# Lab Class NLP: Corpus Linguistics

By 28.05.2020 solutions for the following exercises have to be submitted: 1, 2, 3, 4

*Note:* You can find the documents required for this assignment in the `corpus-linguistics-documents.zip` on the Moodle page for this course.

## Exercise 1 : Inter-Annotator Agreement (1+1+1+1 Points)

You conducted an annotation campaign on humor. Each annotator was given 6 Tweets and rated each on a 4-point scale from 1 (lame) to 4 (hilarious). Since you have a fine grasp of the concept yourself, you also annotated each Tweet with the "true" label. The results of your campaign are as follows:

| **Tweet** | **Annotator** | | | | | | **Inference** | | |
|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | Truth | Majority | Mean | Median |
| 1 | 1 | 2 | 1 | 1 | 1 | 1 | | | |
| 2 | 2 | 2 | 4 | 2 | 3 | 2 | | | |
| 3 | 2 | 3 | 1 | 2 | 2 | 2 | | | |
| 4 | 4 | 3 | 4 | 3 | 3 | 4 | | | |
| 5 | 1 | 1 | 1 | 1 | 2 | 1 | | | |
| 6 | 3 | 2 | 4 | 2 | 2 | 2 | | | |
| Accuracy | | | | | | | | | |
| Fleiss Kappa | | | | | | | | | |

(a) Assess the performance of each annotator by calculating the Accuracy of the annotations against the truth.

(b) Assess the agreement of all annotators by calculating the Flei? Kappa.

(c) Infer the final annotation for each Tweet from the vote of the five annotators (A-E) in three different ways:

  (c1) By calculating the majority vote (the vote which occurs most often for each Tweet)

  (c2) By calculating the mean of each vote and rounding appropriately.

  (c3) By calculating the median of the votes.

(d) Assess the performance of the different inference strategies by calculating the accuracy of the inferred annotations agains the truth. Which one would you chose and why?

## Exercise 2 : Zipf Distribution (2+1 Points)

Download the documents `raven.txt` and `gullivers-travels.txt` from the course website.
In the lecture, you we're introduced to Zipf's Law:

$$P(w) = \frac{c}{(r(w))^a} \qquad \Leftrightarrow \qquad P(w) \cdot r(w)^\alpha = c$$

(a) Write a python program to estimate the constant $c$ from both, `raven.txt` and `gullivers-travels.txt` by averaging the computed `c` for each occuring word. You can assume that $\alpha = 1$. Remove all punctuation from the text, separate the tokens by whitespaces and newlines (\n), and lowercase all letters. Submit your program with your solution.

$c_{\text{raven}} =$

$c_{\text{gulliver}} =$

(b) How and why does your estimation of c differ between the given documents and the estimation of english with $c_{\text{AP89}} = 0.1$?

Exercise 3 : Discriptive Statistics (4 Points)

Download the documents `raven.txt` and `gullivers-travels.txt` from the course website. Write a python program to compute the descriptive statistics about both documents listed in the following table.

**Hint:** For simplicity, remove all punctuation from the text, separate the tokens by whitespaces and newlines (\n), and lowercase all letters. For example

```
Quoth the raven "Nevermore".
```

would be tokenized as

```
["quoth", "the", "raven", "nevermore"]
```

| | raven.txt | gullivers-travels.txt |
|---|---|---|
| Number of tokens | | |
| Vocabulary size | | |
| Type-token-ratio | | |
| Mean token length | | |
| Entropy | | |
| Top 3 most frequent 1-grams | | |
| Top 3 most frequent 2-grams | | |
| Top 3 most frequent 3-grams | | |

Exercise 4 : Assisted authorship attribution (1+3 Points)

After browsing your collection of classic british literature, you notice that one documents, `disputed.txt`, contains no information about the author. After reading the document, you are tied between the candidate authors *Jane Austen* and *Charlotte Brontë*. You are determined to solve this closed-set authorship attribution problem by comparing the disputed-author document to the known-author documents `austen.txt` and `bronte.txt`.

(a) Recalculate the statistics from exercise 3 for the three documents. You can increase the length and number of n-grams. Name the author you suspect has written `disputed.txt` and briefly explain why.

(b) A simple, statistical measure used for authorship attribution is Burrows Delta. It is defined as the Manhattan distance between the normalized, relative frequency differences of the most common words:

$$\Delta_B = \sum_{n=1}^{N} |z(D_1) - z(D_2)|,$$

where $z(D)$ is an $N$-length vector for the document $D$ defined as:

$$z_i(D) = \frac{f_i(D) - \mu_i}{\sigma_i},$$

where $f_i(D)$ is the relative frequency of the word $i$ in the document $D$, $\mu_i$ is the mean relative frequency of the word $i$ in all documents, and $\sigma_i$ is the standard deviation of the relative frequency of the word $i$ over all documents.

(b1) Calculate the common vocabulary for the three documents and determine the z-scores. Plot the z-scores for the 50 most frequent words for each document in one bar chart. Name the author you suspect has written `disputed.txt` and briefly explain why.

(b2) Calculate $\Delta_B$ for each pair between the three documents. Name the author you suspect has written `disputed.txt` and briefly explain why.

$\Delta_B(bronte, austen) =$

$\Delta_B(bronte, disputed) =$

$\Delta_B(austen, disputed) =$