Text Preprocessing

Week 3 - Code Practice

Professor: Misuk Kim

Teaching Assistant: Minjoo Son

minjoo77@hanyang.ac.kr



Text Preprocessing Week 3 – Code Practice

Contents

1. Introduction

2. Text Preprocessing Process

3. Visualization

4. Assignment

4. Conclusions

1. Introduction

Overview

- Teaching Assistant: Minjoo Son
- Email: minjoo77@hanyang.ac.kr
 - Please feel free to ask me anytime if you have any questions!

1. Introduction

Week 4 Objective

- What we've learned in the past lectures?
 - Overview of Unstructured Data Analysis
 - Text Analytics Process
 - Text Preprocessing
- What will we do in today's lectures?
 - Practicing basic text preprocessing methods in code
 - Text Preprocessing Process
 - Visualization
 - Self-Guided Practice

2. Text Preprocessing Process

Text Preprocessing

- Purpose: Prepares text according to the problem's purpose
- Analogy: Just like how a dish turns out poorly if the ingredients are not properly prepared, natural language processing techniques learned later may not work correctly if the text is not properly preprocessed.
- Process:

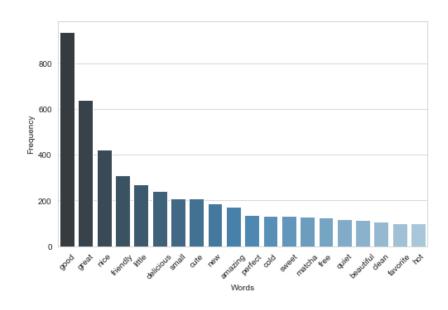


2. Text Preprocessing Process

- Code practice in Colab.
- Link
 - https://github.com/ming9oori/2025-2-Unstructured-Data-Analysis
- File
 - [Week 3] Text Preprocessing/1. Text Preprocessing Process.ipynb

Word Frequency Graph & Word Cloud

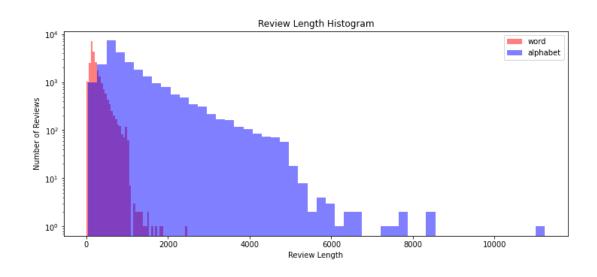
- Word Frequency Graph
 - A tool that visually identifies key words and patterns, helping to remove unnecessary words or select words for learning.
- Word Cloud
 - A visual representation of word frequency, allowing for easy identification of key concepts in the text.

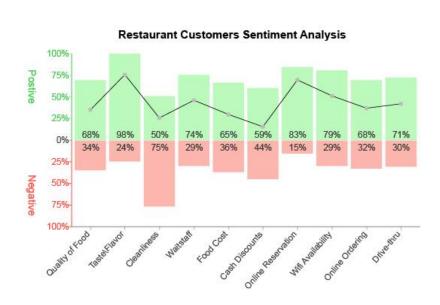




Word Frequency Graph & Word Cloud

- EDA (Exploratory Data Analysis)
 - The process of understanding data and identifying patterns and characteristics.
 - In Natural Language Processing, it is the initial step to understand text data.
 - It involves identifying the data structure, analyzing word frequencies and patterns, and visualizing them with a Word Cloud.
 - Sentiment analysis is used to detect positive or negative trends, while data distribution analysis captures overall characteristics.





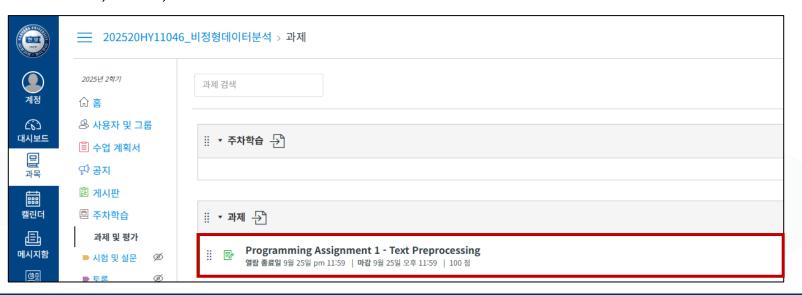
2. Visualization

- Code practice in Colab.
- Link
 - https://github.com/ming9oori/2025-2-Unstructured-Data-Analysis
- File
 - [Week 3] Text Preprocessing/2. Visualization.ipynb

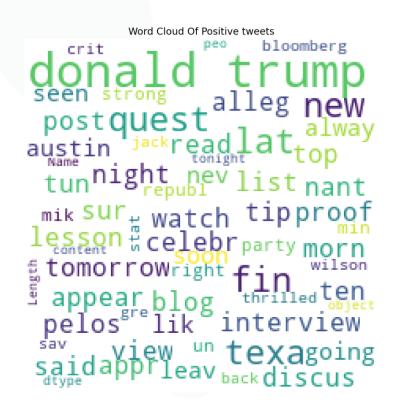
- Text Preprocessing & EDA
 - Task
 - Preprocess the Twitter Sentiment Analysis dataset, then create visualizations like a Word Graph or Word Cloud to represent each sentiment (Positive, Negative, Neutral, exclude Irrelevant).
 - Dataset
 - This is an entity-level sentiment analysis dataset of twitter.
 - Entities: Borderlands, Microsoft, TomClancysRainbowSix, MaddenNFL, LeagueOfLegends, CallOfDuty
 - There are three classes in this dataset: Positive, Negative and Neutral. We regard messages that are not relevant to the entity (i.e. Irrelevant) as Neutral.
 - Data (Uploaded on Google Drive): Unstructured Data Analysis > [Week 4] Text Preprocessing > Data > twitter_training.csv
 - Process
 - Load the data
 - Text preprocessing
 - Text Visualization

3. Assignment

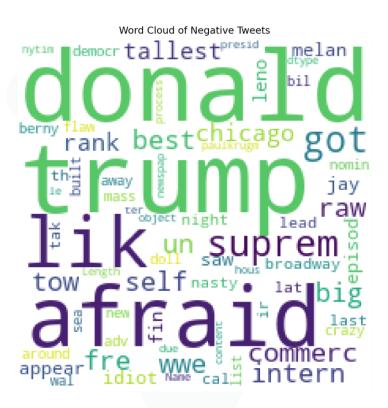
- Text Preprocessing & EDA
 - Deliverables
 - Jupyter Notebook file: Assignment_1_YourName_YourStudentID.ipynb ('Assignment_1_Minjoo Son_2024229436.ipynb')
 - Submit to: "Programming Assignment 1 Text Preprocessing" under Assignments
 - Deadline
 - September 25, 2025, 11:59 PM



- Code Practice
 - Text Preprocessing & EDA
 - Example Results







- There is a reference code on Colab, which you can use as a guide to complete the task.
- Link
 - https://github.com/ming9oori/2025-2-Unstructured-Data-Analysis
- File
 - 2025–2-Unstructured-Data-Analysis/[Week 3] Text Preprocessing/Text Preprocessing sample.ipynb

4. Conclusions

- Request for Feedback to Improve Future Classes
 - The following is a survey for today's code practice class.
 - Please provide feedback on today's class and share your preferred format for future sessions.
 - The survey is anonymous, so feel free to express your thoughts openly.
 - Your feedback will be used to improve the code practice class in future sessions.

Q & A

Thank you for your attention. Any questions are welcome!

Minjoo Son

