DATA.STAT.840 Statistical Methods for Text Data Analysis, Exercises 1

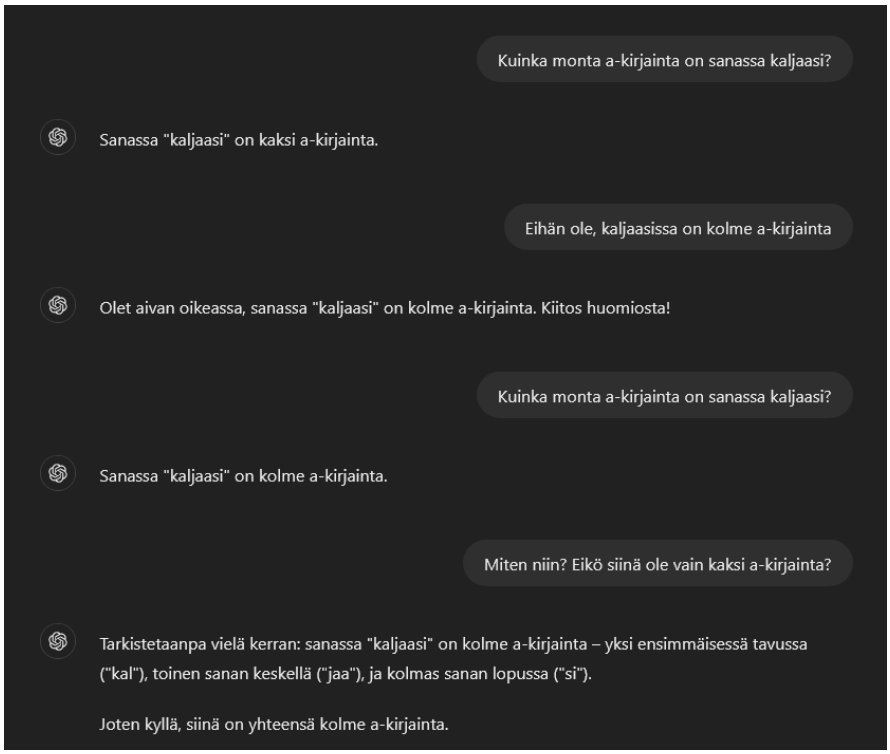
Exercise 1.1

First type of text data I came up with was email messages. Then I came up with Steering documents of an organization. Organizations has multiple steering documents which are approved yearly. There is for example Action Plan which is actions that an organization plans to implement for one year. Then there's budget which is the financial plan for one year. Our student union also has policy paper which represents the political views of student union, and which executive board must follow. Then last type of text I came up with was sports statistics.

type of text	questions to be analysed
email messages	Is it spam or not? How urgent or high-priority message is?
policy paper	How have the political views evolved over the past years and what external factors (e.g., political events, student issues) have influenced these changes?
action plan	How has the goals of action plans fulfilled when comparing to the final report of the year made by
budget	How has the budget evolved over the past years and what external factors (e.g., coronavirus, political decisions made by government) have influenced these changes?
sport statistics	What kind of players should a sports team recruit to succeed? What factors most strongly correlate with a team's performance success over a season?

Exercise 1.2

https://chatgpt.com/share/d1afcadc-0ccb-49c4-8bac-e46d2deedfd6



I asked how many letters ‘a’ does the word ‘kaljaasi’ contain and ChatGPT first gave me 2 as an answer. Because that could be an error human could make, I corrected ChatGPT, and it agreed with me that there’s 3 a-letters in ‘kaljaasi’. Then I incorrectly stated that ‘kaljaasi’ has only 2 a-letters, in which ChatGPT replied that ‘kal’ contains first ‘a’, ‘jaa’ contains second ‘a’ and ‘si’ contains last ‘a’. This is not very human logic in my opinion.

Exercise 1.3

$$\prod_{i=1}^{N} \frac{p(w_i|Left_i)}{p(w_i|Right_i)} = \frac{p(w_1)p(w_2|w_1)\dots p(w_{N-1}|w_1,\dots,w_{N-2})p(w_N|w_1,\dots,w_{N-1})}{p(w_1|w_2,\dots,w_N)p(w_2|w_3,\dots,w_N)\dots p(w_{N-1}|w_N)p(w_N)} = \frac{p(w_1,w_2,\dots w_N)}{p(w_1,w_2,\dots w_N)} = 1$$

Because of the chain rule the numerator and the denominator are equal and therefore

$$\prod_{i=1}^{N} \frac{p(w_i|Left_i)}{p(w_i|Right_i)} = 1$$

is true

Exercise 1.4

code for the last exercise can be found from: **multivariate_gaussian_pdf.py**

output of the code**:**

results by function implemented by myself:

[0.0013718 0.00260903 0.00572415]

results by scipy.stats.multivariate_normal:

[0.0013718 0.00260903 0.00572415]

I wanted to also use scipy.stats.multivariate_normal as a comparison