NLP Assignment Report

Week 4 – Text Analysis and Topic Discovery

---

1. Introduction

Natural Language Processing (NLP) focuses on enabling machines to understand, process, and analyze human language.
In this assignment, we explore different NLP techniques for **text preprocessing, vectorization, embeddings, and topic modeling** using a small dataset related to **nature, environment, and technology**.

The objectives of this assignment were:

- To apply **text preprocessing techniques** (tokenization, stopword removal, lowercasing).
- To analyze text using **TF-IDF** and extract important words.
- To train a **Word2Vec model** and explore semantic similarity.
- To perform **Topic Modeling (LDA)** and interpret discovered topics.
- To visualize results using **PCA** and **pyLDAvis**.

---

2. Dataset

A synthetic dataset of **10–15 short sentences** was created covering topics of **climate change, data science, technology, conservation, and nature**.

Example Sentences:

1. Climate change is affecting forests and wildlife.
2. Technology and data analysis improve weather forecasting.
3. The ocean is full of diverse marine life.
4. Machine learning and data science are transforming industries.
5. Wildlife conservation is crucial for maintaining balance.

The dataset is small but diverse enough to show results for TF-IDF, Word2Vec, and LDA.

---

3. Methodology

3.1 Text Preprocessing

Steps performed on each sentence:

- **Lowercasing** → Converts all words to lowercase.
- **Tokenization** → Splits text into individual words.
- **Stopword Removal** → Removes common words like “is”, “the”, “and”.
- **Filtering** → Keeps only alphabetic words.

Example:
Input: `"Climate change is affecting forests and wildlife"`
Output: `["climate", "change", "affecting", "forests", "wildlife"]`

3.2 TF-IDF Analysis

- Used `TfidfVectorizer` from scikit-learn.
- Extracted top 10 words per document.
- Captures importance of words relative to document and corpus.

Example Results:

- Doc 1 → `{climate, change, wildlife}`
- Doc 2 → `{data, analysis, forecasting}`

3.3 Word2Vec Embeddings

- Trained `gensim.models.Word2Vec` with vector size 50.
- Learned semantic relationships between words.
- Query Example:
  - Similar to **“data”** → `{analysis, science, machine, learning}`
- Visualized embeddings using **PCA (2D scatter plot)**.
- Observed clustering of **nature words** vs **technology words**.

3.4 Topic Modeling (LDA)

- Created dictionary and Bag-of-Words corpus using `gensim.corpora.Dictionary`.

- Built an **LDA Model** with **3 topics**.
- Extracted top 5 words per topic.

Results:

- **Topic 0**: {forest, rivers, ocean, wildlife} → **Nature/Environment**
- **Topic 1**: {climate, biodiversity, conservation, balance} → **Climate/Conservation**
- **Topic 2**: {data, analysis, machine, learning, technology} → **Technology/Data**
- Used **pyLDAvis** to visualize topic distribution interactively.

---

4. Results & Insights

- **TF-IDF** highlighted domain-specific important words in each sentence.
- **Word2Vec** grouped related words (e.g., *data–analysis–machine learning* and *forest–wildlife–climate*).
- **LDA Topic Modeling** successfully separated the dataset into **Nature, Climate, and Technology topics**.
- Visualization confirmed that topics and word clusters are interpretable even for a small dataset.

---

5. Conclusion

This assignment demonstrated key NLP techniques on a small dataset:

- Preprocessing helps clean raw text.
- TF-IDF highlights important keywords per document.
- Word2Vec embeddings capture semantic meaning of words.
- LDA reveals hidden topics and provides interpretable clusters.

These methods are foundational for real-world applications like **sentiment analysis, document clustering, and text classification**.

---

6. References

- Jurafsky & Martin, *Speech and Language Processing*
- scikit-learn Documentation
- Gensim Documentation
- NLTK Toolkit