Ling226/A1: Exploring the Impact of Preprocessing on Text Metrics in Python

I want to leverage my non-computer science knowledge, so have chosen to focus on lexical diversity.

Brown Corpus: Document ‘ca01’

```
Total number of words after preprocessing: 1111
```

```
Overall lexical diversity of the text: 0.5913591359135913
```

```
Average lexical diversity of text sentences: 0.5880733944954128
```

```
Top ten most frequent words: [('said', 24), ('jury', 18), ('county',
15), ('fulton', 14), ('election', 14), ('state', 12), ('city', 9),
('department', 9), ('would', 9), ('resolution', 9)]
```

Document ‘cao1’ from the Brown corpus covers a legal investigation into the primary election In Atlanta, Georgia in the United states. The resulting text metrics reveal a top 10 word list that aligns with expected jargon for a text containing legal and political content.

An overall lexical diversity of 59% and an average sentence lexical diversity of 59% suggest above-average diversity, indicating a good variety of vocabulary but with some repetition (Scott, 2023). In the context of a legal investigation, this repetition might be attributed to the recurrence of specific legal terms, place names and speech. For instance, the top words like 'jury,' 'election,' and 'department' are expected in a law context, ‘fulton’ and ‘county’ account for the place which the court investigation took place and ‘said’ with the most repetitions would indicate discussion and reporting of speech. These results underscore the importance of considering the domain-specific nature of the text when interpreting lexical diversity.

To further analyse the lexical diversity, I ran preprocessing on words that appeared only between 1 and 4 times to see what words are used less frequently and the lexical diversity of these rarer words. The average and overall lexical diversity improved drastically to 71%. This is interesting as only 226 words were removed in this frequency range. The drastic improvement in lexical diversity suggests that the inclusion of less frequent words provides a more comprehensive view of the vocabulary spectrum within the text. The modest number of words removed highlights the influential role that these less common words play in shaping the overall linguistic richness of the content.

The Current: Petrol cars should be banned by 2030.

```
Number of words after preprocessing: 18987
```

```
Overall lexical diversity of the text: 0.22550617933210623
```

```
Average lexical diversity of text sentences: 0.22432575356953993
```

```
Top ten most frequent words: [('cars', 492), ('petrol', 362),
('need', 326), ('think', 274), ('electric', 217), ('better', 202),
('good', 200), ('change', 181), ('would', 176), ('planet', 174)]
```

One question from Te Papa's '*The Current'* dataset centres around the advocacy for banning petrol cars by 2030. Frequent terms include 'cars,' 'petrol,' 'electric,' 'change,' and 'planet,' suggesting a focus on environmental concerns and a transition to electric vehicles. The high frequency of 'need,' 'think,' and 'better' indicates an emphasis on persuasion and opinion formation within the discourse.

Overall lexical diversity of 0.2255 and average sentence lexical diversity of 0.2243 suggest a moderate diversity in the vocabulary used. The prevalence of words like 'electric', 'better' and 'planet' emphasises that a large number of respondents support a shift towards alternative, more sustainable transportation methods.

The Current Topic Question: Nature helps us get through lockdowns

```
Number of words after preprocessing: 13276

Overall lexical diversity of the text: 0.27341162300511895

Average lexical diversity of text sentences: 0.2720888083371092

Top ten most frequent words: [('nature', 621), ('think', 359),
('good', 240), ('would', 212), ('idea', 212), ('people', 190),
('need', 184), ('us', 180), ('great', 160), ('time', 143)]
```

The next question I analysed questions the impact of interacting with nature during lockdowns. Results prominently feature words like 'nature,' 'think,' 'good,' 'idea,' and 'people' in the top 10, highlighting the importance of nature in alleviating challenges posed by lockdowns.

Overall lexical diversity of 0.2734 and an average sentence lexical diversity of 0.2721 suggest a moderately diverse vocabulary. Repeated occurrences of 'nature' and 'need' in the top word list corroborates the emphasis on the role of nature in providing solace during challenging times.

Comparing these results to that of the previous question, which share similar overarching concepts about the importance of nature (petrol car emissions being a risk to nature), the language reflects that responders lean towards a conservationist view on environmental issues.The reason for this could be that the survey itself is likely to be affected by self-selection bias, in that those who go to museums may be more willing to learn and conscious of human effects on the environment.

Gutenberg Library: The Great Gatsby

```
Number of words after preprocessing: 16803
```

```
Overall lexical diversity of the text: 0.28126859454956565
```

```
Average lexical diversity of text sentences: 0.281183051654368
```

```
Top ten most frequent words: [('said', 164), ('one', 130), ('like',
115), ('tom', 115), ('gatsby', 105), ('came', 103), ('daisy', 99),
('little', 91), ('went', 90), ('back', 89)
```

Overall lexical diversity of F. Scott Fitzgerald's 'The Great Gatsby' is considerably low at 28%, even excluding stopwords. This is interesting as this novel is considered as an 'American Classic' by many scholars, including Xiangqi Liu, writing 'Stylistic Analysis of The Great Gatsby from Lexical and Grammatical Category'. Liu notes *'One of the simplest yet most profound reasons The Great Gatsby is considered as an American classic is its use of language.'* (Page 662).

Liu focuses on Fitzgerald's use of adjectives and the implications of repetitive descriptors to paint visuals unique to different characters. One example of this is *'the adjective "bright" seems to imply that [the character] Daisy is bright [Happy]. However, the word "sad" denies this by its meaning; thus the description gives us a suspicious impression.'* (Page 662*).* The most common 10 words do contain adjectives, but not as descriptive as Liu's statement would suggest. Running the text metrics again using the same frequency range as used on Brown ca01, the overall and average lexical diversity both increased to a modest 66%.

The presence of less common words, even within the 1 to 4 frequency range, enhances the overall linguistic texture of the novel. Adjectives in the top 10 in this range include ('intimate', 4), ('uniform', 4),('dignified', 4) and ('likely', 4). Further analysis may be needed to determine why the original diversity is low - but as 'The Great Gatsby' does contain many conversational and narrative passages - ('said', 164), ('like', 115), ('tom', 115), ('gatsby', 105), ('daisy', 99) this indicates that much of this text is dialogue, which tends to be uncomplex.

References

Scott, B. (2023, October 20). What are Lexical Density and Lexical Diversity. Readability Formulas. https://readabilityformulas.com/what-are-lexical-density-and-lexical-diversity/#:~:text=Purpose%3A%20Measures%20the%20variety%20of,repetition%20of%20the%20same%20words.

Liu, X. (2010). Stylistic Analysis of The Great Gatsby from Lexical and Grammatical Category. Journal of Language Teaching and Research, 1(5), 662-667. https://doi.org/10.4304/jltr.1.5.662-667