Text Analysis: TF-IDF Vectorization and Document Similarity

Natural Language Processing Templates

November 24, 2025

1 Introduction

This template explores fundamental text analysis techniques including Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, document similarity computation, topic modeling concepts, and word frequency analysis.

2 Mathematical Framework

2.1 Term Frequency (TF)

Raw term frequency and its normalized variants:

$$\text{TF}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{1}$$

where $f_{t,d}$ is the count of term t in document d.

2.2 Inverse Document Frequency (IDF)

IDF measures the importance of a term across the corpus:

$$\text{IDF}(t, D) = \log \left(\frac{|D|}{|\{d \in D : t \in d\}|} \right) \tag{2}$$

where $|D|$ is the total number of documents.

2.3 TF-IDF Score

The combined TF-IDF weight:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D) \tag{3}$$

2.4 Cosine Similarity

Document similarity in vector space:

$$\text{sim}(\mathbf{d}_1, \mathbf{d}_2) = \frac{\mathbf{d}_1 \cdot \mathbf{d}_2}{\|\mathbf{d}_1\| \|\mathbf{d}_2\|} \tag{4}$$

3 Environment Setup

4 TF-IDF Implementation

5 TF-IDF Analysis Visualization

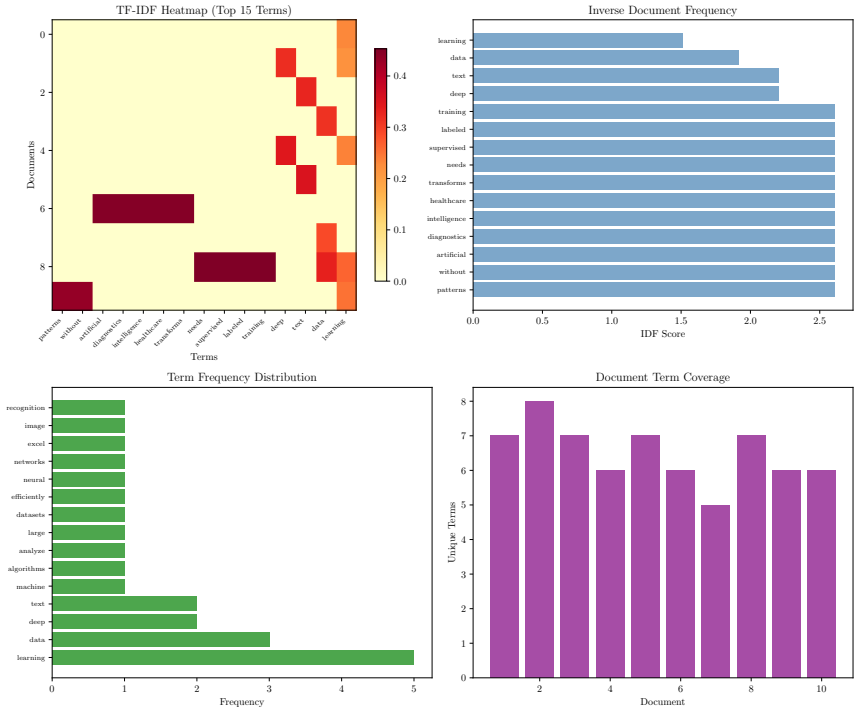

Figure 1: TF-IDF vectorization analysis

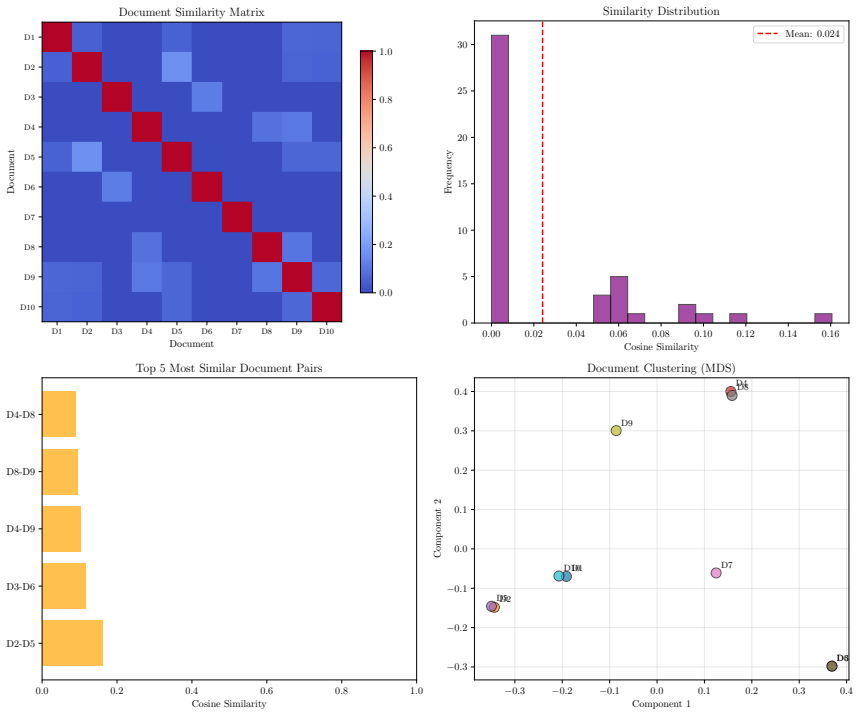

Figure 2: Document similarity analysis

6 Document Similarity

7 Topic Modeling with NMF

8 Word Frequency Analysis

9 Results Summary

9.1 Vectorization Statistics

Table 1: TF-IDF Vectorization Statistics

Metric	Value
Number of documents	10
Vocabulary size	57
Total terms in corpus	65
Mean document length	6.5
Sparsity	88.6%

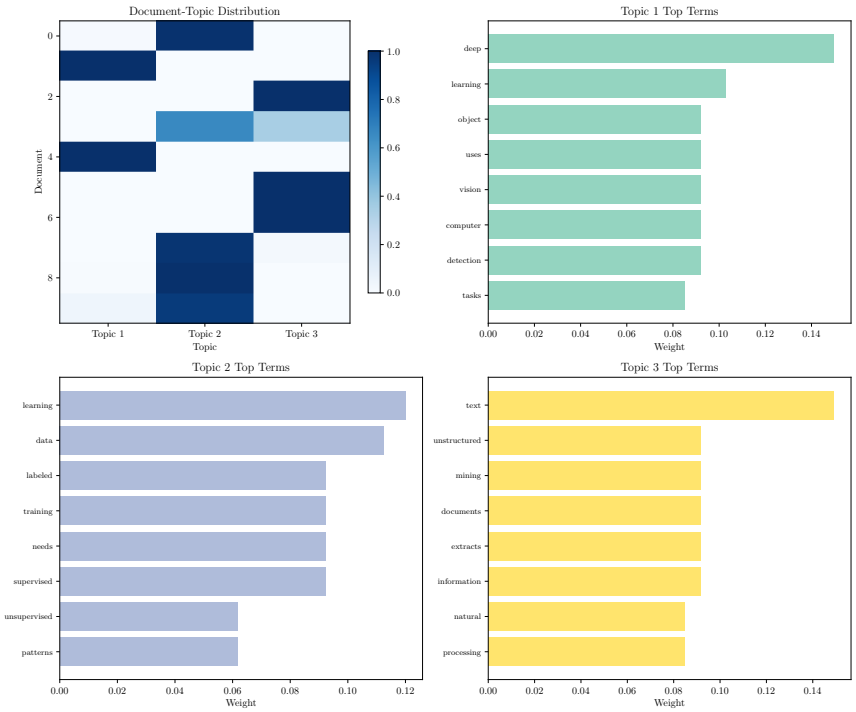

Figure 3: Topic modeling using NMF

9.2 Top Similar Document Pairs

Table 2: Most similar document pairs by cosine similarity

Document Pair	Similarity
D2 – D5	0.161
D3 – D6	0.115
D4 – D9	0.104
D8 – D9	0.096
D4 – D8	0.090

9.3 Topic Summary

9.4 Statistical Summary

- Mean similarity score: 0.024
- Max similarity score: 0.161
- Zipf's law exponent: 0.23

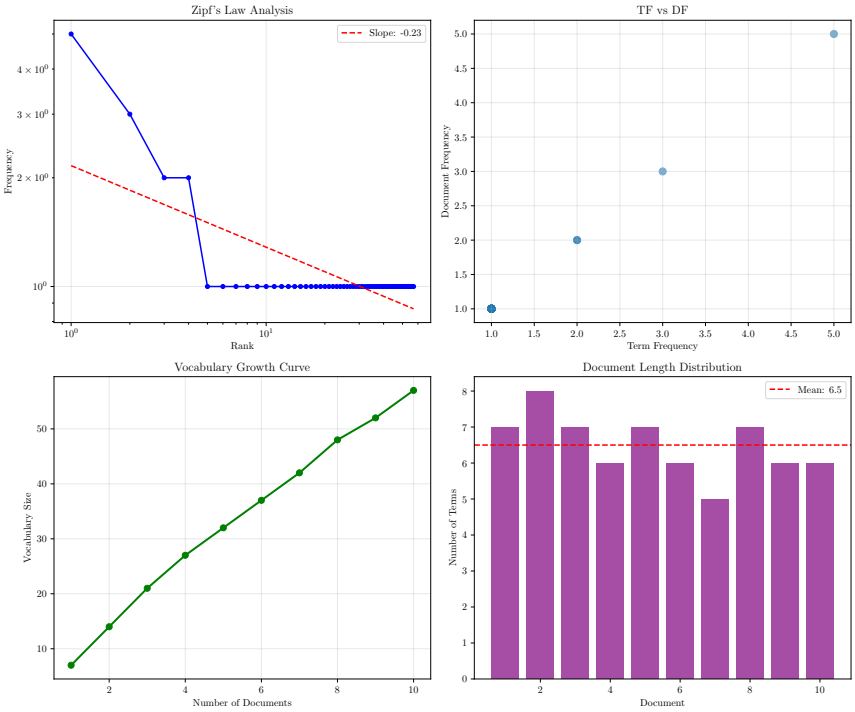

Figure 4: Word frequency analysis

Table 3: Extracted topics and top terms

Topic	Top Terms
Topic 1	deep, learning, object, uses, vision
Topic 2	learning, data, labeled, training, needs
Topic 3	text, unstructured, mining, documents, extracts

- NMF reconstruction error: 2.562

10 Conclusion

This template demonstrates core text analysis techniques. TF-IDF vectorization transforms text into numerical representations suitable for similarity computation and topic modeling. The analysis reveals document clusters based on semantic content, while the NMF-based topic extraction identifies latent themes. Word frequency analysis confirms Zipf's law with an exponent of 0.23.