Class Notes on Text Classification and Vector Space Models

Introduction

In this class, we discussed the concepts of document-term matrix, frequency-based embeddings, and techniques for text classification, specifically using vector space models to represent documents in numerical vector form.

Document-Term Matrix

Definition and Construction

A **document-term matrix** is a representation of a text corpus where:

- **Documents** are in the rows (D1, D2, …, Dn).
- **Terms** or unique tokens in your corpus are in the columns (T1, T2, ..., Tk).

Each matrix entry (i, j) represents the count (frequency) of token Tj in document Di【4:1†transcript】.

Problems of Document-Term Matrix

1. **High Dimensionality**: The number of unique tokens results in a high number of dimensions which makes it computationally expensive.
2. **No Context or Order**: The order of words and context are not captured. Sentences with reversed word orders look identical 【4:2†transcript】.
3. **Lack of Semantic Understanding**: Words with different meanings but the same token (e.g., "Python") aren't distinguished

Solutions to Problems

- Reduce dimensions by removing stop words, stemming, lemmatization, etc.
- Use techniques like **n-grams** to retain order: Bigram, trigram, etc. allow capturing sequences【4:5†transcript】.

Term Frequency-Inverse Document Frequency (TF-IDF)

Calculating TF-IDF

TF-IDF is used to measure how important a word is to a document in a corpus. It's a weight used in information retrieval and text mining.

1. **Term Frequency (TF)**: Measures how frequently a term occurs in a document. If a term occurs multiple times, its importance is high. [\text{TF}(t) = \frac{\text{Number of times term } t \text{ appears in a document}}{\text{Total number of terms in the document}}]
2. **Inverse Document Frequency (IDF)**: Measures the importance of a term. It is calculated as: [\text{IDF}(t) = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents with term } t}\right)]
3. **TF-IDF Calculation**: [\text{TF-IDF}(t, d) = \text{TF}(t) \times \text{IDF}(t)]

Importance of a term increases proportionally to the number of times a word appears in the document and is offset by the frequency of the word in the corpus【4:10†transcript】【4:11†transcript】.

Example and Application

Similarity Metrics

1. **Euclidean Distance**: Measures the root of square differences between vectors.
2. **Cosine Similarity**: Measures the cosine of the angle between two vectors, useful for measuring similarity irrespective of size.

Cosine similarity is preferred for document analysis where the magnitude of the word count is less significant compared to the direction defined by the word frequencies【4:11†transcript】.

Conclusion

Understanding document-term matrices, n-grams, and TF-IDF provides critical tools for handling and analyzing textual data in a meaningful way. This class gave an introduction to how text is converted into numerical formats that machine learning models can process to perform tasks like text classification and similarity computations 【4:1†transcript】【4:13†transcript】.