Report: The Analects Corpus Compilation and Exploration

Part 1 Research Questions

1. What is the overall statistical summary of **the Analects corpus**?
2. What is the word frequency list for the corpus?
3. What themes are represented by the high-frequency words in the corpus?
4. What does the dispersion plot reveal about several key terms in the corpus?
5. What significant collocations can be identified in the corpus, and what themes do they represent?
6. Does the corpus conform to Zipf's law distribution?
7. What is the dependency grammar of sentences related to "Junzi," and what meanings do they convey?

Part 2

1. **Web Crawling:** We crawl the webpage at https://ctext.org/analects/ to retrieve 20 raw English texts, which are stored in the file "raw_data."
2. **Data Pre-processing:** We clean **the 20 texts** to compile the Analects corpus by removing noise that does not meet the corpus compilation standards. This includes eliminating excessive punctuation marks, blank lines, spaces, and garbled text. Specifically, we employ regular expressions such as "**\[Show all\]\s([\s\S]+)\sURN**" to filter out noise and "[\u4e00-\u9fa5]" to remove Chinese characters.
3. **Basic Statistics:** Using the NLTK and SpaCy libraries, particularly the "en_core_web_sm" model, we analyze the texts to identify **the top 30 most frequent words,** generate concordances for terms like **"virtue," "benevolence," "ritual," "wisdom," and "harmony,"** and extract **the top 15 significant collocations.** The results are presented below:

Basic Statistics	
Total Words:	**35572**
Vocabulary size	**3096**
Total Sentences:	**1751**
Unique Words:	**2913**
Average Sentence Length:	**20. 32 words**

Frequency List (Top 30…)	
master	**528**
Virtue	**97**
superior	**95**

The most frequent word in the Analects is "Master" (referring to Confucius), appearing 528 times, which suggests the text primarily consists of dialogues between the master and his students, while other frequent words related to morality, such as "virtue" (97 times) and "superior" (95 times), indicate that the main focus is on teaching individuals how to cultivate and maintain good character to become morally upright.

Concordance	
virtue	**132 occurrences**
benevolence	**37 occurrences**
ritual	**31 occurrences**
wisdom	**20 occurrences**

The concordance analysis of the Analects shows that VIRTUE (132) is the central concept, frequently collocating with humanity and propriety, while benevolence (37) and ritual (31) highlight the importance of social relationships

and moral conduct. Consequently, virtue serves as the core of the Analects, with wisdom (20) also emphasizing the need for self-discipline and high standards in society.

Collocation (Top 15…)	
Total Bigrams	**1867**
Unique Bigrams	**291**
Superior man	**621.2614**
Rules propriety	**333.3011**

The most significant collocation in the Analects is **"superior-man" (Junzi),** which embodies the central theme of Confucius' teachings on the virtues necessary to become a gentleman, including benevolence, righteousness, propriety, wisdom, and faithfulness. Another key collocation, **"rules-propriety,"** highlights Confucius' emphasis on individual etiquette and social norms, serving as essential principles for both personal interactions and governance, while other notable collocations include names of individuals and chapter titles.

4. **Stanza Pipeline:** We utilize the Stanza pipeline to process the corpus, performing tokenization, lemmatization, part-of-speech tagging, and dependency parsing on all 20 texts. The results are documented in the file "stanza_pipeline_The_Analects." Specifically, we selected four sentences from "Xue Er," extracted their data, and saved it as an XLSX file named "xue_er_4_sentences_nlp_analysis" for further analysis.
5. **Visualization:**
   a) **Zipf's Law Distribution:** We plotted the distribution for the Analects corpus.
   b) **Dispersion Plot:** We visualized the positions of four key words----"virtue," "superior," "government," and "master"----within the corpus.
6. **Dependency Parsing Analysis of Four Sentences Related to "Junzi"**
   a) **Seek (Head 0): The core word and root**
      i. This demonstrates that the gentleman consistently seeks good virtue, which includes not being particular about his diet and living conditions, embracing simplicity, and avoiding extravagance.
   b) **Man : head of "Virtue"**
      i. This indicates that these virtues are inherently linked to individuals; thus, people must strive relentlessly to attain them.
7. **Other Statistical Data**

TTR	**0.0863**
Ratio of content words	**0.4007**
Hapax legomena ratio	**0.0379**
Relative frequency of 'virtue':	**0.0029**
Synsets for "virtue"	
virtue.n.01	**the quality of doing what is right and avoiding what is wrong**
merit.n.01	**any admirable quality or attribute**
virtue.n.03	**morality with respect to sexual relations**
virtue.n.04	**a particular moral excellence**
Synonyms	**"virtue", "virtuousness", "moral excellence"**
Hypernyms	
good.n.02	**moral excellence or admirableness**

https://github.com/mbt1909432/CBS5501_1124.git

You could access our project through this github link.