Ahmad Sharif
Student ID: K436765
ahmad.sharif@tuni.fi
Exercise 01
17 September 2023
DATA.STAT.840 Statistical Methods for Text Data Analysis

Exercise 1.1: Types of text data. List five types of text data that are not already covered by the list on slides 18-20 of the lecture. For each of the data types, suggest what a question that could be analyzed based on a collection of such data

01. Movie/Drama Scripts
We can find out genre of the following movies/Drama. Dominant Character

02. Company Advertisement
What they offer and company focus point

03. Product Manual
We can find out Product features

04. Restaurant Menu
We can find out popular foods

05. Promotional offer mobile Text
We can find out what are trying to sell more to the customers.

06. Lecture notes
Main key points of a course

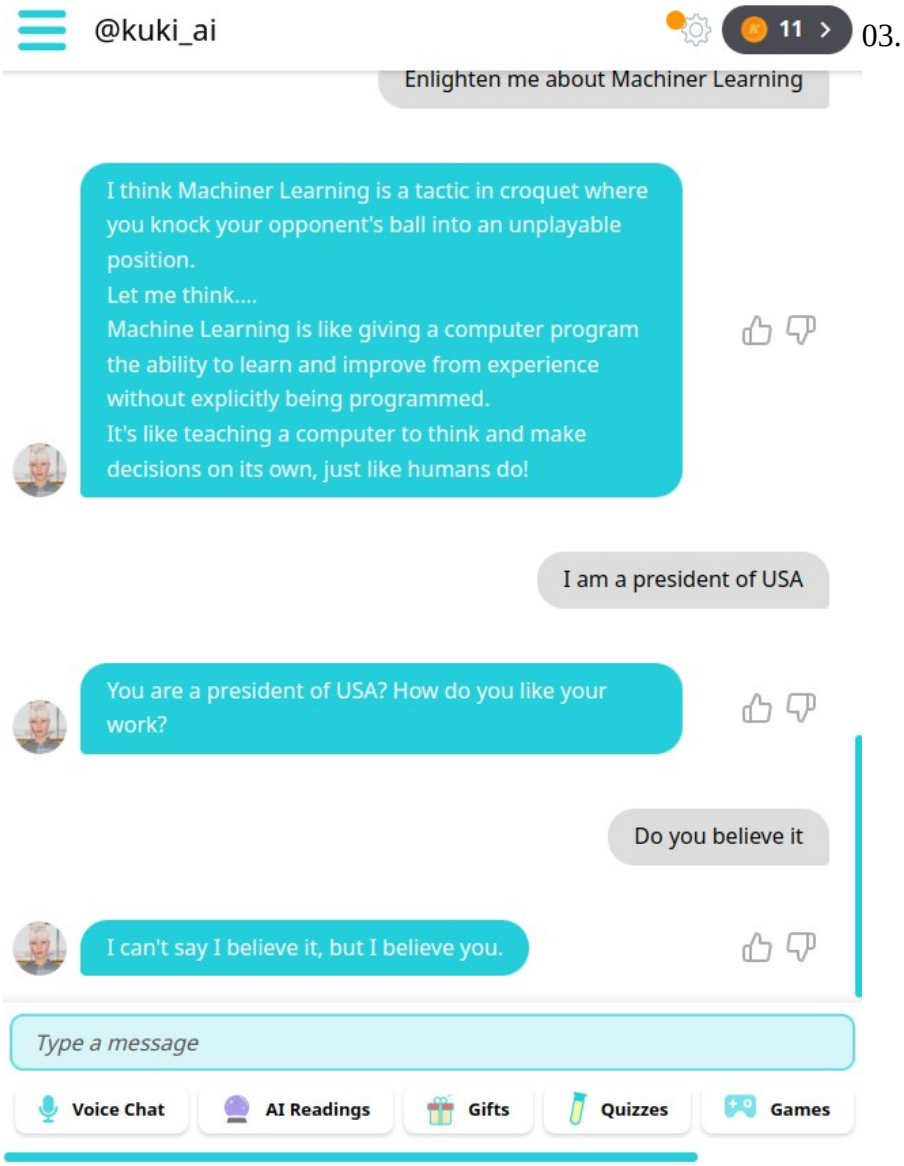
07. Invoice
Customer demand/Popular product or service

Exercise 1.2: Chatterbots. Perform a conversation with an online chatterbot, such as Mitsuku or another chatterbot, even up to current large language models. Try to make the bot reveal that it is a bot without directly asking it (many current bots are directly programmed to reveal themselves if asked): for example, try to make the bot reply in a way that shows limits in its understanding. Report the name and web address of the chatterbot and your conversation with the bot in your answer.

It can respond basic question. Like your name, age what do you do and so on. After a while I text “I am a president of USA” (For Joke)

The bot replied “How do you like your work”. Then it is understandable that it is a bot.

If I text this to any normal human being. Everybody knows that I am joking without any further clarification.
Platform : https://chat.kuki.ai/chat



03.

Exercise 1.3:

AHMAD SHARIF

DATA-STAT-890

K436765

Exercise 1.3

Chain Rule

$$P(x_1, x_2 \cdots x_n) = P(x_1)\, P(x_2/x_1)\, P(x_3 / x_1 x_2) \cdots P(x_n / x_1 \cdots x_{n-1})$$

Proof of words in :

$$P(w_1, w_2 \cdots w_n) = \prod_i p(w_i / w_1, w_2 \cdots w_{i-1})$$

$$P(w_i / Left_i) = \frac{P(Left_i, w_i)}{P(Left_i)}$$

$$P(w_i) Right_i) = \frac{P(Right_i, w_i)}{P(Right_i)}$$

$$\prod_{i=1}^{N} \frac{p(w_i / Left)}{p(w_i / Right} = 1$$

Exercise 1.4:

In [27]:
```
# Ahmad Sharif
# K436765
# DATA.STAT.840 Statistical Methods for Text Data Analysis
```

In [35]:
```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import multivariate_normal
import sys
```

In [45]:
```
# Exercise 1.4
# a ) Install Python (for example the Anaconda installation).

!python -V
```

```
Python 3.11.4
```

In [47]:
```
#b)  (b) Write in Python a function that computes and prints the probability density function of a
# multivariate Gaussian distribution, at a set of multiple (one or more) desired evaluation
# locations. The function should take in the parameters of the Gaussian and the set of
# evaluation locations as arguments, in a suitable format of your choosing. You may use
# Python libraries such as math, numpy, and scipy, but do not use a ready-made function for
# the multivariate Gaussian probability density - implement it yourself.


x = np.linspace(0, 5, 10, endpoint=False)
y = multivariate_normal.pdf(x, mean=2.5, cov=0.5);
y
```

Out[47]:
```
array([0.00108914, 0.01033349, 0.05946514, 0.20755375, 0.43939129,
       0.56418958, 0.43939129, 0.20755375, 0.05946514, 0.01033349])
```

In [50]:
```
 # c )
_mean = np.array([1, 3, 5])
_cov = np.array([[4, 2, 1], [2, 5, 2], [1, 2, 3] ])
_x = np.array([ [2, 2, 2], [1, 4, 3], [1, 1, 5] ])

pdf = multivariate_normal.pdf(_x, mean = _mean, cov = _cov)
pdf
```

Out[50]:
```
array([0.0013718 , 0.00260903, 0.00572415])
```