**Attn: Dr. Sun Aixin**



**AI6122 Text Data Management and Processing**

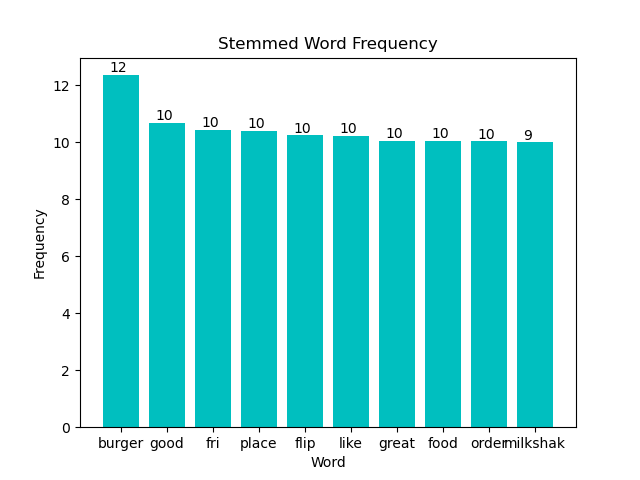
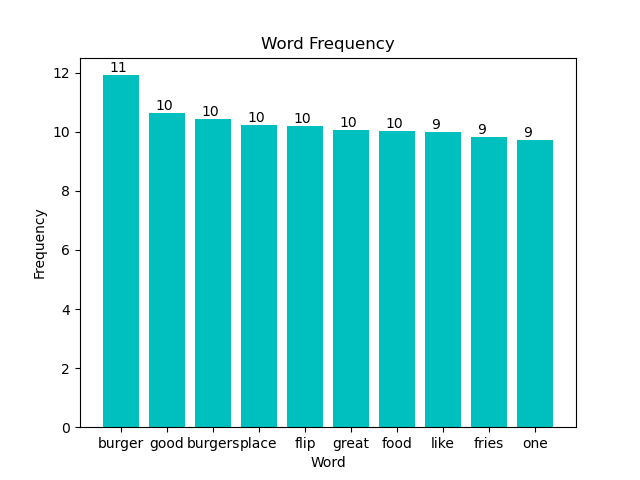
We hereby declare that the attached group assignment has been researched, undertaken, completed and submitted as a collective effort by the group members listed below. We have honored the principles of academic integrity and have upheld Student Code of Academic Conduct in the completion of this work. We understand that if plagiarism is found in the assignment, then lower marks or no marks will be awarded for the assessed work.

Important note: Name must **EXACTLY MATCH** the one printed on your Matriculation Card. Any mismatch leads to **THREE (3)** marks deduction.

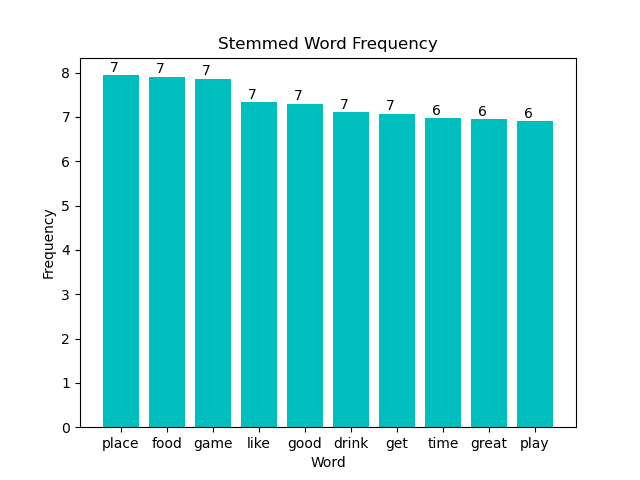
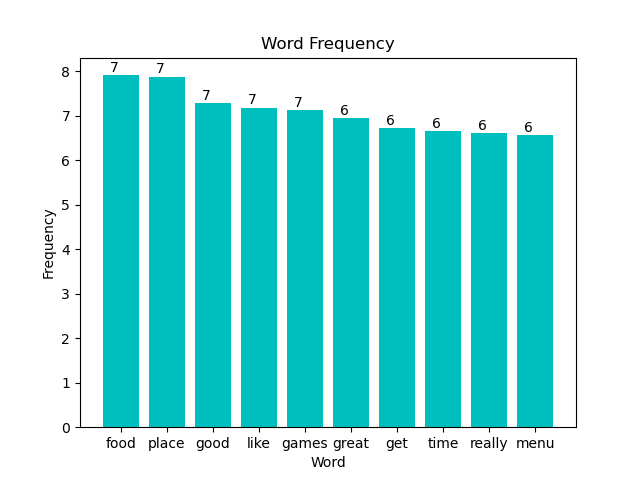
|  |  |
| --- | --- |
| Name | Signature / Date |
| Yipeng Bao | Yipeng Bao/2021/10/21 |
| Kanyu Wang |  |
| Yuting Deng |  |
| Tianyi Peng |  |
| Zhiyuan Zhou |  |

1. **Dataset Analysis**
   1. **Tokenization and Stemming**

Business id b1 = ‘TA1KUSCu8GkWP9w0rmElxw’



Business id b2 = ‘6Hm2FmfLcU\_M91TrZI5htA’



Stemming reduces inflected and derived words to their word stem:

In b1: Taking ‘fri’ in stemmed diagram as an example, the stemming process considers ‘fry’, ‘’fries’ and ‘fried as the same word ‘fri’, that’s the reason why ‘fries’ is missing in stemmed diagram and the frequency of ‘fri’ in stemmed diagram is higher than the frequency of ‘fries’ in non-stemmed diagram. The stem is different from any words.

In b2: Taking ‘like’ in stemmed diagram as an example, the stemming process considers ‘like’, ‘’liking’, ‘liked’ and ‘likes’ as the same word ‘like’, that’s the reason why the frequency of ‘like’ in stemmed diagram is higher than the frequency of ‘like’ in non-stemmed diagram.

* 1. **POS Tagging**
  2. **Writing Style**
  3. **Most frequent Noun - Adjective pairs for each rating**

**Algorithm describtion:**

The basic idea is coming from that adjective are words that modify nouns, if an adjective exists, there must be a noun that it modifies. The reason why our group do not choose a noun to find an adjective is because noun can also be modified by other words like verbs, in case, an incorrect adjective may be paired with the noun. In practice, in a sentence (exclude stopwords) if an adjective is detected, the target noun is its closest noun under the condition that there is no punctuation between them.

However, an adjective can modify many nouns, for instance, ‘Cute dogs, cats’, both ‘Cute, dogs’ and ‘Cute, cats’ should be recognized.

Also, there are negation cases, for example, ‘not good service’, ‘not good, service’ should be recognized instead of ‘good, service’.

Cases like ‘very good service’ is recognized as ‘good, service’ since ‘very’ only modifies the level of ‘good’, which does not change the meaning.

**Limitations:**

1. It does not support negation of a phrase or a sentence:

‘No waiters are nice’ or ‘The waiters are nice? I don’t think so.’

1. It does not support phrases like ‘in general’
2. **Development of a Simple Search Engine**
3. **Extraction of Indicative Adjective Phrases**
4. **Application**