are still low." -> Dominant Topic: 3 (Score: 0.8499)
Document: "Renting an apartment can be more flexible." -> Dominant Topic: 4 (Score: 0.8123)
Document: "Property taxes vary by location." -> Dominant Topic: 2 (Score: 0.8498)
Document: "Commercial real estate is also growing." -> Dominant Topic: 2 (Score: 0.8744)
Document: "Investing in real estate requires research." -> Dominant Topic: 1 (Score: 0.8715)
```



- Using Latent Dirichlet Allocation (LDA), the 26 documents were grouped into **4 latent topics**.
- **Topic 1** top words are investing, spectators, crossed, finish, horses, cheered, line, cause, 1, cars. A mix of **investing** terms with some **horse racing / Formula 1** keywords.
- **Topic 2** top words are estate, real, race, around, driver, track, sped, car, also, commercial. Overlap of **real estate** and **car racing** terms.
- **Topic 3** top words are rates, market, still, expected, mortgage, rise, low, interest, waved, investing. **Finance and interest rate** theme. No overlapping of data seen here.
- **Topic 4** top words are hot, weather, signs, showing, economy, recovery, renting, drink, apartment, dengerous. A mix of **weather/heat** and **economic indicators**.
- Topics are **not perfectly separated** because the dataset is small and contains **three different domains** (weather, racing, finance/real estate).
- Words from unrelated domains sometimes appear together in a single topic.
- With **more documents per domain**, LDA would separate these themes more cleanly.

Conclusion

NLP techniques like TF-IDF, Word2Vec, and topic modelling effectively extracted key terms, revealed hidden word relationships, and uncovered latent themes in the text corpus, demonstrating their value for deeper text understanding.