11

Machine Learning with NLP

NLP for short!

11.1
Feature Engineering in NLP.

Introduction to Feature Engineering

- Feature Engineering is the process of converting **raw data into a numerical format** that algorithms can utilize for prediction or classification
- Natural language is inherently complex and unstructured. Feature engineering in NLP is crucial for transforming text into a structured, machine-readable form

Bag of Words (BoW) - Countvectorizer

- BoW: Counts the occurrence of each word in a document, transforming text into a numerical vector. Each word becomes a feature
- **Strengths**: Simplicity and ease of implementation. It's a good starting point for basic text classification tasks
- **Limitations**: Ignores grammar and word order, treats every word equally regardless of its relevance to the text's meaning

Term Frequency-Inverse Document Frequency (TF-IDF)

- TF-IDF addresses one of BoW's key limitations by considering **not just frequency but the importance of words** within a document set
- TF-IDF reduces the weight of common words like 'the' or 'is' across documents, which are less informative, and increases the weight for words that are unique to a specific document

Term Frequency-Inverse Document Frequency (TF-IDF)

- **Advantages Over BoW**: This method allows us to surface more relevant terms in our analysis and to better distinguish between documents based on their unique content
- Despite its sophistication over BoW, **TF-IDF still does not account for the semantics of word order or context** — essential elements for true language understanding

Word2Vec Embedding

- Word2Vec represents words by their **context**, capturing **semantic relationships** in a dense vector space
- Through **neural networks**, Word2Vec predicts a word from its neighbors, or vice versa, learning vectors that place semantically similar words close together

Word2Vec Embedding

- **Semantic Insights**: This model goes beyond frequency, allowing algorithms to understand similarity and analogy based on word usage patterns
- Though powerful, **Word2Vec requires significant data and computational power**, and it **does not inherently capture the meaning of larger text** structures like sentences or paragraphs

SentenceBert

- SentenceBert adapts the powerful **BERT** model to generate embeddings that **represent the meaning of entire sentences, not just words**
- By using siamese and triplet network structures, SentenceBert is trained to understand the nuanced differences and similarities between sentences

SentenceBert

- Fit for Complex Tasks: These embeddings excel in tasks requiring deep semantic understanding, such as semantic text similarity, clustering, and information retrieval
- The trade-off for this depth of understanding is the need for greater computational resources and more complex model fine-tuning

GloVe (Global Vectors for Word Representation)

- GloVe builds a co-occurrence matrix that records how often each word appears in the context of every other word
- It combines the advantages of matrix factorization methods (like LSA) with the contextual benefits of Word2Vec, offering a rich, nuanced view of word meanings
- By capturing global corpus statistics, GloVe provides insight into the collective context that individual Word2Vec or BoW models might miss

11.2
Tasks in NLP.

Text Classification

- Utilizes algorithms to understand the subject or theme of a text and classify it into predefined categories.
- Applications: Email filtering (e.g., spam or non-spam), news article categorization, language identification.
- Example: A machine learning model classifies product reviews as either 'electronics', 'books', 'clothing', etc.

Named Entity Recognition (NER)

- The task of identifying and classifying key information (entities) in text into predefined categories.
- Entity Types: Common categories include names of people, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.
- Applications: Extracting information from resumes, automating customer support, content classification.
- Example: From the sentence "Apple Inc. was founded by Steve Jobs in California", NER would identify 'Apple Inc.' as an Organization, 'Steve Jobs' as a Person, and 'California' as a Location.

Sentiment Analysis

- The process of determining the emotional tone behind a body of text.
- How It Works: Classifies the sentiment of a text as positive, negative, or neutral using NLP and machine learning techniques.
- Applications: Brand monitoring, product analytics, customer feedback, market research.
- Example: Analyzing tweets mentioning a brand to gauge public sentiment towards its latest product launch.

Document Classification

- Similar to text classification, but focused on larger bodies of text, like entire documents or articles.
- How It Works: Involves analyzing the document as a whole, often considering structure and context more than in shorter text classification.
- Applications: Library cataloging, legal document sorting, automatic document indexing for search engines.
- Example: Classifying academic papers into different scientific fields like 'Biology', 'Computer Science', or 'Physics'.

END.