Assignment 2: Regular Expressions

Sam Zhang 261072449

2024-02-06

Problem 2

Problem 2.1

What is the CQL query for modifiers of Covid (all forms)?

```
[word="(?i)covid(-|\s)?(\d+)?"]
```

Include the snapshot of the 20 most frequent modifiers modifiers (top four are shown above):

Problem 2.2

What is the CQL query for modifiers of covid (all forms)?

```
[tag="JJ"] [word="(?i)covid(-|\s)?(\d+)?"]
```

Include the snapshot of the 20 most frequent modifiers modifiers (top four are shown above):

What is the CQL query of words that are modified by Covid (all forms)?

```
[word="(?i)covid(-|\s)?(\d+)?"] [tag="N.*"]
```

Include the snapshot of those words:

What is the CQL query for words that occur in right coordination with Covid (all forms) (e.g., in Covid-19 , SARS-2002 , and HCoV-NL63, the words iSARS-2002 and HCoV-NL63 are the right conjuncts/coordinates).

```
[tag="N.*" & (word !="(?i)covid(-|\s)?(\d+)?('s)?")] within
[word="(?i)covid(-|\s)?(\d+)?"] ([tag="CC" | word=","][(tag="N.*")]){0,9999}
```

Include the snapshot of those words:

What is the CQL query for verbs that can take Covid (all forms) as subject?

```
[word="(?i)covid(-|\s)?(\d+)?"][]{0,2}[(tag != "VH.* | VB.*") & (tag = "VV.*")]
```

Include the snapshot of verbs that take Covid as subject:

What is the CQL query for verbs that can take Covid (all forms) as object?

```
[(tag ="V.*")&(tag!="VB.*|VH.*")][]{0,2}[word="(?i)covid(-|\s)?(\d+)?"]
```

Include the snapshot of verbs that take Covid as object:

Problem 2.3

Show the collocations sorted according to what you think is the best metric (T-Score, MI, LogDice). Indicate the metric you used.

For this COVID-19 corpus, the LogDice score appears to be the most effective metric for identifying and ranking collocations The ranking from Mutual Information (MI) score is unique but prioritizes rare word combinations (≤ 10 co-occurrences). This characteristic of MI can lead to highlighting less frequent, therefore potentially less relevant collocations in the context of a prevalent and significant term like "COVID."

The T-Score, places emphasis on more common words such as "of" In this specific corpus, these common words provide relatively minimal informational value about the unique linguistic patterns associated with COVID-19.

The LogDice score seems to offers a more balanced approach. It appear to effectively normalize the frequency of word pairs and addresses the biases towards extremely rare or extremely common words. This produced good results, where overly common words and overly scarse words are ranked lower. Among all three rankings, the LogDice score gives more nuanced and contextually relevant ranking of collocations that balances statically significance and content richness well.

Problem 3

Problem 3.1