Vocabulary & Term-Document Matrix Analysis Lab Work

Prosanto Mojumdar, Student ID: 200226

Utsa Debnath, Student ID: 200242

Khulna University, Khulna

**Introduction:**

The project involves analyzing text data from a dataset titled **British\_Airway\_Review.csv**. The objective is to:

1. Extracting vocabulary from the dataset.

2. Build a full term-document matrix.

3. Extract the top 100 most frequent terms from the dataset.

4. Generate a partial term-document matrix for the top 100 terms.

The dataset contains customer reviews, and this analysis allows us to understand the frequency distribution of terms used across reviews.

**Dataset Overview:**

**Dataset source** **:** <https://www.kaggle.com/datasets/praveensaik/british-airways-passenger-reviews-2016-2023>

**Dataset Snapshot:**

****

**Number of Samples:**

The dataset consists of 2500 samples (rows), where each sample represents an individual customer review.

**Number of Categories:**

The data used for this project belongs to a single category – customer reviews.

**Total Number of Terms:**

After processing the reviews, the total vocabulary size (unique words) across all reviews is 11982 unique terms.

**Vocabulary and Matrix Overview:**

**Full Term-Document Matrix:** Contains 11982 terms and 2500 samples.

**Partial Term-Document Matrix:** Contains the top 100 most frequent terms.

**Results and Findings:**

**Term Document Matrix:**

Below is the term-document matrix snapshot-

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 0 | 38 | 53 | 55 | 00am | 00o |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Top 100 Terms:**

The following table shows the top 100 most frequent terms extracted from the

vocabulary-

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| flight | ba | verified | service | london |
| british | airways | seat | food | time |
| trip | crew | seats | class | good |
| cabin | heathrow | one | business | staff |
| would | economy | get | check | first |
| review | us | flights | airline | hours |
| back | boarding | plane | even | could |
| passengers | lounge | return | experience | club |
| meal | told | 2 | new | fly |
| hour | aircraft | customer | like | airlines |
| airport | long | premium | well | 3 |
| drinks | really | minutes | two | also |
| got | flying | offered | luggage | better |
| poor | made | served | much | another |
| last | still | way | due | full |
| board | booked | friendly | take | flew |
| bag | grat | 5 | gatwick | via |
| arrieved | nice | however | asked | next |
| comfortable | old | breakfast | never | people |
| pay | late | entertainment | delayed | given |

**Partial Term-Document Matrix for Top 100 Terms:**

Below is the partial term-document matrix for the top 100 terms snapshot

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| flight | ba | verified | service | london | british | airways | seat | food |
| 5 | 6 | 1 | 0 | 2 | 0 | 0 | 1 | 2 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 1 | 2 | 1 | 1 | 1 | 0 | 0 |
| 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 8 | 5 | 1 | 0 | 5 | 2 | 2 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 1 | 2 | 0 | 1 | 1 | 2 | 1 |
| 12 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 2 | 1 | 1 | 1 | 2 | 2 | 0 | 0 |
| 4 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 2 |
| 1 | 3 | 1 | 2 | 0 | 0 | 0 | 0 | 1 |
| 3 | 5 | 1 | 3 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 2 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 2 | 1 | 3 | 0 | 2 | 2 | 1 | 2 |

**Conclusion:**

In this project, we processed a text dataset to generate a term-document matrix that captures the frequency of words across customer reviews. From this matrix, we derived a partial matrix with the top 100 most frequent terms to facilitate more focused analysis. The results provide insights into the most commonly used words, potentially offering value in understanding customer sentiment and identifying key discussion points.

**Files Generated:**

term\_document\_matrix.csv: The full term-document matrix.

term\_document\_matrix\_top\_100.csv: The partial term-document matrix with the top 100 terms.