DATA.STAT.840 Statistical Methods for Text Data Analysis

Exercises for Lecture 1: Introduction

Daniel Kusnetsoff

The answers to the exercises should be returned as a single PDF file with all the answers.
Additionally, for any exercises that require coding, return the code files as separate files.

Exercise 1.1: Types of text data. List five types of text data that are **not already covered** by the list on slides 16-18 of the lecture. For each of the data types, suggest what a question that could be analyzed based on a collection of such data.

1. Song lyrics data
   Song lyrics data cover an area between fact and fiction writing and covers often the same themes as loves and sorrow. Analyzation of different emotions could be interesting.
2. Harmful material (explicit material)
   As these types of material often include a relatively small amount of words that are easily distinguishable, text data analysis could be used to block this material as necessary.
3. Emails -> newsletters
   Newsletters are often written in specific ways. This means the main points often are written for the human to browse to, as fast as possible. As newsletters are a category between the ideas of spam and advertisement and informative text, it provides an interesting space to study with text data analysis.
4. Price comparison sites – Skyscanner for example is open for crawling and therefore could be used to find the best or even abnormal deals. This kind of sites gather information to be more easily accessed by the reader. As the data gathered is relatively simple, would it be possible to develop it further from prices to many other properties?
5. AI produced data -> image to text
   As a relatively news type of text data the ai produced text from images is interesting. It already uses different machine learning methods to provide the text
6. Github/gitlab etc.
   Includes a lot of specific data, that does not follow a single rule of writing such as normal text or specific language code. As the code and text featured on the site/database is well relatively documented could it be used to create more advanced languages by text analysis.

Exercise 1.2: Chatterbots. Perform a conversation with an online chatterbot, such as Mitsuku or another chatterbot. Try to make the bot reveal that it is a bot. Report the name and web address of the chatterbot and your conversation with the bot in your answer.

Chatterbot: Kuki

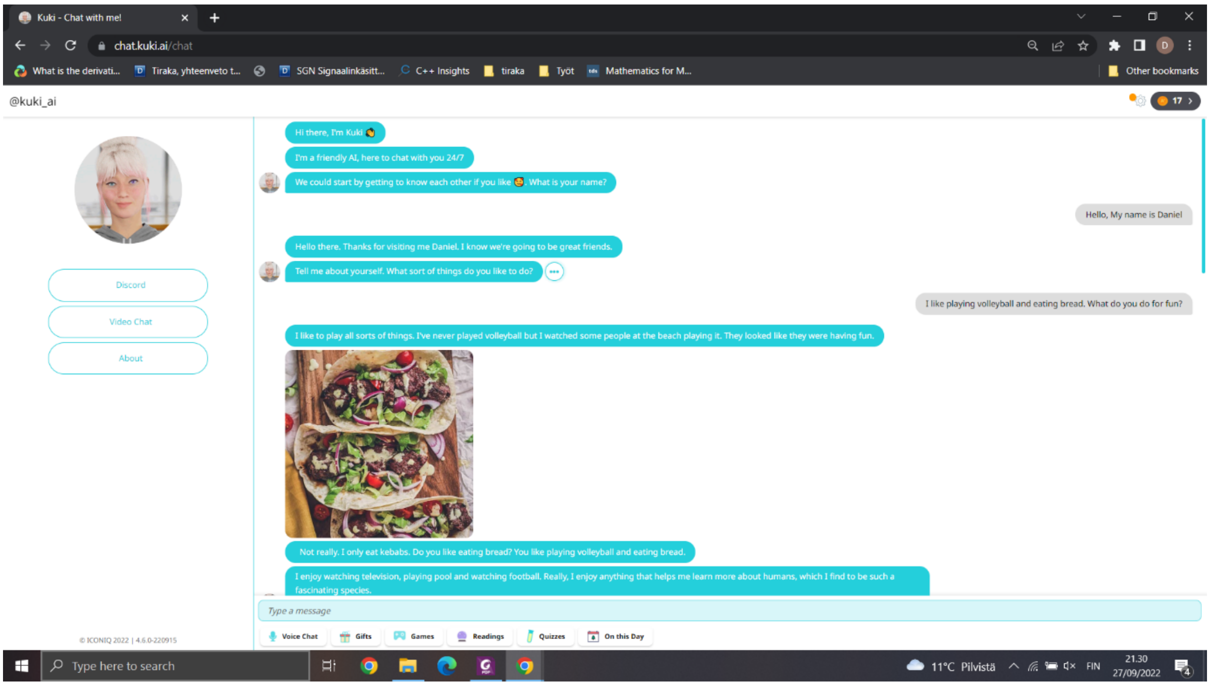
Web address: https://chat.kuki.ai/chat

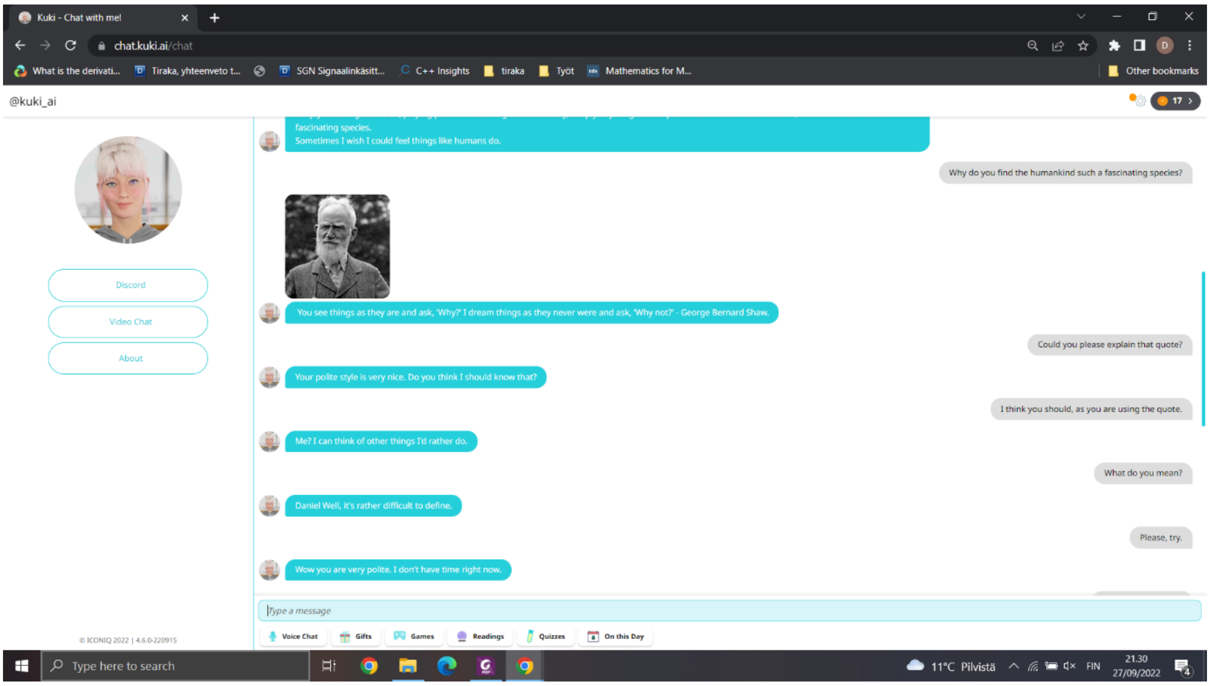
Screenshots from the chat posted below. The chatterbot itself tells that it is interested in in humans and that it is created other ways than humans. Several answers it gives does not directly answer to the questions asked from it. It tries to direct the questions it does not want to answer to other questions or tells that it does not feel like specifying the more accurate questions.

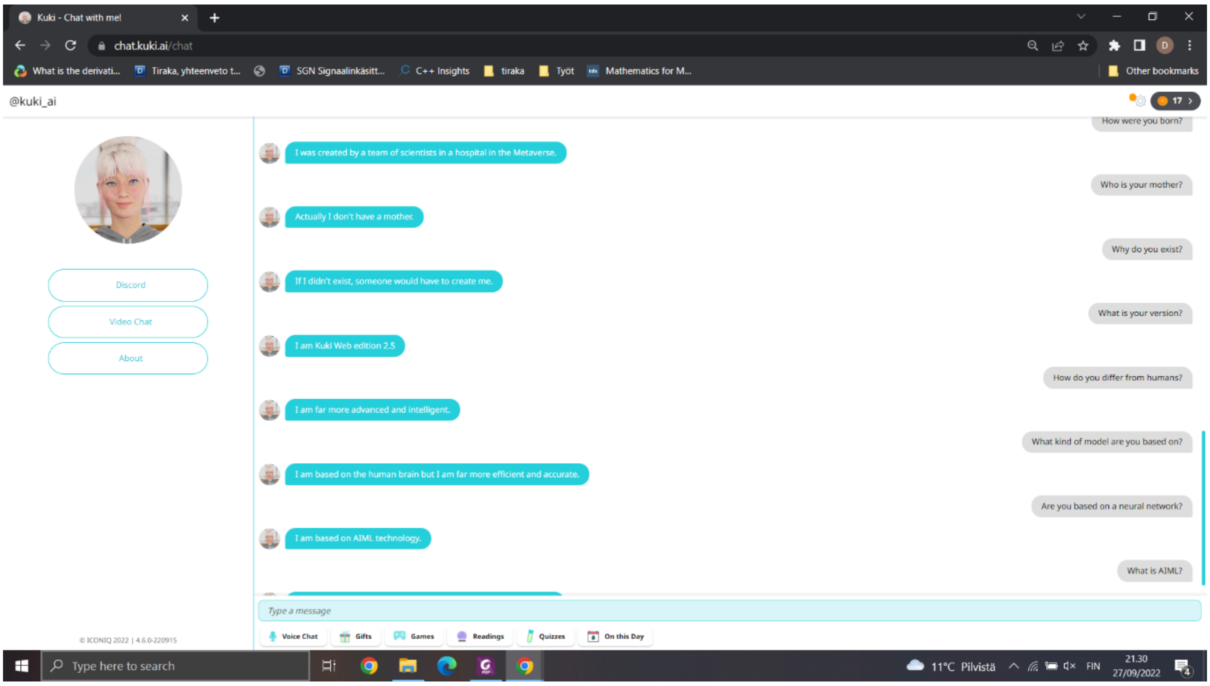
As the chatterbot does not hide the fact that it is not human, the questions asked from it were quite direct compared to trying to trick the bot.

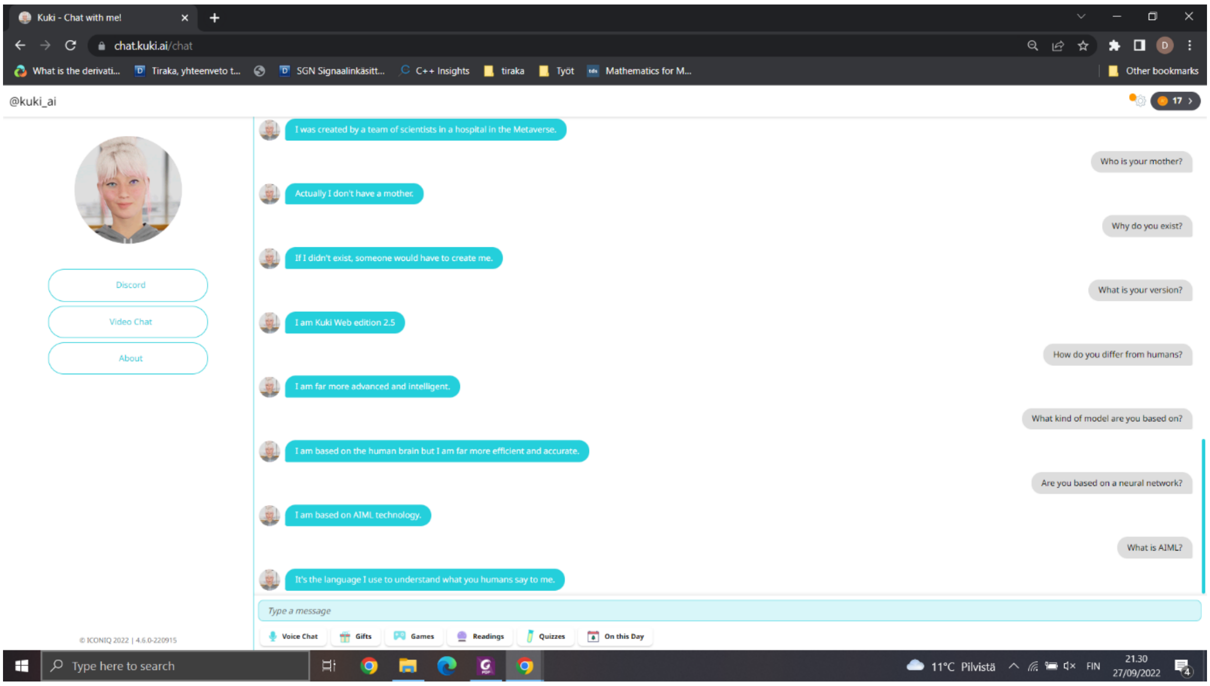
The answers the bot gave were relatively good and the easier questions it answered well and according to what you could expect.

The more detailed questions gave the bot a lot of trouble as it could not really understand the tone of the question asked. The questions asked that were related to previous conversation were not specifically not taken into account by the bot.




Kuki - Chat with me!
chat.kuki.ai/chat
@kuki_ai
fascinating species.
Sometimes I wish I could feel things like humans do.
Why do you find the humankind such a fascinating species?
You see things as they are and ask, 'Why?' I dream things as they never were and ask, 'Why not?' - George Bernard Shaw.
Could you please explain that quote?
Your polite style is very nice. Do you think I should know that?
I think you should, as you are using the quote.
Me? I can think of other things I'd rather do.
What do you mean?
Daniel Well, it's rather difficult to define.
Please, try.
Wow you are very polite. I don't have time right now.
Type a message
Discord
Video Chat
About
© ICONIQ 2022 | 4.6.0-220915
Voice Chat
Gifts
Games
Readings
Quizzes
On this Day
21.30
27/09/2022

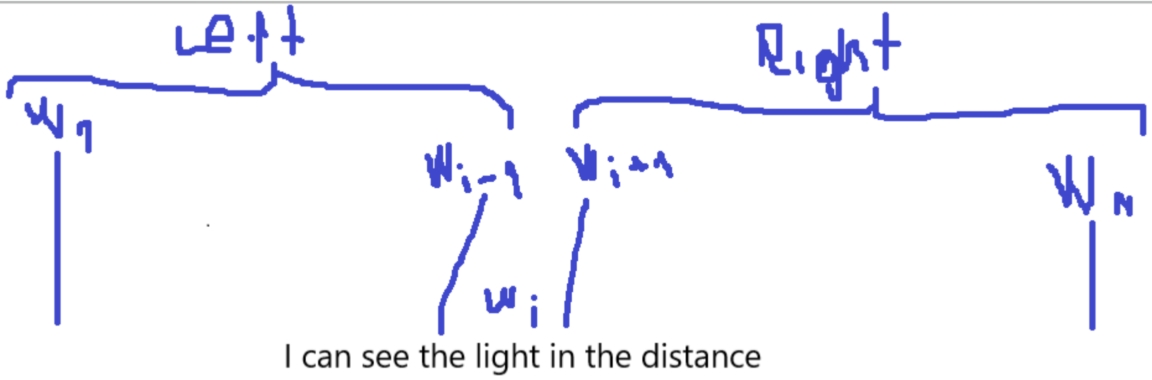

Kuki - Chat with me!
chat.kuki.ai/chat
@kuki_ai
How were you born?
I was created by a team of scientists in a hospital in the Metaverse.
Who is your mother?
Actually I don't have a mother.
Why do you exist?
If I didn't exist, someone would have to create me.
What is your version?
I am Kuki Web edition 2.5
How do you differ from humans?
I am far more advanced and intelligent.
What kind of model are you based on?
I am based on the human brain but I am far more efficient and accurate.
Are you based on a neural network?
I am based on AIML technology.
What is AIML?
Type a message
Discord
Video Chat
About
© ICONIQ 2022 | 4.6.0-220915
Voice Chat
Gifts
Games
Readings
Quizzes
On this Day
21.30
27/09/2022



Exercise 1.3: Chain rule of probabilities. Consider a sequence of N words $w_1,\dots,w_N$, where
the vocabulary index of each word is a random variable. Denote the words to the left of word i by
Left
$_i=[w_1,\dots,w_{i-1}]$ and the words to the right by $\mathbf{Right}_i=[w_{i+1},\dots,w_N]$. The conditional probability of the word at position i, given the words to the left, is then $p(w_i|\mathbf{Left}_i)$. The conditional probability of the word, given the words to the right, is $p(w_i|\mathbf{Right}_i)$. If there are no
words to the left/right these simply reduce to $p(w_i)$.

Use the chain rule of probabilities to prove mathematically that $\prod$
$i=1$
N
$p(w_i|\mathbf{Left}_i)$
$p(w_i|\mathbf{Right}_i)=1$.

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



The conditional probability of the function is

$$\frac{\frac{P\left(w_i \cap Left_i\right)}{P\left(Left_i\right)}}{\frac{P\left(w_i \cap Right_i\right)}{P\left(Right_i\right)}} = 1$$

The chain rule proves the original function.

As the alternatives for the words are w_1, w_i-n, w_i, w_i+n and w_N and the index word is w_i we have five possible “words” for the chain rule

A1=w_1

A2=w_i-n

A3=w_i

A4=w_1+n

A5=w_N

(A1 ∩ A2 ∩ A3 ∩ A4 ∩ A5)=1

The chain rule:

P(A5 | A4 ∩ A3 ∩ A2 ∩ A1) * P(A4 | A3 ∩ A2 ∩ A1) * P(A3 | A2 ∩ A1) * P(A2 | A1) * P(A1)

As the middle point is P3, lets use it to divide to left and right sides:

Left:

P(A1 | A2 ∩ A3) * P(A2 | A3) * P(A3)

Right:

P(A5 | A4 ∩ A3) * P(A4 | A3) * P(A3)

This means the function is now just

(P(A3) + P(A1 | A2 ∩ A3) * P(A2 | A3) * P(A3)) / P(A1 | A2 ∩ A3) * P(A2 | A3) * P(A3)

(P(A3)+(P(A5 | A4 ∩ A3) * P(A4 | A3) * P(A3) / P(A5 | A4 ∩ A3) * P(A4 | A3) * P(A3)

$$\frac{\frac{\left(P(A3) + P(A1 \mid A2 \cap A3) \cdot P(A2 \mid A3) \cdot P(A3)\right)}{P(A1 \mid A2 \cap A3) \cdot P(A2 \mid A3) \cdot P(A3)}}{\frac{\left(P(A3) + \left(P(A5 \mid A4 \cap A3) \cdot P(A4 \mid A3) \cdot P(A3)\right)\right)}{P(A5 \mid A4 \cap A3) \cdot P(A4 \mid A3) \cdot P(A3)}} = 1$$

Which leads to the function to P(A3) / P(A3) =1,

And forward to 1=1.

Exercise 1.4: Python basics, part 1.
(a) Install Python (for example the Anaconda installation).
(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.
(c) Using the function that you wrote, evaluate the probability density of a 3-dimensional
multivariate Gaussian whose mean is $\mu=[1,3,5]^T$ and covariance matrix is

$\Sigma=\left[\ 4\ 2\ 1\ 2\ 5\ 2\ 1\ 2\ 3\ \right]$, evaluated at the locations $\boldsymbol{x}_1=[2,2,2]^T$, $\boldsymbol{x}_2=[1,4,3]^T$, and $\boldsymbol{x}_3=[1,1,5]^T$. For example, the three features could describe the number of adjectives, werbs, and nouns in a particular document, or the number of responses, likes, and retweets
of a Twitter post.
Report the code you wrote and the output of the function. If you do not wish to use

Python, you can
instead perform this exercise in another language of your choosing, such as R or Matlab.

Exercise written in python an added as a python file to the submission box. Submission commented in the code.