How James Joyce's style developed from *Dubliners* to *Ulysses*

1. Research Questions

As a pioneer in the use of stream of consciousness, James Joyce is reowned for his narrative style that seeks to depict the inner thoughts and feelings of characters in a flowing and fragmented manner. In his first collection of fifteen short stories, Dubliners (1914), Joyce employed a straightforward, realist style without incorporating stream of consciousness. However, some traces of this technique can be observed in A Portrait of the Artist as a Young Man (1916), including the use of interior monologue and a focus on a character's psychological reality rather than their external environment. After the publication of Ulysses in 1920, it exemplified his fully developed modernist style, blending myth, realism, and stream of consciousness.
This project aims to apply stylometric and distant reading techniques to compare these three books and identify stylistic similarities and differences between them. I will utilize computational stylistic approaches to explore the following questions:

- To what extent these three literary works differ in style?
- How James Joyce's writing style evolved from 1914 to 1920?
- What quantitative features characterize the use of stream of consciousness?

2. Data

The data used in this project includes plain texts of three books from James Joyce, which will be retrieved from Project Gutenberg:
1. *Dubliners* (1914)
2. *A Portrait of the Artist as a Young Man* (1916)
3. *Ulysses* (1920)

3. Preprocessing

In the first step of preprocessing, I manually remove the metadata at the beginning and copyright information at the end of each text file retrieved from Project Gutenberg. And then, with the means of regular expression, I processed three text files in the following six aspects:

1. Remove extra whitespace and newlines.
2. Replace curly quotes with normal quotes. Because these three files are all encoded in UTF-8, I recognized that double curly quotes and single quotes exist in the texts.

```
cleaned_text = re.sub(r'[\"\"]', '"', cleaned_text)
cleaned_text = re.sub(r'[\‘\’]', "'", cleaned_text)
```

3. Remove non-essential characters. In this step, non-essential characters were removed to ensure that only alphanumeric ones and the puntuations I selected are present in the files.

```
cleaned_text = re.sub(r'[^a-zA-Z0-9.,!?\'\-\-\"\s]', '',  cleaned_text)
```

4. Replace em dash with a white space. In the files, some em dashes are placed between two sentences without any whitespace. If the em dash is removed as a non-essential character, two words on the either side of the dash will be combined together, which will affect the subsequent calculations. For example, in *A Portrait of the Artist as a Young Man*:
"... by the force of its delight in rude bodily **skill—for** Davin had sat at the feet of Michael Cusack, the **Gael—repelling** swiftly ..."

```
cleaned_text = re.sub(r'-', ' ', cleaned_text)
```

5. Replace the ellipsis with only one period. The ellipsis will affect the accuracy of sentence tokenization as `nltk` will tokenize two sentence on the either side of the ellipsis as only one sentence.

```
cleaned_text = re.sub(r'\.{2,}', '.', cleaned_text)
```

6. Lowercase the whole text.

4. Analysis

I. Lexical Variety and Richness:

The first step in the analysis is to look at word level. TTR is a useful measurement of linguistic diversity. It is defined as the ratio of unique tokens divided by the total number of tokens. A higher TTR indicates that a wider range of vocabulary was used in the text.

For The below table presents the basic statistics of the type-token ratio (TTR) in three texts.

Statistic	**Dubliners**	**Portrait**	**Ulysses**
Tokens	67,328	84,774	262,800
Types	7,272	9,026	28,949
Type-Token Ratio	0.11	0.11	0.11
Standardized Type-Token Ratio	0.41	0.42	0.50

Althoug simple TTRs for three etxts remain relatively same, there're big differences between standardized TTRs. Ulysses has the highest STTR, which is 9% and 8% higher than Dubliners and Portrait. This results indicate that James Joyce used more vocabulary items in Ulysses than in Dubliners and Portrait. This is because Joyce invented lots of new words in Ulysses, such as 'contransmagnificandjewbangtantiality', "ripripple", "weggebobble" and so on. Joyce aimed to capture the intricacies of human consciousness through stream-of-consciousness narration. These invented words often mirror the spontaneity of thought and the fluidity and fragmentation of memory and perception. Words like "snotgreen" vividly encapsulate sensations or emotions in a single, striking image. Hence, the differences of STTRs significantly imply Joyce's heavy use of stream of consciousness in Ulysses.

In the next, percentages of hapax legomena and lexical density are calculated, which also indicate that the use of stream of consciousness increased a lot through Joyce's works.

Hapax legomenon refers to a word or an expression that occurs only once within a context. Although this type of words could be overlooked due to their infrequency in the text, they are important for the analysis of lexical richness. As the below table illustrates, percentage of hapax legomenon is higher than the other two works.

Statistic	**Dubliners**	**Portrait**	**Ulysses**
Hapax Legomena Count	3,751	4,515	15,671
As Percentage of Tokens (%)	5.57	5.32	5.96
As Percentage of Types (%)	51.58	50.02	54.13

Lexical density is a way to measure the structure and complexity of a language. It estimates the linguistic complexity in a written or spoken composition from the functional words (grammatical units) and content words (lexical units, lexemes). One method to calculate the lexical density is to compute the ratio of lexical items to the total number of words, which is the means utilized here.

The below table displays lexical density of each work. It can be clearly observed that lexical density of Ulysses is higher than other two works. The language used in Ulysses is more complex and diverse, which is consistent with the language characteristics stream of consciousness possesses: the text is complicated and difficult to understand.

Work	Lexical Density (%)
Dubliners	55.67
Portrait	54.46
Ulysses	59.04

Meanwhile, I counted the top 100 frequent words in each work and visualized the results as word clouds. These frequent words can unveil some interesting stylistic features. I will leave them for the future further close-reading analysis. During the calculation of top 100 frequent words, a stopword list was introduced to remove less relevant words. Here're the visualization results:

Figure 1: Wordcloud for Dubliners

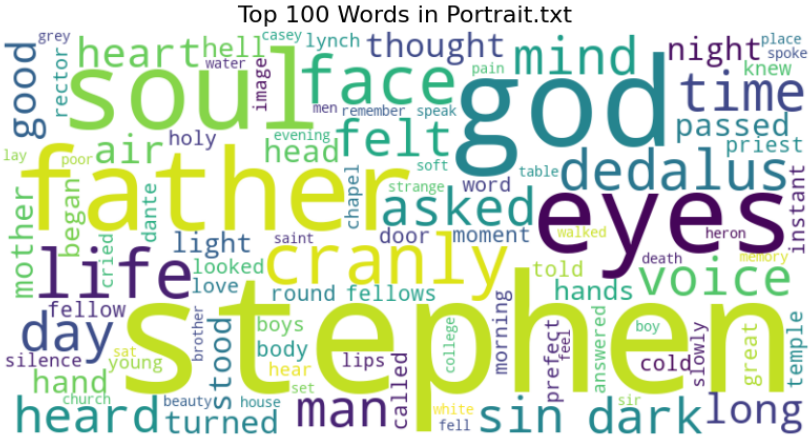

Figure 2: Wordcloud for Portrait

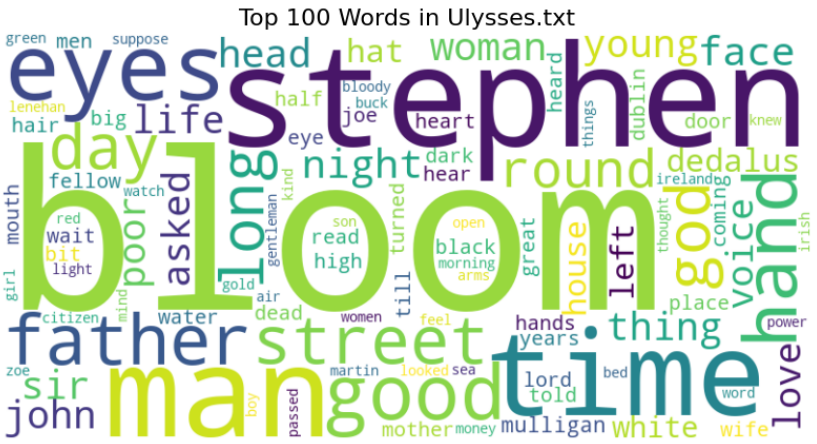

Figure 3: Wordcloud for Ulysses

II. Syntactic Features:

The second step in the analysis is to look at sentence level. As there're no any puctuations used in almost forty pages of Ulysses' last chapter, sentence tokenization will be not fully correctly conducted, which is bound to lead to biases in calculations of syntactic features for Ulysses. Manually tokenizing these forty pages into sentences will be a huge challenge, so I couldn't find a better solution. Although the final results contain biases, to some extent they indeed uncover some important syntactic differences between three works.

Here's the table for the stats of sentence lengths in each work.

Statistic	**Dubliners**	**Portrait**	**Ulysses**
Total Sentences	4,774	5,381	24,102
Mean Length	14.10	15.75	10.90
Median Length	12.0	12.0	6.0
Min Length	0	0	0
Max Length	94	148	12,817

The maximum sentence length in *Ulysses*, 12,817, is the biased result influenced by the punctuation-free style of Chapter 18. While the stats for Ulysses are not totally accurate, the differences between the mean and median length of three works can still suggest that sentence lengths on avergae in Ulysses are shorter than in Dubliners and Portrait. It also closely aligns with one of the characteristics of stream of consciousness: sentences are often incomplete, fragmented, or broken, mimicking the disjointed nature of thought.

And then, calculation of 3-grams to 6-grams which occurred a minimum of 5 times in each text file was conducted. The table below presents the results of the number of n-grams at each length.

N-grams	**Statistic**	**Dubliners**	**Portrait**	**Ulysses**
3-grams	Tokens	2,993	3,670	10,115
	Types	371	471	1,262
4-grams	Tokens	202	363	642
	Types	31	50	98
5-grams	Tokens	12	89	65
	Types	2	12	11
6-grams	Tokens	6	27	5
	Types	1	4	1

N-gram frequency can be used as a metric for language and translation creativity. An author who is more creative may be expected to use fewer fixed expressions. Ulysses and Dubliners demonstrates more linguistic and translational creativity, as they rely less on repeated fixed expressions. It can indicated that the conventional grammer is used very less in the text full of stream of consciousness.

III. Stylometric Analysis Based on Delta Method

In this aspect, stylometric analysis will be initially implemented with John Burrows' Delta method. This algorithm will first assemble a corpus of different texts and identify the n most frequent words as features. It calculates z-scores for these features by comparing their frequency against corpus norm. Finally, a delta score is computed by averaging the absolute differences between z-scores of each text.

Equation for the z-score statistic:

$$Z_i = \frac{C_i - \mu_i}{\sigma_i}$$

Equation for John Burrows' Delta statistic:

$$\Delta_c = \sum_i \frac{\left| Z_{c(i)} - Z_{t(i)} \right|}{n}$$

Here're the delt scores calculated:

	Dubliners	Portrait	Ulysses
Dubliners	0.000000	1.239246	1.249384
Portrait	1.239246	0.000000	1.319008
Ulysses	1.249384	1.319008	0.000000

From the table, it can be seen that the stylistic distance between Dubliners and Portrait is smaller than the one between Dubliners and Ulysses. Smaller Delta values indicate higher stylistic similarity between texts. Thus, it can be inferred that use of stream of consciousness increased through three works.

IV. Stylometric Analysis Based on Machine Learning

When it comes to implementing stylometric analysis by machine learning, there're a lot to consider. One of the most important step is feature selection. While in many research style is generally understood as the distribution of most frequent words, I would like to also add syntactic features and lexical variety in the feature selection period. Besides, which classification model or cluster model to use also needs to be carefully considered. This task is time-consuming. And I will leave this part for the future further exploration.

5. Conclusion

With the calculation of lexical variety, syntactic features and John Burrows' Delta method, some characters of stream of consciousness can be observed from the computational methods. They also display how James Joyce's style developed through these three works.