Introduction - 6 seconds

Good afternoon, my name is Conor and today I'll be talking you through the progress I've made in my final year project entitled "Markson's Memory".

Project Outline - 25 seconds

This is a research project on the subject of digital humanities, specifically the field of text analytics and in the course of the study, I've been familiarising myself with various approaches to text analytics and in particular, the use of a Python library called Natural Language Toolkit, which I'll refer to as NLTK.

Using this toolkit, I've developed data analytics techniques with a view to analysing one particular text called *Wittgenstein's Mistress*, a novel by American author David Markson.

Wittgenstein's Mistress - 45 seconds

Markson is known most for his unconventional style of writing, presenting the narrative of his books in experimental ways and Wittgenstein's mistress is no exception

This novel is set in a world with only one survivor, a woman who serves as the narrator.

Through the text, we see from her thoughts that she's becoming more and more insane, but something that remains constant is certain phrases she seems to keep coming back to, as well as a constant mention of figures from history.

While the content of the narrative is unique, the structure of the writing itself is also interesting. Here's a small sample of the book.

The entire thing is laid out like this, with no chapter breaks. Just an ongoing string of short paragraphs like these.

It's certainly an interesting novel and people have been analysing it's bizarre structure ever since it was published.

Close vs Distant - 30 seconds

These types of analyses are generally the classic format of an individual reading the text, retaining a certain amount of information and trying to keep track of those thoughts as they progress through it. This is what's known as a "close reading".

As an alternative to that lengthy and expensive process, with the age of digital humanities, another approach has emerged called "distant reading".

Here, a computer is the one reading the text, automatically extracting information from it, and presenting it in some format to the user, allowing them to make comprehensive conclusions about it.

Approach - 42 seconds

In my study, I performed this type of distant reading on Wittgenstein's Mistress, in the hopes of discovering some kind of repetition to the structure of the narrative particularly the associations between entities.

To achieve this, my approach involved four main areas of investigation.

First, I extracted all repeated phrases from the text, to find sayings she keeps coming back to.

Then I analysed those repeated phrases to find a list of the named entities that are repeatedly mentioned.

With those entities found, I then constructed networks showing connections between them.

Then lastly, I applied that network data to the text, trying to find some patterns to the author's train of thought.

I'll go through each of these steps now and demonstrate what's been achieved.

Phrase Repetition - 50 seconds

Since we're trying to find repetition in the book, it seemed a like good first choice to look for repeated strings in the text.

To get all of these possible sub-strings, I found every n-gram, where an n-gram is a sequence of n words, such as this set of 3-grams, 6-grams, and the single 9-gram, which is the whole paragraph.

And then after repetitions are found, only the most significant mutation of each phrase is kept.

We really didn't expect to find very long phrases but were surprised when we ran the program to find phrases of over 30 words perfectly intact, sometimes separated by half a book.

Clearly from this, Markson was very deliberate in his repetition of ideas, suggesting that a distant reading could actually find some significant information.

These phrases are put to use in the following task, but further work is still being done to provide visualisations of the data, and those will be featured in my final report.

Named Entity Recognition - 35 seconds

Named entities are essentially proper nouns that feature in a text.

Recognising these entities in text is a simple task for a human, as in this example, we can clearly see that Renoir and Cézanne are the names of people, but it's a lot more difficult for a computer to recognise this, since the process involves a combination of rules for the language structure, as well as a large set of training data.

Thankfully Stanford’s NER software is included in NLTK, so this process was actually made very easy for me.

I parsed through the set of repeated phrases found earlier, and used this to extract a set of entities, and where in the text they occur.

Networks - 1 minute, 24 seconds

Having found these entities, I was then able to move onto constructing networks to represent how they are all connected.

Here, two entities are connected if they occur sufficiently close together in the text, so to start this investigation off, they have to occur in the exact same paragraph.

We can see this particular network, how Achilles was found to be connected to both Odysseus and Alexander the Great, with much more weight attached to the connection to Odysseus, since this connection occurs more often.

Expanding this whole network, we can see all of the characters connected to Achilles, with many entities from Homer’s Odyssey grouped together here.

Revealing all of these networks, we discover that there are 5 disjoint networks present in the text, representing their own little clustering of ideas.

This result is so significant to the study because it shows definite structure to the way these entities are mentioned, repeatedly clustering ideas together, instead of having many small disjoint networks.

What surprised us is what happened when we allowed for connections between entities at most 1 paragraph away…we can see that the five disjoint networks have become one large network.

From this, we can conclude that not only are entities consistently grouped into their own concepts, but that these concepts aren’t independent, actually occurring one after another in sequential paragraphs.

Now one key piece of information is still missing from these graphs - an idea of time, since there’s no way to know from these graphs where in the text these relations occur.

Sequence Graph - 1 minute

So to find this, I took the network data and searched for these five groups in the whole text, to try to find a pattern to their occurrence.

These graphs show the timeline of the novel, with each point on the x axis representing a single paragraph, and the legend on the right shows the colours for our five networks.

Here, the network with Helen is shown and it single-handedly takes up a lot of the timeline, and similarly for the network featuring Achilles.

By plotting these networks' occurrences together in the timeline, we can establish a full sequence of jumps between them being mentioned, and by searching this sequence for common patterns, the following data was found.

Clearly, jumps between Helen's network and Achilles' are the most frequent, with over 74 instances of one being mentioned immediately after the other.

And there are even repeated sequences of the same 10 jumps between ideas visible in the last two lines on the right.

If we're ever going to get a definite answer as to whether or not there's deep repetition in the book, I think this is it.

Conclusion - 1 minute

So, to conclude, we've determined that distant reading is definitely applicable to this text, since in searching for repetition, we've found 3 types.

We've found a significant amount of word-for-word phrase repetition in the text, revealing a clear high level intent by the author.

We've also found a repeated grouping of entities, which verifies what we've been looking for: that there are deep regularities to the structure of associations between ideas.

And lastly, we've seen that there is also significant repetition to the order in which these groups occur, showing exactly how and when this deep repetition occurs between the associations, which may not have even been known to Markson.

Another significant thing to note is that the processes used in this study are applicable to any text, as the algorithms used are completely generic and automated.

This may suggest a possible use for this work in other studies, particularly those related to literary analyses of social networks, which is a common area of investigation in digital humanities.

End

That's it for my presentation. Thank you for listening and I'm open to any questions you might have about the project.