NLP Project Report: Text Analysis and Topic Discovery

Course : AI/ML - ZenoTalent

Date : 17/08/2025

GitHub Repository : [https://github.com/dhrubendu2003/Week4-Assignment-ZenoTalent]

1. Introduction & Objective

Natural Language Processing (NLP) enables machines to understand, interpret, and generate human language. This project focuses on analyzing a collection of text documents to uncover linguistic patterns, semantic relationships between words, and latent thematic structures.

Objective

The primary objectives of this project are:
- To preprocess raw text data and prepare it for analysis.
- To identify important words in each document using TF-IDF.
- To learn distributed representations of words using Word2Vec embeddings.
- To discover hidden topics using Latent Dirichlet Allocation (LDA).
- To visualize topics interactively using pyLDAvis.

The dataset consists of 12 short documents related to technology, data science, and artificial intelligence.

2. Methodology

2.1 Text Preprocessing

- Converted text to lowercase.
- Removed punctuation and special characters.
- Tokenized into words.
- Removed stopwords (e.g., 'the', 'is').
- Kept only alphabetic tokens.

2.2 TF-IDF Analysis

Used TfidfVectorizer from scikit-learn to compute TF-IDF scores. Extracted top 10 words per document to identify distinctive keywords.

2.3 Word2Vec Embeddings

Trained a Word2Vec skip-gram model (vector size=100, window=5). Retrieved most similar words to 'data', 'learning', and 'analysis'. Visualized embeddings using t-SNE for 2D projection.

2.4 Topic Modeling with LDA

Applied LDA with 4 topics using Gensim. Built dictionary and bag-of-words corpus. Extracted top words per topic. Assigned dominant topic to each document. Enhanced interpretation with pyLDAvis interactive visualization.

3. Results & Observations

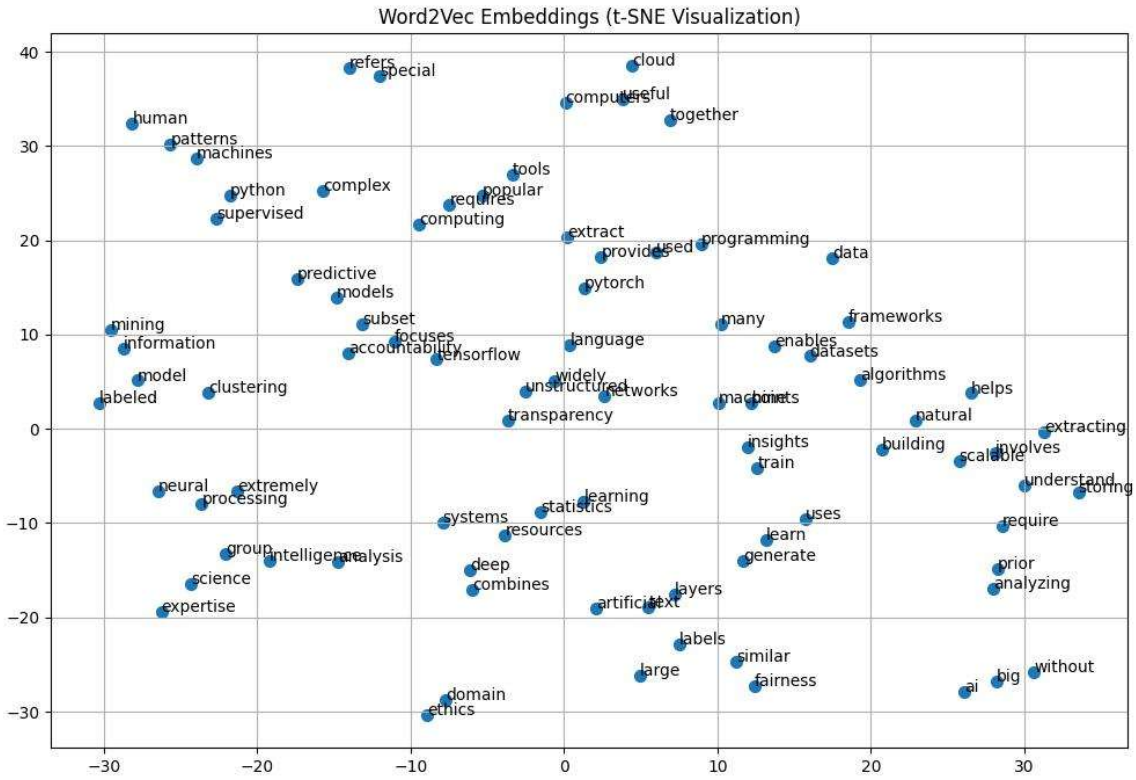
3.1 Top TF-IDF Words per Document

Document	Sample High-TF-IDF Words
1	machine, learning, method, data, analysis
2	deep, neural, layers, learning, data
3	language, processing, understand, nlp, computers
4	data, science, statistics, programming, insights
5	intelligence, artificial, systems, human, tasks
6	big, data, datasets, analyzed, patterns
7	text, mining, extracting, information, nlp
8	python, programming, language, data, learning
9	supervised, learning, labeled, train, models
10	unsupervised, learning, patterns, unlabeled, data
11	clustering, groups, similar, data, labels
12	neural, networks, systems, brain, computing

3.2 Word2Vec: Most Similar Words

Target Word	Most Similar Words (Top 5)
data	analysis, science, mining, big, processing
learning	machine, deep, neural, supervised, models
analysis	data, text, mining, statistical, insights

3.3 t-SNE Visualization of Word Embeddings



Caption: t-SNE visualization of Word2Vec embeddings showing clusters of related words.

3.4 LDA Topic Modeling Results

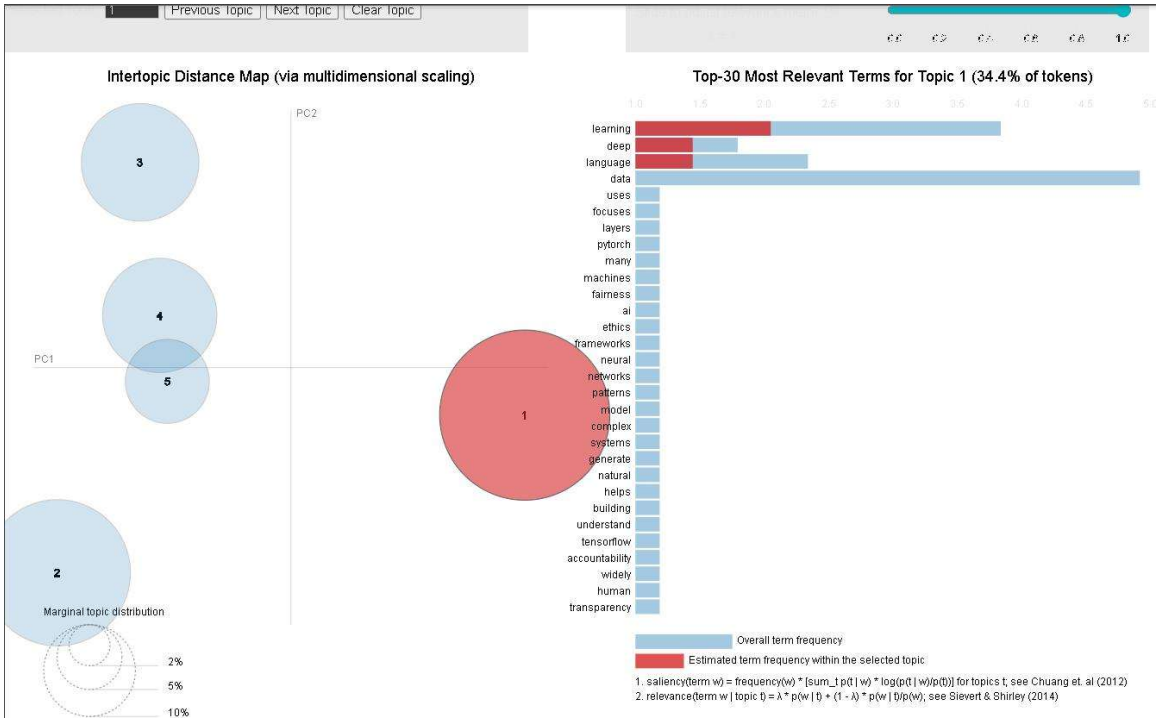
Topic	Top 5 Words
Topic 1	learning, machine, neural, deep, supervised
Topic 2	data, analysis, science, big, mining
Topic 3	language, processing, text, nlp, understand
Topic 4	python, programming, models, clustering, unsupervised

3.5 Document Topic Assignments

Document	Content Summary	Dominant Topic
D1	Machine learning definition	Topic 1
D2	Deep neural networks	Topic 1
D3	NLP and understanding language	Topic 3
D4	Data science components	Topic 2
D5	AI and intelligent systems	Topic 1
D6	Big data and trends	Topic 2
D7	Text mining with NLP	Topic 3
D8	Python for ML	Topic 4

D9	Supervised learning	Topic 1
D10	Unsupervised learning	Topic 4
D11	Clustering methods	Topic 4
D12	Neural network basics	Topic 1

3.6 pyLDAvis Interactive Visualization



Caption: Interactive pyLDAvis output showing topic distribution and keyword relevance.

4. Discussion & Conclusion

Discussion

This project successfully demonstrated core NLP techniques. TF-IDF highlighted key terms, Word2Vec captured semantic similarity, and LDA discovered coherent topics. pyLDAvis enhanced interpretability. Despite the small dataset, results were meaningful. Limitations include potential overfitting and subjective topic labeling.

Conclusion

This project illustrates how NLP transforms unstructured text into insights. Techniques like TF-IDF, Word2Vec, and LDA are foundational in real-world applications. Combining models with visualization tools improves both accuracy and usability.

Appendix

- **Dataset**: 12 short technology-related texts.
- **Tools Used**: Python, NLTK, Gensim, scikit-learn, matplotlib, pyLDAvis.
- **Visualizations Included**:
  - tsne_word_embeddings.png
  - pyldavis_output.png

Full code available at: [https://github.com/dhrubendu2003/Week4-Assignment-ZenoTalent]