**Assignment 6**

**1 Text Preprocessing**

**1.1 Input and Output**

**1.1.1 Text chosen for text preprocessing**

**Text 1:**

Super Bowl viewers could hardly miss an ad blitz from e-commerce company Temu promising low prices that allow customers to “shop like a billionaire.”

The company’s app vaulted to second place among the most downloaded free apps on Apple devices, Apple rankings showed Monday.

**Text 2:**

Surprise, surprise: SpaceX plans to set more spaceflight records this year.

Elon Musk's company launched 96 orbital missions in 2023, a big jump from its previous high of 61,

which was set a year earlier. And SpaceX is planning another big leap in 2024, one that will take it well above the century mark.

**Text 3:**

With its latest earnings report, Amazon (NASDAQ:AMZN) delivered the news and results that analysts and investors had been waiting to hear, and the company’s stock is responding.

**1.1.2 Text preprocessing output**

**Result of Text 1:**

[['super', 'bowl', 'viewer', 'could', 'hardly', 'miss', 'ad', 'blitz', 'ecommerce', 'company', 'temu', 'promise', 'low', 'price', 'allow', 'customer', '“', 'shop', 'like', 'billionaire', '”', 'company', '’', 'app', 'vault', 'second', 'place', 'among', 'downloaded', 'free', 'apps', 'apple', 'device', 'apple', 'ranking', 'show', 'monday']]

**Result of Text 2:**

[['surprise', 'surprise', 'spacex', 'plan', 'set', 'spaceflight', 'record', 'year'], ['elon', 'musk', 'company', 'launch', 'orbital', 'mission', 'big', 'jump', 'previous', 'high', 'set', 'year', 'earlier'], ['spacex', 'planning', 'another', 'big', 'leap', 'one', 'take', 'well', 'century', 'mark']]

**Result of Text 3:**

[['late', 'earnings', 'report', 'amazon', 'nasdaq', 'amzn', 'deliver', 'news', 'result', 'analyst', 'investor', 'wait', 'hear', 'company', '’', 'stock', 'respond']]

**1.2 Question 1**

**What is the purpose of text preprocessing? Are all text preprocessing steps needed for all analysis tasks? Why or why not?**

Text preprocessing is essentially to create machine-friendly corpus. No matter to use dictionary-based method or take advantage of language models, the computer/program takes a single word as basic units of analysis, i.e. a unique dimension or a factor to understanding the meaning underlying the corpus. Text preprocessing is trying to eliminate the noise made by *human language habits* as much as possible.

To determine whether a preprocessing step needed or not for a specific analysis, we should first determine whether specific kinds of *human language habits* are the noise for our research or potentially valuable information for our research. Segmentation, tokenization and non-word character removal are likely to be useful for all analysis tasks. Case conversion, stop word removal and stemming/lemmatization should be implemented when the research question can theoretically ignore the information eliminated by these operations. Token replacement (including expanding some contractions or abbreviations) is somehow more complicated to my opinions. Researchers should take more consideration about the research content and the characteristics of the texts.

**1.3 Question 2**

**Did the output from the text preprocessing code cell differ from your prediction? How so?**

Some punctuations are failed to be removed, such as single and double quotation marks. It seems that '“' is a Chinese punctuation, which is probably why the **string.punctuation** not recognize it. But I copied the news from English website.

**2 Dictionary Analysis**