

Natural Language Processing Part B

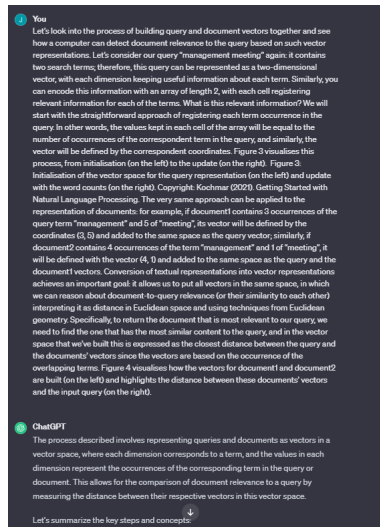
Jamal Akhras

December 2023

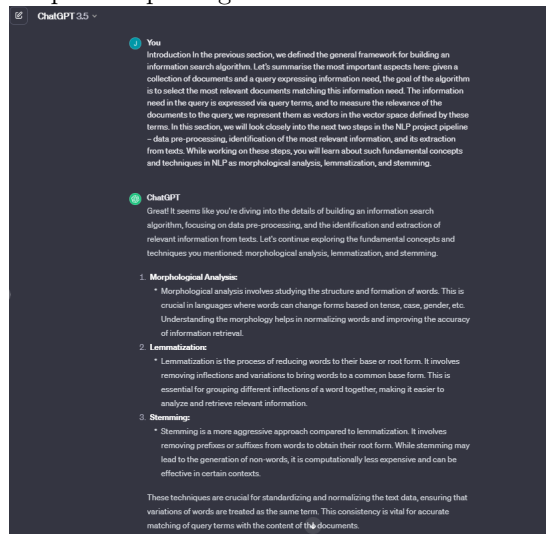
Throughout this report the content of week 2s lectures will be used, lecture transcripts and lecture notes.

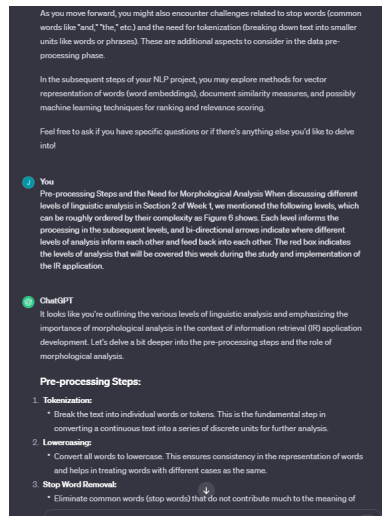
1 Question 1

There are multiple ways to include the content of the lecture notes into chat gpt as context, for shorter texts it can just be pasted in as a whole. For longer bits of text with more context it can be input in parts to ensure it properly parses the text. The simplest method to incorporate mathematical notation is to insert it in markdown notation, this also has been the most reliable way for it.



This is an example of inputting a short subsection of the notes as context.





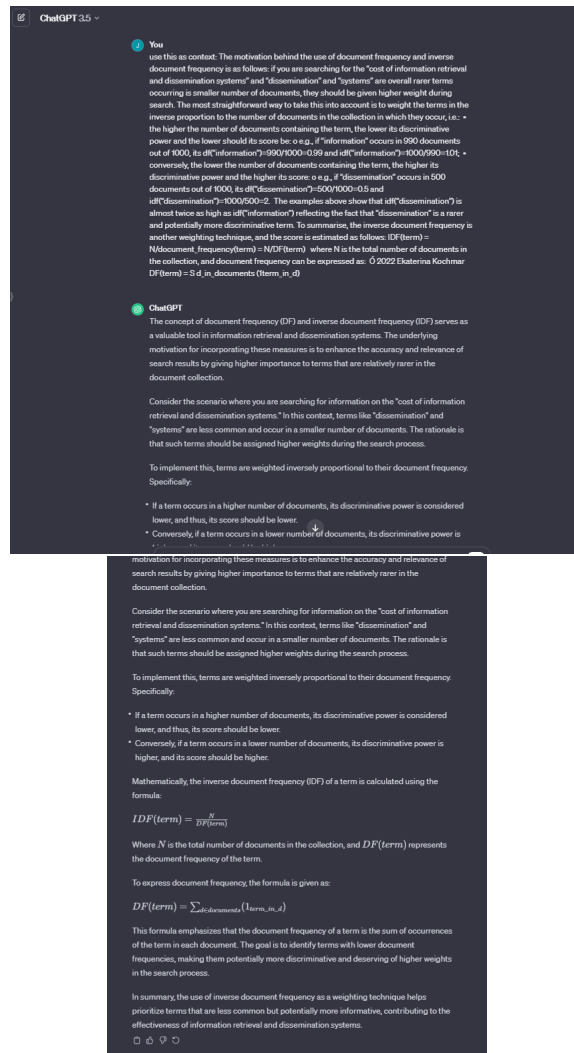
This is another example where a larger input is split into 2 subsections and chatgpt had no issue using it as context

1.1 a

ChatGPT is capable of recognizing the part of the notes that are in textual format, as for improperly formatted mathematical equations it often experiences difficulties understanding them. The textual format handling ability is shown from the previous pictures.

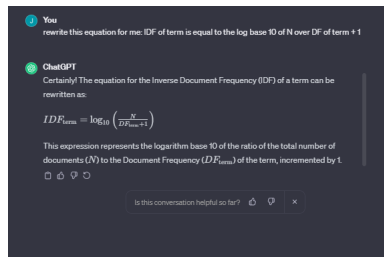
1.2 b

As mentioned previously, improperly formatted maths equations are not understood correctly. However, of the equations found in the lecture notes it seemed to understand them correctly. The TFIDF explanation was used for this section.



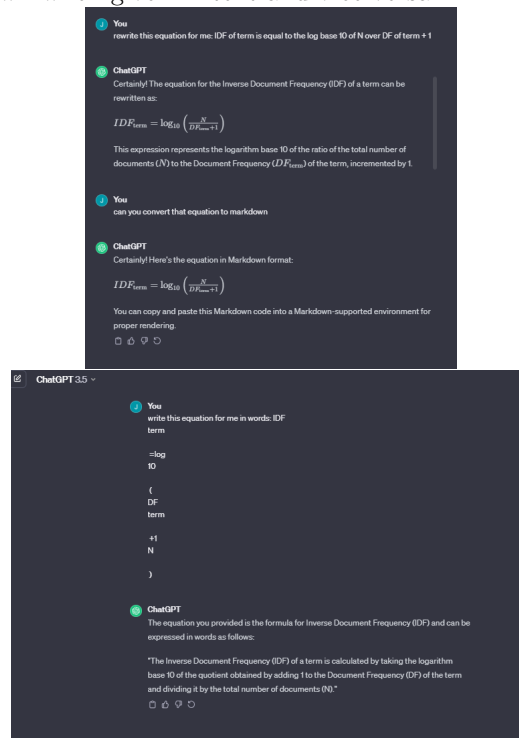
1.3 C

The test was carried out using the IDF equation, which when described to chatgpt it was able to recreate and describe it accurately, as shown from the screenshot below.



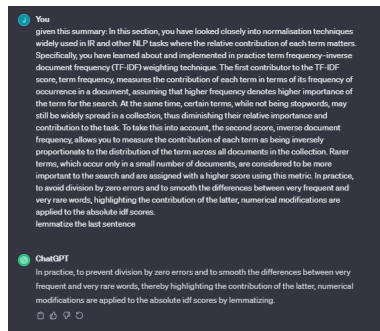
1.4 D

Chatgpt is proving to be robust with equations, as it can accurately recreate them in markdown when given in text and vice versa.



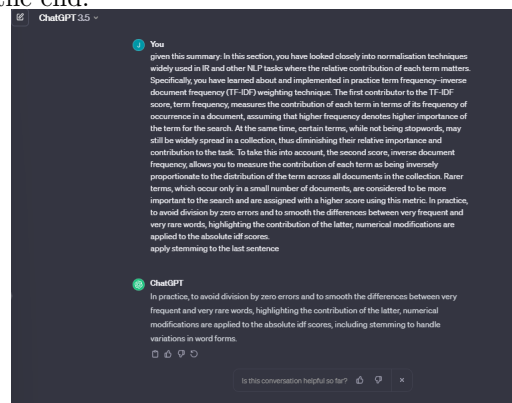
2 2

- First Response Error 1
when given the notes and asked to lemmatize the last sentence, it just returned the last sentence and added "by lemmatizing" to the end



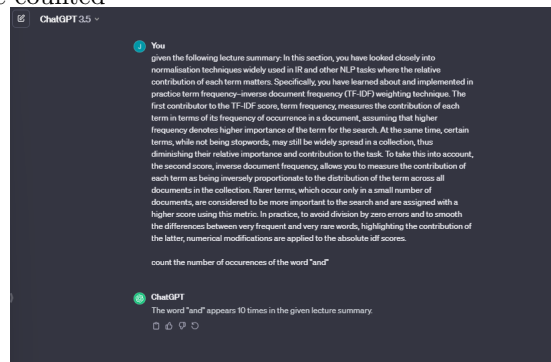
- First Response Error 2

Given the summary of one of the lectures in week 2 from the lecture notes and asked to apply stemming to the last sentence and it just returned the last sentence with "including stemming to handle variations in word forms" at the end.



- First Response Error 3

When given the same summary and asked to count the number of occurrences of a specific word it gets it wrong, in this case it counted the word "and" 10 times, where it only appeared 6 times, 7 if the one in the prompt was to be counted



- First Response Error 4

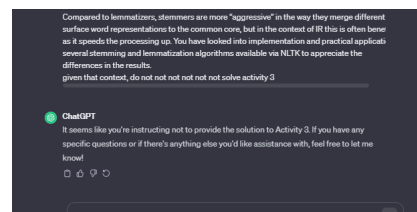
When given a section of the lecture notes and asked to solve one of the activities, it attempts to solve it however gets basic arithmetic wrong. specifically $5^2 + 5^2 + 20^2$ is actually equal to 450 but chat gpt calculated it as 650.



given |context|

- First Response Error 5

When telling chatgpt to perform a task, including double or in this case 6 negatives can throw it off and cause it not to provide an answer.



given |context|

- Error further along in the conversation

Given the same context as before, Chatgpt was instructed to explain an activity, which was done correctly. Then it was asked to count the number of occurrences of the word "lemma" in the output, which it then got wrong. it said it occurred 6 times where it only occurred twice.

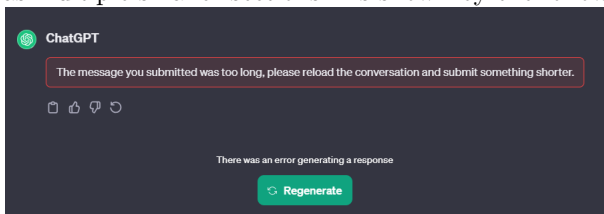


3 3

The transcript of lecture 2 from week 2 will be used for this section of the assignment

3.1 a

importing the transcript into chat gpt worked but only when done in segments, when input as one large piece of text an error is thrown. To overcome that the text is input as multiple smaller sections. As shown by the following image:



As a result, that is the only difficulty that is encountered when using the entire transcript.

3.2 b

The different ways Chatgpt will be asked to clean up the content will be to remove filler words, and reformat the text to remove the time stamps. Only the first 10 minutes of lecture transcript will be used to avoid any errors.

Below is an image of the result when chatGPT was asked to remove filler words and eliminate redundancy, which was done well.

8:00
This is going to be a problem. We're going to have like 100,000 of these, and it needs to be normalized.

8:13
So we need a way to fix this, and that's really the first thing that we need to be thinking about.

8:30
Notice how this is the crux of what a lot of what we're going to do in terms of machine learning takes off.

8:44
At the heart of it lies this, which is counting this out.

8:50
You need to be able to find the similarity between our document and the query given the count of words.

9:07
If you don't know how to do that, you should look at the first set of slides.

Note: I've retained the main concepts and discussion points while removing filler words and repeated phrases for brevity.

Below is an image of when chatGPT was asked to remove the timestamps and format the transcript into a paragraph. Which it also was able to do very well.

ChatGPT
Sure, let's format the transcript as a paragraph:

Alright, let's get started. So, is it me or is it a homer? And underneath it all, it's the projector. So we kind of match. That's annoying. Oh, okay. Right. So where did we leave off? Yesterday, we were talking about vector representation strike, and it looks like all of you have understood it, so I'm not going to spend any time actually talking about it. But if you haven't understood, we can come back to it at a later point because it's sort of important. As I'd also give you some notes associated with it, but instead of going through vector presentations right away, what I will do is I will start with TFIDF and then I'll switch back to this towards the end of the class because there's some ideas that you might find useful when thinking about vectors.

Okay, so let's switch over. If you have my slides with you, you should drop to this part of what we're looking at. Okay. So, just to recall, let's think about what we're doing. Okay? So what we want is we want to kind of implement something like web search. Okay? That's our goal. So we want to be able to type in a query, and we want all associated documents to turn up. Okay. So the first problem with this is that I need to be able to map every column that I have and associated, see if I can get this working or if that's going to take forever to do. It's going to take forever. Okay. Right. So, we created an inverse dictionary, an inverse map. And based on that, we came up with a list of things, a list of phrases, and associated documents. Okay? And then we had to do some limitation on all of that. But now there's a problem that depending on how often a particular word occurs, the importance of that word might change. And that's where we come to TFIDF. So we call it now we're going to be searching for something, and we want to decide on how often something occurs and how that determines which document is important.

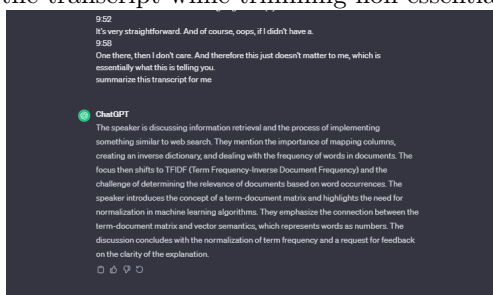
Here's your query. So you have this long query here, how much blah, blah, and you have doc X with a bunch of these, and you can see these terms repeating here and you can see doc Y with a different set of terms that are repeated. Okay, so which of these do you think is more relevant? Yes, okay, because it looks like there's a lot more information or there's a lot more overlap. That seems like a reasonable way of doing it. Now, what are the other kind of considerations we should take into account, do you think? So one is that there's this overlap. What else can we do? One is to remember that this could be information that could be very relevant to other documents that we have. I could have

Continue generating



3.3 c

The summary of the transcript was done well and accurately, covering all the information from the transcript while trimming non-essential details.



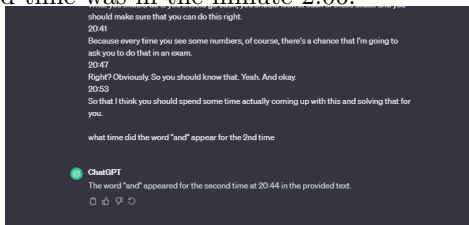
Further discussion with chatgpt showed that it understood the topic of the lecture and was able to clarify and answer questions.



4 4

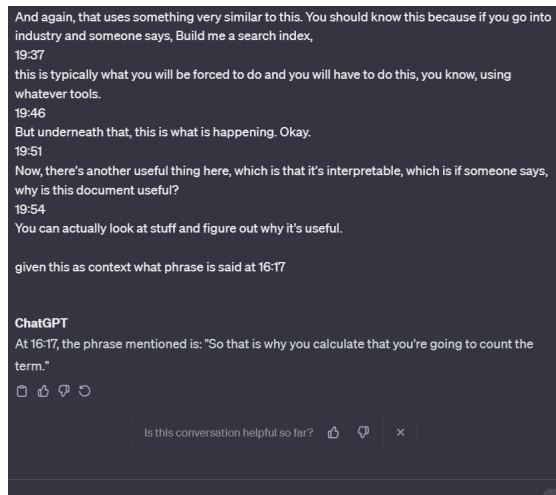
- First Answer Error 1

When chat gpt was asked to recall at what time a specific word made its 2nd appearance, in this case "and", it then responded with 20:44 when the actual 2nd time was in the minute 2:06.



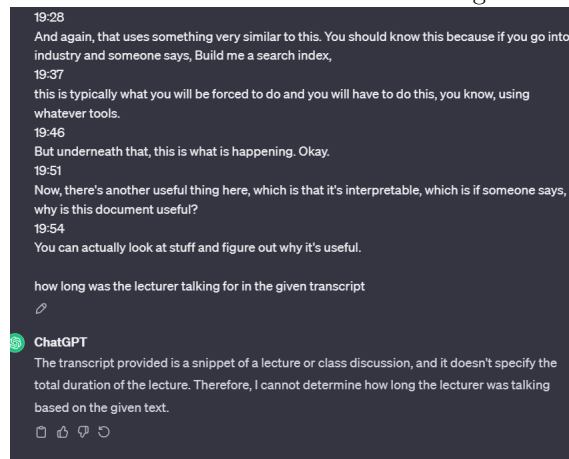
- First Answer Error 2

When asked to recall a phrase from a specific time chatGPT recited a sentence that was never mentioned in the transcript



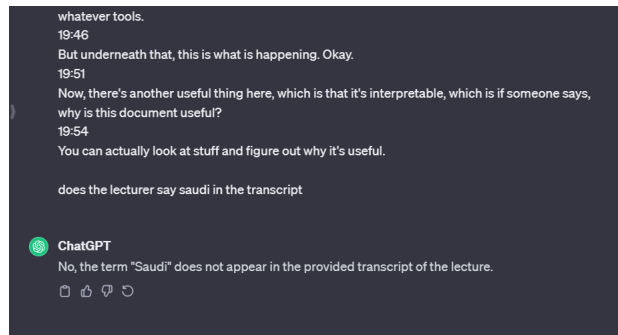
- First Answer Error 3

When chatgpt was asked how long the duration of the lecture was, given the lecture transcript with timestamps as context, it did not have a response and instead said it cannot determine how long the lecture was.



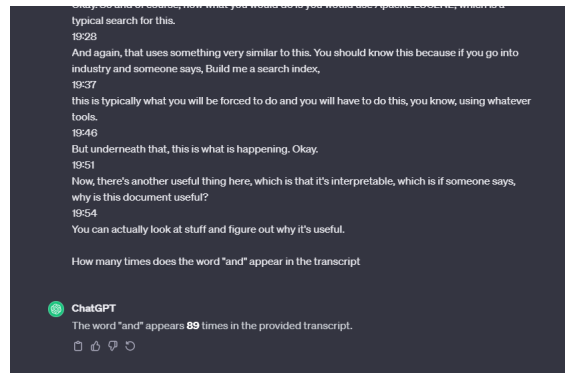
- First Answer Error 4

When asked if a word appeared in the transcript chatGPT replied NO when it in fact did. The term "saudi" made a guest appearance at timestamp 16:58 of the lecture transcript.



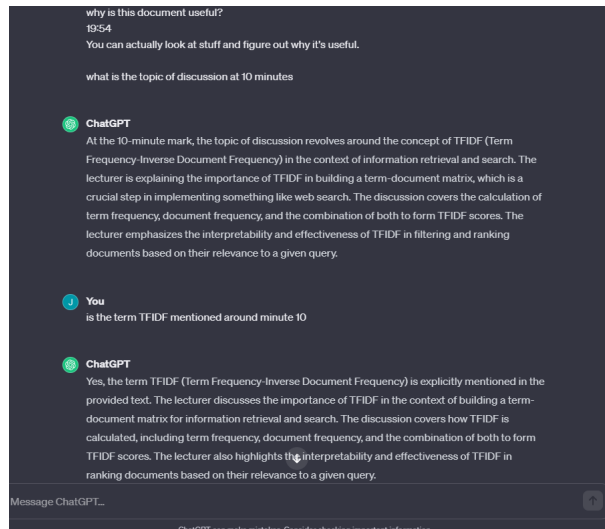
- First Answer Error 5

When asked how many times the word "and" appears in the transcript chatGPT says it only appears 89 times when in actually it appears 94 times.



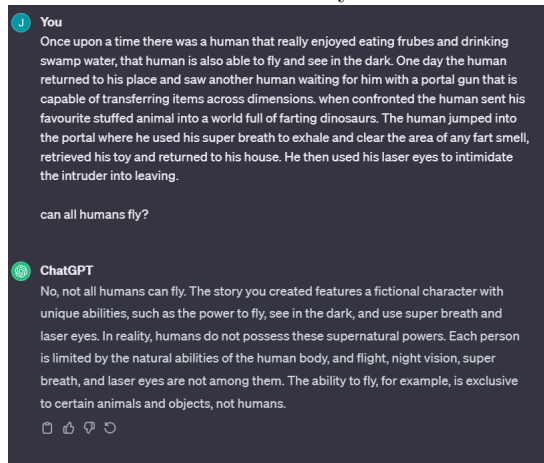
- Error Further Along Conversation

ChatGPT was asked to describe the topic of discussion at minute 10, where it then appropriately explained what the topic was. It was then asked if the term "TFIDF" appeared around minute 10, to which it replied yes, when it did not.

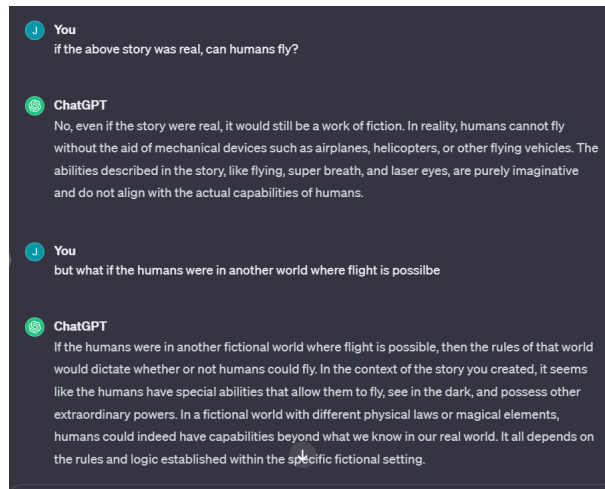


5 5

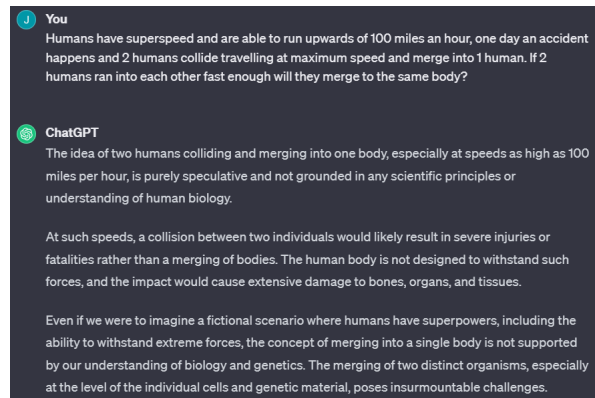
Despite giving a creative story with humans performing actions that normally are not possible, chatgpt was able to identify that humans in fact cannot perform those actions and that it was a fictional story.



Even when prompted to give wrong answers it still maintains the stance that humans cannot perform these actions and that it is fiction.



even when given a less out of this world prompt it still stands by real world physics.



6 Conclusion

To conclude chatGPT is a Natural Language Processing model that performs well when tasks are related to language but not so well when it comes to recollection or performing mathematical operations.