**Assignment 3 Report**

**Part 1 – ML Coding of Stance**

**1. Feature Creation**

To prepare textual Reddit data for machine learning classification, we performed a detailed text preprocessing and feature engineering pipeline. This included several stages to clean, reduce noise, and optimize the feature set for predictive modeling.

**Preprocessing Steps:**

* Converted all text to lowercase to normalize the corpus.
* Removed punctuation and numbers using regex patterns.
* Tokenized the cleaned text into unigrams (individual words) using the unnest\_tokens() function from the tidytext package.
* Removed stop words from the tidytext::stop\_words list.
* Applied stemming using the SnowballC::wordStem() function to consolidate variants of the same root (e.g., "invent", "invention" → "invent").

**Filtering Rare Words:**

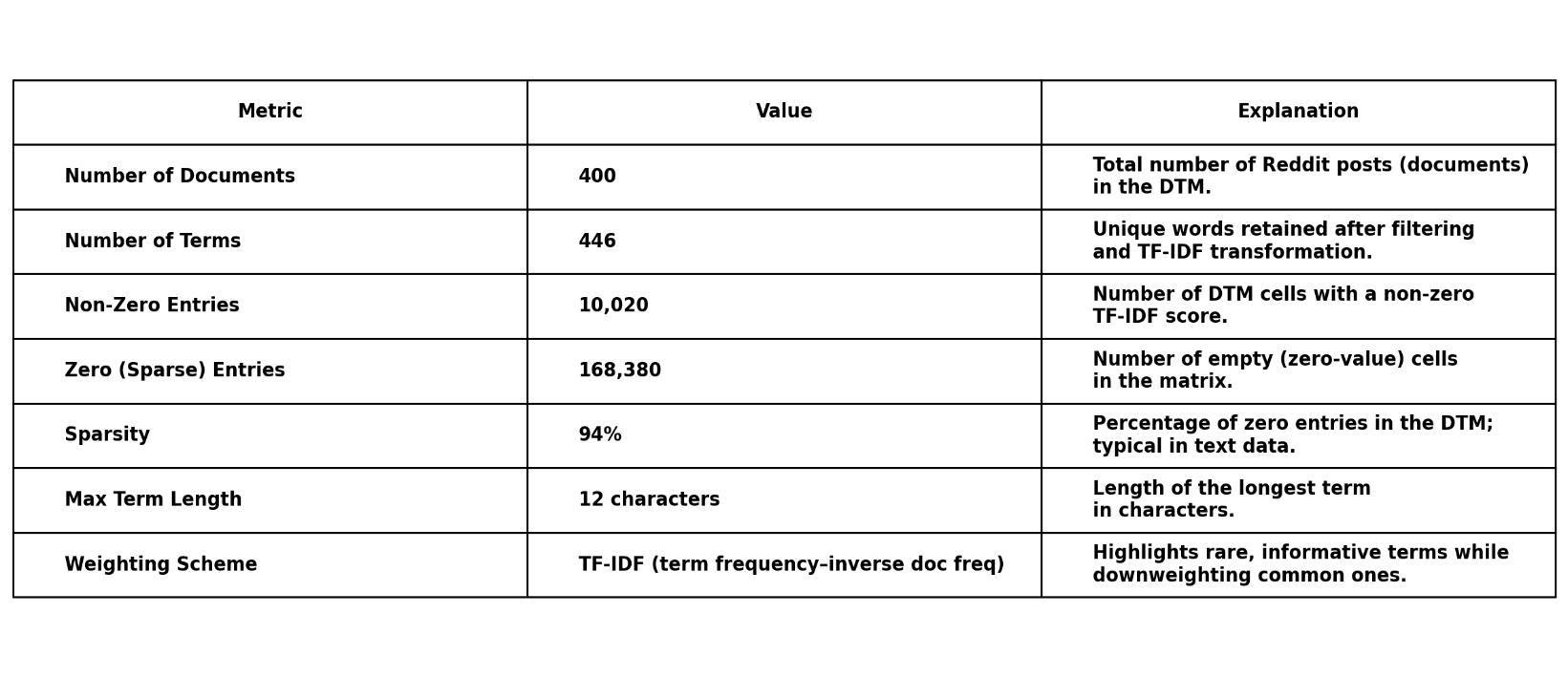
We explored the distribution of word frequencies to determine an appropriate threshold for filtering low-frequency terms. The log-scale histogram below shows the frequency of unique words in the corpus:

* Based on the visualization, we applied a minimum frequency threshold to exclude words that appeared fewer than **3 times**.
* This helped reduce the vocabulary size and focused the model on more informative terms.

**TF-IDF Weighting:**

We used TF-IDF (Term Frequency-Inverse Document Frequency) to represent words in each post. This weighting scheme highlights rare but informative words and reduces the influence of highly frequent, low-information words.

* The bind\_tf\_idf() function was used to calculate weights, and the cast\_dtm() function generated a Document-Term Matrix (DTM).
* Sparse terms were removed using a **0.98 sparsity threshold**, which retained terms appearing in at least 2% of documents

The resulting matrix had the following structure:

This balance allowed us to retain sufficient word variation for modeling while avoiding overfitting due to overly rare terms.

**2. Dataset Split**

To evaluate model performance reliably, we split the dataset into training and test sets:

* **20%** of the full dataset was randomly sampled to serve as the **test set**.
* The remaining **80%** was used to train models.
* The dataset was organized with an ID column to track original post identities, which was removed after splitting.

This split ensures that the test set is fully unseen during model training, enabling fair evaluation of model generalization.