Lab Class NLP:II

By May 26, 2025, solutions for the following exercises have to be submitted: 1, 2, 4, 3

Exercise 1 : Zipf Distribution

The lecture introduced Zipf's Law as a statistical law that describes the frequency of words in a language.

(a) What does Zipf's law state?

(b) What implications does it have for developing statistical models in NLP?

(c) Consider the following table of words that appeared in a corpus:

Rank	Word	Frequency
1	the	36000
2	that	18000
3	for	12000
4	is	9000
5	said	7200
6	on	6000
8	in	4500
9	it	4000
10	by	3600
12	from	3000
15	million	2400
16	at	2250
18	as	2000
20	with	1800
24	a	1500
25	was	1440

Assuming that there are a total of 500,000 words out of which 50,000 are unique, do the word frequencies satisfy Zipf's law in this case? Explain why or why not.

Exercise 2 : Data Acquisition

You decide to study the problem of computational argumentation. Name two possible sources of argumentative text like arguments, debates, or claims as text or speech transcripts. For each, describe how you would collect this data and what problems you could face collecting them. Inform yourself about common sources of bias in language corpora or datasets. Name the biases you found and how they can be mitigated during data acquisition.

Exercise 3 : Corpora and Annotation

In this exercise, you will explore the challenges involved in annotating, both for humans and automated systems. Consider the following corpus:

1. Paris Hilton stayed at the Hilton in Paris.
2. Quentin Bloody Tarantino.
3. Max works for Berlin & Brandenburg Post.

Identify and annotate all named entities in the sentences above. For this task, a named entity is any word or phrase that refers to a person (PER), location (LOC), or organisation (ORG). Use the following format for your annotations: [Mary Smith]$_{\text{PER}}$ went to [Tokyo]$_{\text{LOC}}$.

Discuss any instances you found challenging to annotate. Which cases might be particularly difficult for an automated system to annotate? Why?

> To solve the following task, please make yourself familiar with Cohen's Kappa, Fleiss' Kappa, and Krippendorff's Alpha first.

Exercise 4 : Inter-Annotator Agreement

Fleiss' Kappa (κ) is often used in inter-annotator reliability studies as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where:

$$p_e = \sum_{c \in C} P(c)^2, \qquad P(c) = \frac{1}{m \cdot n} \sum_{i=1}^{n} m_{i,c}$$

$$p_o = \frac{1}{n} \sum_{i=1}^{n} \text{agr}_i, \qquad \text{agr}_i = \frac{1}{m(m-1)} \sum_{c \in C} m_{i,c}(m_{i,c} - 1)$$

$c \in C$ – class
n – number of examples;
m – number annotators;
$m_{i,c}$ – number of annotators that rated example i with class $c \in C$;

(a) What is measured by p_o and p_e in the equation above?

(b) In the following table, each cell lists the number of times class c was assigned to example i. The values for agr$_i$ and $P(c)$ are already provided to help you with the calculations:

Example	Class 1	Class 2	Class 3	Class 4	agr_i
1	4	1	0	0	0.6
2	0	3	1	1	0.3
3	1	3	1	0	0.3
4	0	0	3	2	0.4
5	4	1	0	0	0.6
6	0	3	1	1	0.3
Total	9	11	6	4	–
$P(k)$	0.3	0.37	0.2	0.13	–

(b1) Calculate the inter-annotator agreement κ using the provided table.

(b2) Interpret the κ value you obtained. What does it mean in terms of agreement between the annotators?

(c) Briefly explain the main differences between Fleiss' κ, Cohen's κ, and Krippendorff's α in terms of their use cases and applicability.