2024-2 Unstructured Data Analysis
2024.09.23

Text Preprocessing

Week 4 - Code Practice

Professor: Misuk Kim
Teaching Assistant: Minjoo Son
minjoo77@hanyang.ac.kr

HANYANG UNIVERSITY

Contents

1. Introduction

2. Text Preprocessing Process

3. Visualization

4. Self-Guided Practice

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

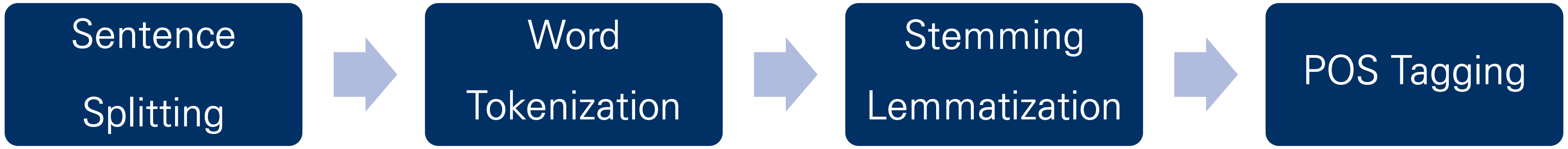
- What we've learned in the past lectures?
  - Overview of Unstructured Data Analysis
  - Text Analytics Process
  - Text Preprocessing

- 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:



Sentence Splitting → Word Tokenization → Stemming Lemmatization → POS Tagging

2. Text Preprocessing Process

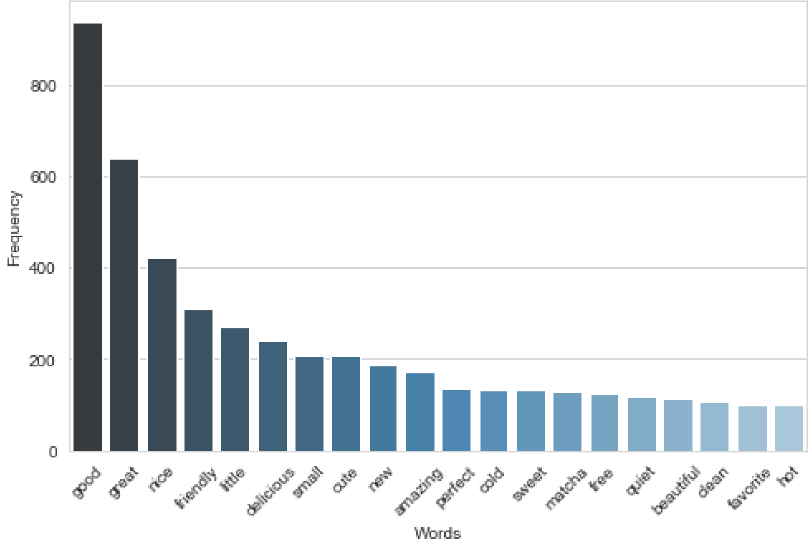
❖ Code Practice

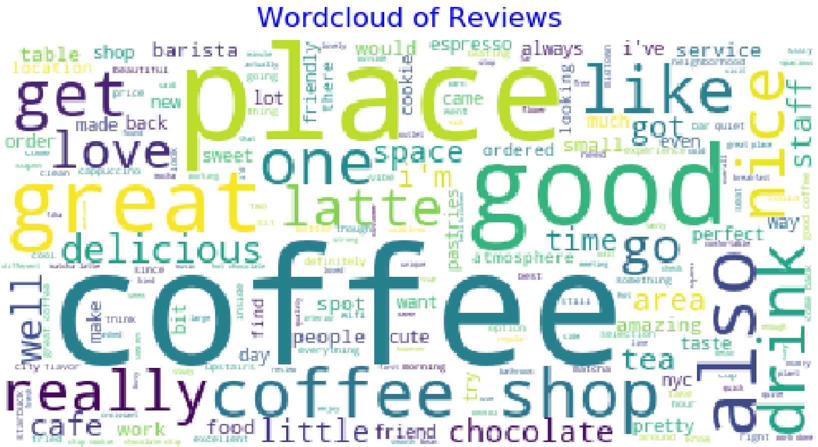
- Code practice in Colab.
- Link
  - https://drive.google.com/drive/folders/1astmYfYGZKk3mfoiPMb5muRbfCOTo4hI?usp=drive_link
- File
  - Unstructured Data Analysis > [Week 4] Text Preprocessing > 1. Text Preprocessing Process.ipynb

2. Visualization

❖ 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.



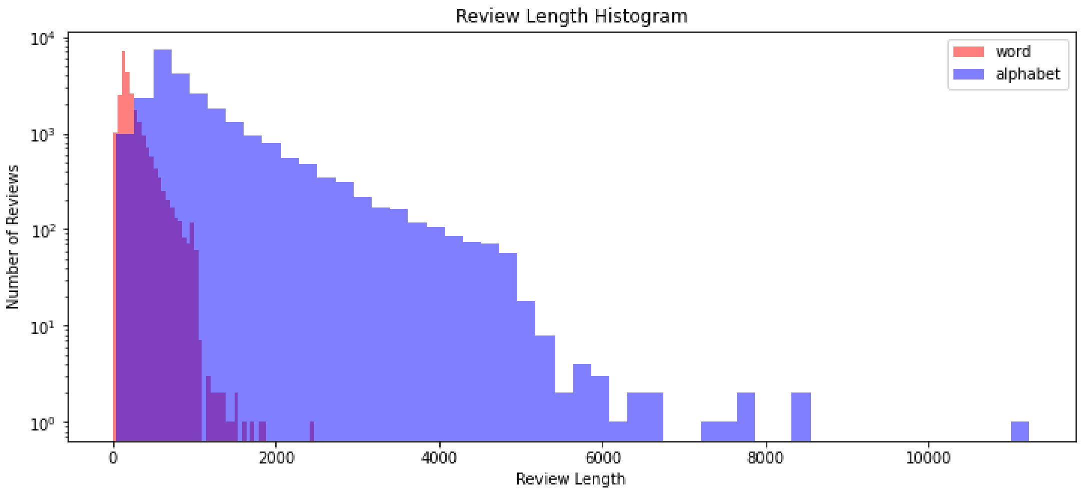


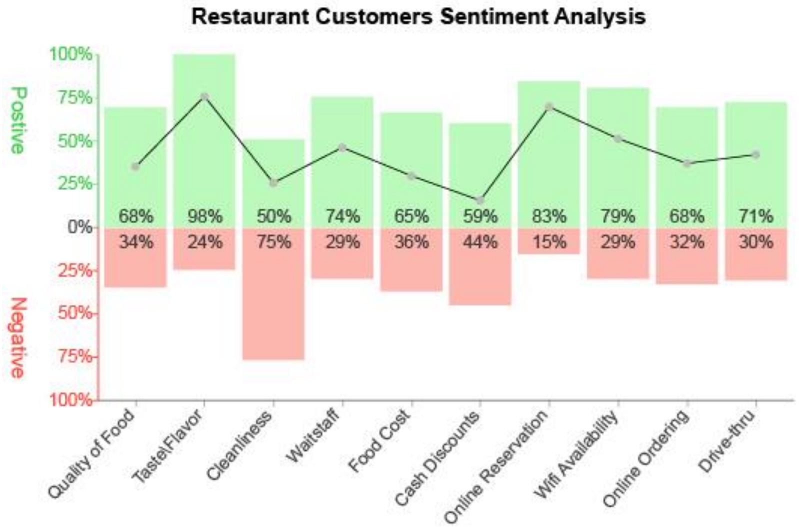
Source: https://github.com/skajonpr/Natural-Language-Processing-and-EDA

2. Visualization

❖ 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.





Source: https://buly.kr/CWsg0Fj / https://chartexpo.com/charts/sentiment-analysis

2. Visualization

❖ Code Practice

- Code practice in Colab.
- Link
  - https://drive.google.com/drive/folders/1astmYfYGZKk3mfoiPMb5muRbfCOTo4hI?usp=drive_link
- File
  - Unstructured Data Analysis > [Week 4] Text Preprocessing > 2. Visualization.ipynb

3. Self-Guided Practice

❖ Code Practice

- Text Preprocessing & EDA
  - – Task
    - ▪ Preprocess the Twitter Sentiment Analysis dataset, then create visualizations like a Word Graph and 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

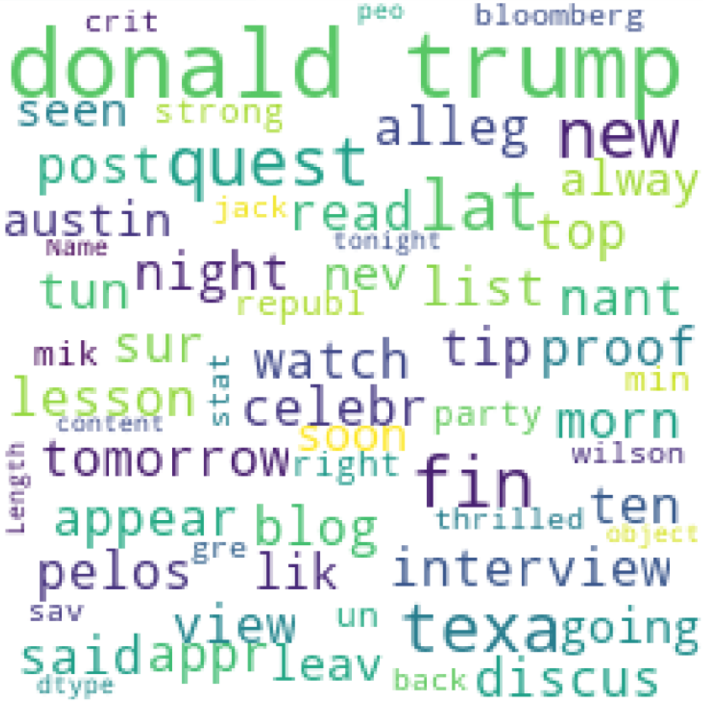
Source: https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis/data

3. Self-Guided Practice

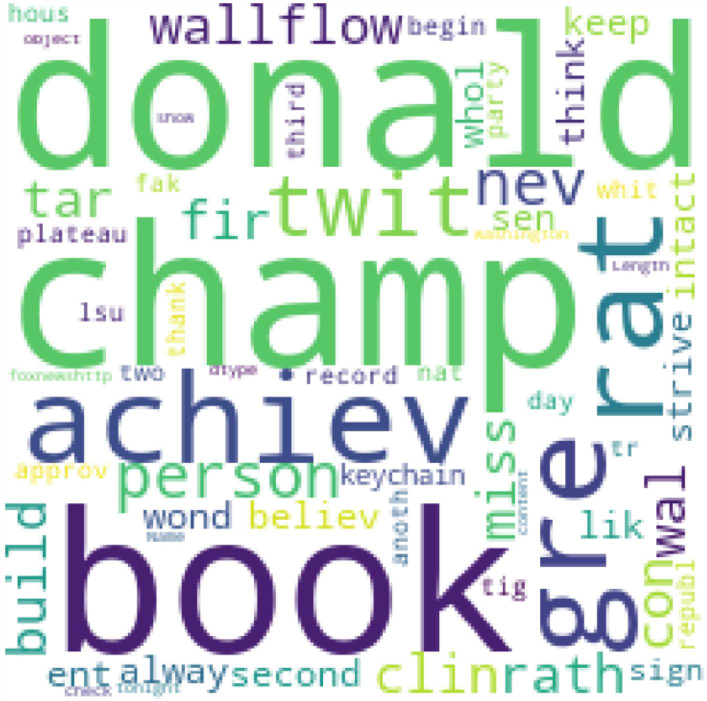
❖ Code Practice

- Text Preprocessing & EDA
  - – Example Results

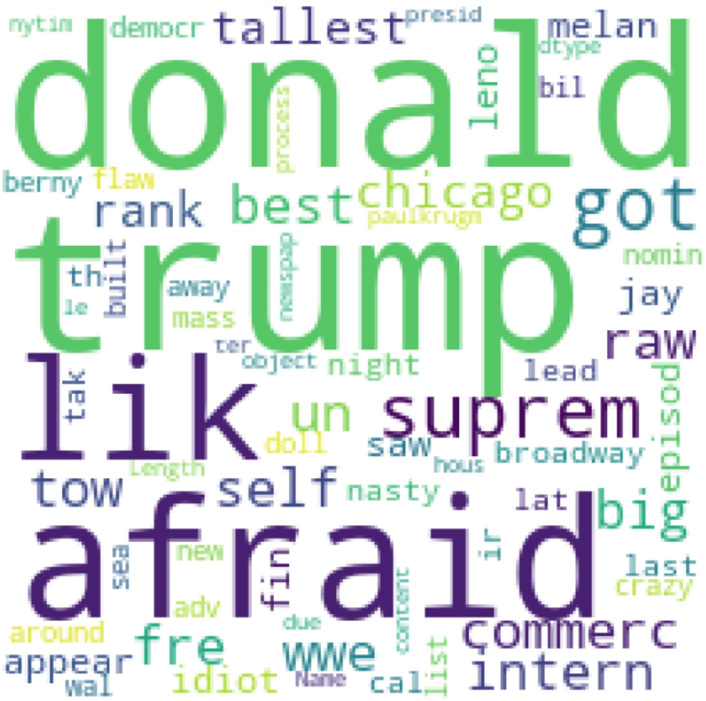
Word Cloud Of Positive tweets



WordCloud of Neutral Tweets



Word Cloud of Negative Tweets



Source: https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis/data

3. Self-Guided Practice

❖ Code Practice

- There is a reference code on Colab, which you can use as a guide to complete the task.
  - Link
    - https://drive.google.com/drive/folders/1astmYfYGZKk3mfoiPMb5muRbfCOTo4hl?usp=drive_link

Source: https://www.kaggle.com/code/sudalairajkumar/getting-started-with-text-preprocessing

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.

Text Preprocessing
Week 4 – Code Practice

2024-2 Unstructured Data Analysis
2024.09.23

Q & A

Thank you for your attention. Any questions are welcome!

Minjoo Son

HANYANG UNIVERSITY