**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…thus really strange.

**2 Dictionary Analysis**

**2.1 Question 3: What is the purpose of the dictionary-based analysis? What are some of the limitations of this approach?**

The purpose of the dictionary-based analysis is to generate a reliable dictionary for the content analysis of some specific constructs. The method makes it possible to take a word-counting strategy in measuring constructs through texts.

As for limitations, Reid, McKenny & Short (2023) highlight that “whether language associated with the construct is likely to manifest in organizational texts” and “whether the presence of individual words or short phrases would be indicative of the construct” should be considered seriously. In short, limitations of this approaches concern whether the measurement of a specific construct can be reduced to measuring a bag of related words, i.e., the bag-of-words assumption. For instance, none of the two constructs I proposed in assignment 5, i.e., *the level of free translation of film titles* and *the additional explicitation of actors in the translated titles*, can be measured in this way. Because coding these constructs doesn’t only rely on what the words used mean, but also on how these words are used (e.g., as a who-class noun in the titles) and the relationship of these words with each other (e.g., the similarity between titles).

**2.2 Question 4: Reflect on the dictionary development process you completed as part of part 5 above. What were some of the challenges you encountered? What would you do differently if you were doing this 'for real'?**

I take first 50 samples of film overview (from the [TMDB 5000 data set](https://www.kaggle.com/code/kerneler/starter-tmdb-5000-movies-d190dc26-3)) into [my dictionary development](https://colab.research.google.com/drive/1pUdm0lKIXiZiqWzrBoz4NP7KQQDMfh2O?usp=sharing). For training purpose, I plan to measure ‘masculinity’ content of a film, which I think, (a) can be measured through a bag of words and (b) may have great presence in these film overview because the films are mostly action films.

The main problem in this process is the preprocessing codes perform not well, which seems unable to delete stop words. This led to poorly-generated inductive words, no matter how to set the threshold frequency.

As for a real dictionary, the corpus used for inductive word generation should first be carefully selected or sampled, then well preprocessed including stop word removal and token replacement.

**2.3 Question 5: Take a look at the Excel file created in the dictionary-based analysis code cell at the end of part 5 above. What do these numbers 'mean'? How would you interpret these numbers to someone who is not familiar with the dictionary-based analysis process?**

In the final analysis, *masculinity* represents the number of words recognized by the masculinity dictionary we have developed. *dlnorm\_masculinity* represents the relative presence of these words in respective corpus, which may be explained as an indicative score for how much the story of the focus film related to masculinity.

For people who are not familiar with the dictionary-based analysis process, e.g. reviewers, the interpretation can be:

The overviews of film often reveal critical elements (if not all the elements) of the story, which aim to attract specific audience. Therefore, the content mentioned in these overviews provide a good manifestation for whether a target film involves masculinity elements. Based on this assumption, we then implement a self-developed masculinity dictionary (*Krippendorff's Alpha* = .77) to measure, for each film, how frequently masculinity words would show in its overview text. We divided this frequency by the total numbers of meaningful words of the overview. Meaningful words are the set where words like *is*, *an*, *of*, etc. have been excluded. This operation can strengthen the validity of our measurement. Therefore, the higher the resulted *dlnorm\_masculinity* of a film, the more the story of the film is related to masculinity than other films.

**2.4 Question 6: How might 'transformers' be used to improve upon what is possible with dictionary-based analyses?**

Transformer model would do better in recognize context condition for targeted words. It can be use to increase the validity, in particular when words can be polysemy. For example, the word *toughness* might refer to masculinity only when it is used to mark characteristics of men. It might not be a good masculinity indictor when it is used in a context to describe the task or mission in the films. To distinguish the two conditions may involve complexed regulation finding and setting by coders. However, transformer can learn these easily through finetune process with adequate coded corpus.