ECON6087: Textual Analysis for Economists

Assignment 1: Bag of Words and TF-IDF

Spring 2026

General Instructions

- **Submission:** Please submit a short report and your source code.
  - **Report Format:** Accepted formats include `.md`, `.doc`, `.docx`, `.pdf`, or embedded directly within a Jupyter Notebook (`.ipynb`).
  - **Code:** You may use any programming language.
- **File Format:** This assignment uses **.parquet** files. Ensure your environment has the necessary dependencies installed to read this format (e.g., `pyarrow` or `fastparquet`).
- **Reproducibility:** Ensure your code is well-commented. If your algorithms involve randomization (e.g., classifier initialization), set a fixed **random seed** to ensure your results are reproducible.

Question 1: Text Vectorization and Supervised Learning (English)

In this exercise, you will work with the `ag_news` dataset, a collection of news articles categorized into four topics: World (0), Sports (1), Business (2), and Sci/Tech (3).

1.1 Data Preparation

Download the following files from Moodle:

- `ag_news_train.parquet`: The training dataset.
- `ag_news_test.parquet`: The testing dataset.

Both files contain two primary columns: `text` (the news snippet) and `label` (the category index).

1.2 Methodology

(a) **Preprocessing:** Load the training and testing files. Perform basic text preprocessing on the `text` column for both datasets:

- Convert text to lowercase.
- Remove punctuation and special characters.

- Remove standard English stop words.

(b) **Vectorization:** Implement two methods to transform the raw text into numerical feature vectors. You may use standard libraries (e.g., `scikit-learn`).

- **Bag of Words (BoW):** Represent each document by the count of words occurring in it.
- **TF-IDF:** Represent each document using Term Frequency-Inverse Document Frequency weights.

*Note: Fit your vectorizers (i.e., build the vocabulary) using only the **training** set, then transform both the training and testing sets.*

(c) **Modeling:** Train a classification model to predict the news category using the vectors generated in step (b). You may use **K-Nearest Neighbors (KNN)** or another standard classifier (e.g., Logistic Regression, Naive Bayes).

1.3 Reporting Requirements

For both vectorization methods (BoW and TF-IDF), report the following:

1. **Dictionary Size:** The total number of unique tokens (features) in your vocabulary derived from the training set.
2. **Top 10 Words:** A list of the 10 most frequent words found in the training corpus.
3. **Model Performance:** Report the classification accuracy on both the **Training Set** and the **Testing Set**.

Question 2: Textual Analysis with Chinese Data

You will now apply the classification techniques from Question 1 to a Chinese-language dataset: `tnews`, which is part of the CLUE benchmark. You can find more details about the dataset structure at the Hugging Face repository: `https://huggingface.co/datasets/clue/clue`.

2.1 Data Preparation

Download the following files from Moodle:

- `tnews_train.parquet`: The training dataset.
- `tnews_test.parquet`: The testing dataset.

The files contain:

- `sentence`: The news title text.
- `label`: The category of the news.

2.2 Methodology

(a) **Preprocessing (Segmentation):** Since Chinese text is not delimited by spaces, use a specialized library (e.g., `jieba`) to segment the continuous strings of characters in the `sentence` column into meaningful tokens. Apply this to both training and testing sets.

- Segment the text.
- Remove stop words (using a Chinese-specific list like Baidu or HIT) and punctuation.

(b) **Vectorization:** Implement two methods to transform the segmented text into numerical feature vectors.

- **Bag of Words (BoW)**
- **TF-IDF**

*Note: Fit your vectorizers using only the **training** set, then transform both the training and testing sets.*

(c) **Modeling:** Train a classification model to predict the news category using the vectors generated in step (b). Use the same classification algorithm chosen in Question 1.

2.3 Reporting Requirements

For both vectorization methods (BoW and TF-IDF), report the following:

1. **Dictionary Size:** The total number of unique Chinese tokens in your vocabulary after segmentation and filtering.
2. **Top 10 Words:** A list of the 10 most frequent Chinese words in the training set.
3. **Model Performance:** Report the classification accuracy on both the **Training Set** and the **Testing Set**.

Technical Note: Ensure your coding environment supports UTF-8 encoding to correctly process and display Chinese characters.