Prediction Competition 6: Text Analysis, Economic News and ML Research

Anonymized name: KentoNanami

Q1

The assignment aims to classify which of the 15,578 news article text snippets contain words related to “economic”, “economy”, “economics”, or the character sequence “econom”. Prior to importing the dataset, adjustments were made including adding a 'text' header, resolving duplicate issues caused by text starting with "=–", and removing problematic symbols like '<' which affected HTML interpretation. After ensuring data integrity, NLP functions were tested and a cleaning function was developed. The Bag of Words (BoW) method was then used to represent text as word occurrences, followed by TF-IDF to evaluate word importance relative to a corpus. Subsequently, a neural network with batch normalization and two hidden layers was constructed, utilizing 'relu' activation function. Output layer with 'softmax' activation was employed for binary classification. Assessing predictions involved creating a confusion matrix and monitoring loss graphs. To address misclassification issues, a threshold based on neural network output probability was implemented and fine-tuned.

Q2

Figure 1: A wordcloud of most common words scaled by size in the text data with a word containing 'econom'

Figure 2: A wordcloud of most common words scaled by size in the text data without a word containing 'econom'

Q3

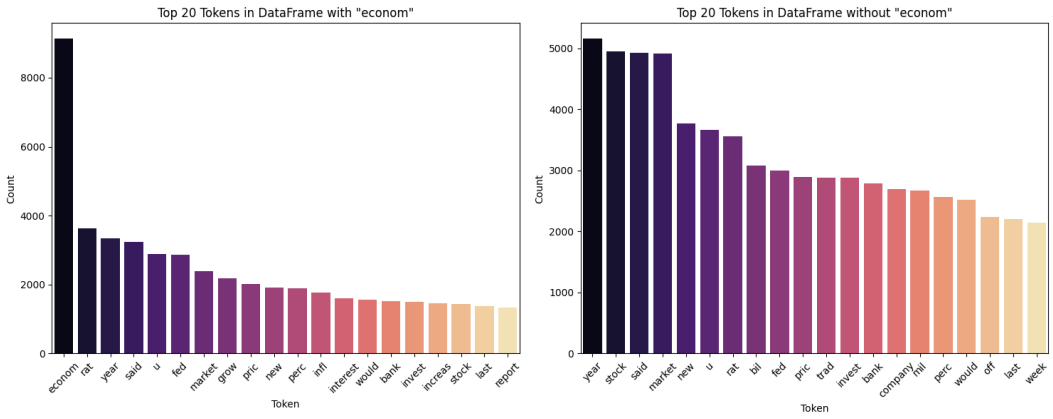


Figure 3: A multigraph plot of the counts of 20 most common words in the text data with a word containing 'econom' in it vs text data without 'econom' in it.

Q4

You

Im working on an nlp project and I am using bag of words and tdidf, can you explain those to me

ChatGPT

Certainly! In natural language processing (NLP), Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) are two popular techniques used for text processing and feature extraction.

Bag of Words (BoW)

The Bag of Words model is a simple and commonly used technique for feature extraction in NLP. It represents text data as a collection of words, disregarding grammar and word order. In BoW, a text document is represented as a "bag" (multiset) of its words, where each word is associated with a count or frequency. The order of words is not considered, only their presence or absence.

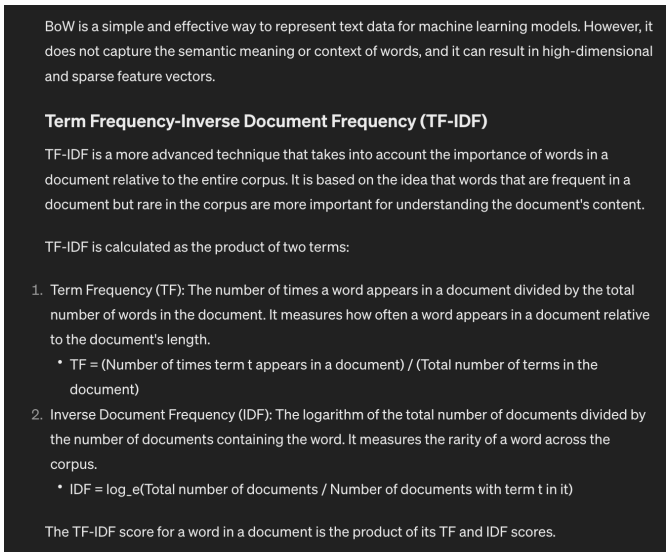
For example, consider the following two sentences:

1. "The cat sat on the mat."
2. "The dog chased the cat."

The BoW representation for these sentences would be:

1. {"the": 1, "cat": 1, "sat": 1, "on": 1, "mat": 1}
2. {"the": 1, "dog": 1, "chased": 1, "cat": 1}

Figure 4: Conversation with Chat GPT



BoW is a simple and effective way to represent text data for machine learning models. However, it does not capture the semantic meaning or context of words, and it can result in high-dimensional and sparse feature vectors.

Term Frequency-Inverse Document Frequency (TF-IDF)

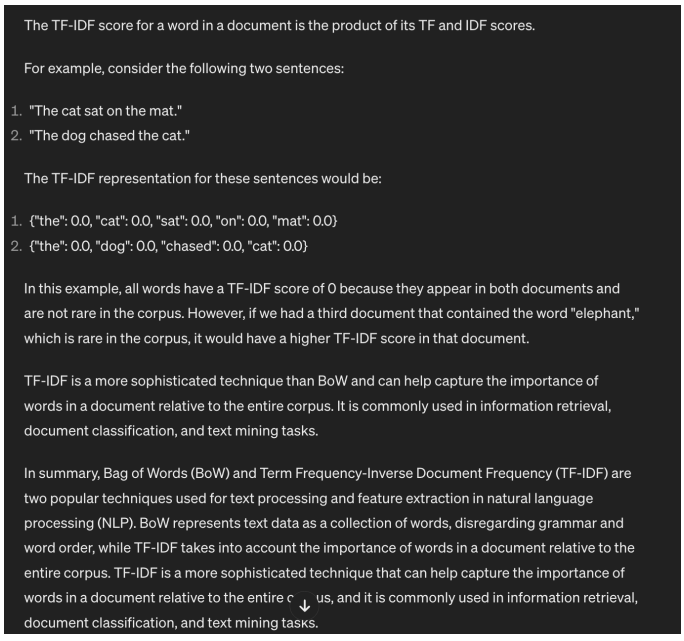
TF-IDF is a more advanced technique that takes into account the importance of words in a document relative to the entire corpus. It is based on the idea that words that are frequent in a document but rare in the corpus are more important for understanding the document's content.

TF-IDF is calculated as the product of two terms:

1. Term Frequency (TF): The number of times a word appears in a document divided by the total number of words in the document. It measures how often a word appears in a document relative to the document's length.
   - TF = (Number of times term t appears in a document) / (Total number of terms in the document)
2. Inverse Document Frequency (IDF): The logarithm of the total number of documents divided by the number of documents containing the word. It measures the rarity of a word across the corpus.
   - IDF = log_e(Total number of documents / Number of documents with term t in it)

The TF-IDF score for a word in a document is the product of its TF and IDF scores.

Figure 5: Conversation with Chat GPT



The TF-IDF score for a word in a document is the product of its TF and IDF scores.

For example, consider the following two sentences:

1. "The cat sat on the mat."
2. "The dog chased the cat."

The TF-IDF representation for these sentences would be:

1. {"the": 0.0, "cat": 0.0, "sat": 0.0, "on": 0.0, "mat": 0.0}
2. {"the": 0.0, "dog": 0.0, "chased": 0.0, "cat": 0.0}

In this example, all words have a TF-IDF score of 0 because they appear in both documents and are not rare in the corpus. However, if we had a third document that contained the word "elephant," which is rare in the corpus, it would have a higher TF-IDF score in that document.

TF-IDF is a more sophisticated technique than BoW and can help capture the importance of words in a document relative to the entire corpus. It is commonly used in information retrieval, document classification, and text mining tasks.

In summary, Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) are two popular techniques used for text processing and feature extraction in natural language processing (NLP). BoW represents text data as a collection of words, disregarding grammar and word order, while TF-IDF takes into account the importance of words in a document relative to the entire corpus. TF-IDF is a more sophisticated technique that can help capture the importance of words in a document relative to the entire c ↓ us, and it is commonly used in information retrieval, document classification, and text mining tasks.

Figure 6: Conversation with Chat GPT