Yuxuan Zhou
2023 Spring
DIGS 20032 1 (Spring 2023) Digital Texts II

Final Project Written Explanation: Comparative Literature Analysis of Drama Scripts by Oscar Wilde and Joe Orton

Abstract

Purpose-This project explores the comparative literature analysis of drama scripts by Oscar Wilde and Joe Orton. Both playwrights are known for their sharp wit and satirical examination of societal norms. This study aims to shed light on the intriguing similarities between Wilde's and Orton's dramatic works. It also seeks to highlight the unique aspects of their approach to social critique. This study aims to determine whether this perceived similarity holds up under statistical scrutiny using natural language processing (NLP) techniques.

Design/methodology- A comparative analysis of selected plays by Wilde and Orton was conducted. Generative AI and NLP techniques were used to analyze thematic patterns, linguistic styles, and rhetorical devices. Sentiment analysis, motif detection, and character interaction analysis were performed to provide a data-driven perspective. Statistical measures were applied to evaluate the extent of similarity in their use of language, themes, and narrative structures.

Findings-The statistical analysis supported the notion of a stylistic similarity between Wilde and Orton. The text analysis revealed parallel themes and comparable usage of linguistic structures, corroborating the subjective feeling of similarity often noted by critics and scholars.

Conclusion-The findings validate the perceived auteur-like similarity between Wilde's and Orton's works, demonstrating how modern computational tools can substantiate traditional literary critiques. The integration of generative AI and NLP techniques provides a novel approach to literary analysis, offering new insights and enhancing the rigor of the comparative study.

Keywords- Comparative literature, Drama analysis, NLP, Oscar Wilde, Joe Orton, Societal critique

Introduction and Scope

This project conducts a comparative literature analysis by cross-examining drama scripts written by Oscar Wilde and Joe Orton. The primary focus is on Wilde's "*The Importance of Being Earnest*" and Orton's "*What the Butler Saw,*" a play often criticized for its stylistic similarities to Wilde. As John Bull notes, "*It is not particularly surprising that Orton should take Wilde as a model in his first years of attempting to act and then later to write*" (Bull 50). However, topic similarity analysis using HCA clustering revealed that the characters from "What the Butler Saw" exhibit similarities not only to those from "The Importance of Being Earnest" but also to characters from other Wilde plays. Thus, the study evolved by separating the characters from "What the Butler Saw" into different clusters based on HCA results, followed by an in-depth characterization investigation of Orton's auteur

Data

The data collection involved manually gathering text files from the Drama Online Library(https://www-dramaonlinelibrary-com.proxy.uchicago.edu/). The selected plays include "Visitors," "Fred and Madge," "What the Butler Saw," "Entertaining Mr. Sloane," and "Loot." Each play script was split by paragraph, aligning with the format where each character's lines start a new paragraph. This approach ensured an accurate corpus for each character.

While a typical NLP program would split by lines, I split the script by paragraphs to gain an accurate corpus for each character. The split data contains the following: 1) voiceover of the play, 2) stage setting or description, and 3) speech lines for each character.

The play scripts were processed by selecting paragraphs containing speech lines, indicated by the format "XXX: ", where "XXX" refers to character names. Only paragraphs with this format were selected to ensure that the dataset included only explicit speech lines. The character names were extracted to create a new column in the data frame, forming the basis for the "speech_line_O.csv" database for Orton's plays.

Similarly, speech lines for Wilde's characters were collected, resulting in the "speech_line_W.csv" file. The combined data structure formed a corpus folder containing speech line collections for each character in text file format for both Wilde and Orton.

Analysis

The Exploratory/Descriptive Analysis:

TF-IDF similarity check: Similarity index between Orton and Wilde's corpus: 77.44%

Conducting a TF-IDF similarity check for the wording similarities in Wilde's and Orton's corpora reveals their unique vocabulary, thematic focus, and stylistic choices. TF-IDF analysis quantifies the importance of terms within their respective works, allowing for a comparative measurement of semantic similarity between the two writers. TF-IDF was chosen because it accounts for both the frequency of a term in a document (term frequency) and the rarity of the term across the entire document collection (inverse document frequency). This method assigns higher weights to terms that are more frequent within a specific document but less common in the overall collection. For instance, character names, often unique and frequent, could add noise to the results. By applying TF-IDF, the relative importance of terms in a document is captured, enabling the measurement of similarity based on their term distributions.

On the other hand, Doc2Vec is an algorithm that learns vector representations (embeddings) for entire documents, capturing both the semantic meaning and context of the text. However, for my research, this function is unnecessary since other semantic analyses will follow. Doc2Vec, an extension of Word2Vec used in class, learns embeddings for individual words and represents documents as fixed-length vectors, allowing similarity computations. However, Doc2Vec breaks the corpus context into sentences, which is not desirable for this analysis.

As the result of TF-IDF similarity check, 77.44% is relatively high and I think it implies the similarities of wording between Orton and Wilde's corpus existed.

PCA stylometric analysis:

I applied codes from out week3 assignments to achieve a general stylometric analysis on each playwrights characters.

According to PCA stylometric analysis, Wilde's characters setting has more diversity and different between each other, which matches his play style that in generally maintain more characters, instead of Orton, only conduct 5-6 main characters, less background/supporting characters. Compare to Orton, Wilde has more character differentiation.

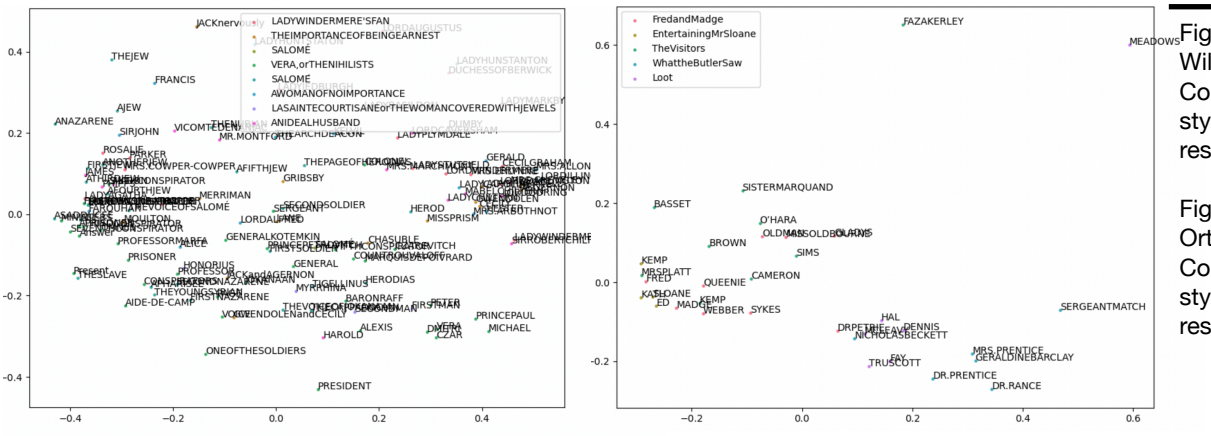


Figure 1.1.1 (on left) Wilde Characters Corpus PCA stylometric analysis result

Figure 1.1.2 (on right) Orton Characters Corpus PCA stylometric analysis result;

By applying the PCA among all of there characters, Figure 1.1.3, a cluster of Orton and Wilde's characterization overlapping to each other, which possibly leading the idea of characterization of similarities of Orton's is only applied to a partial to Wilde's.

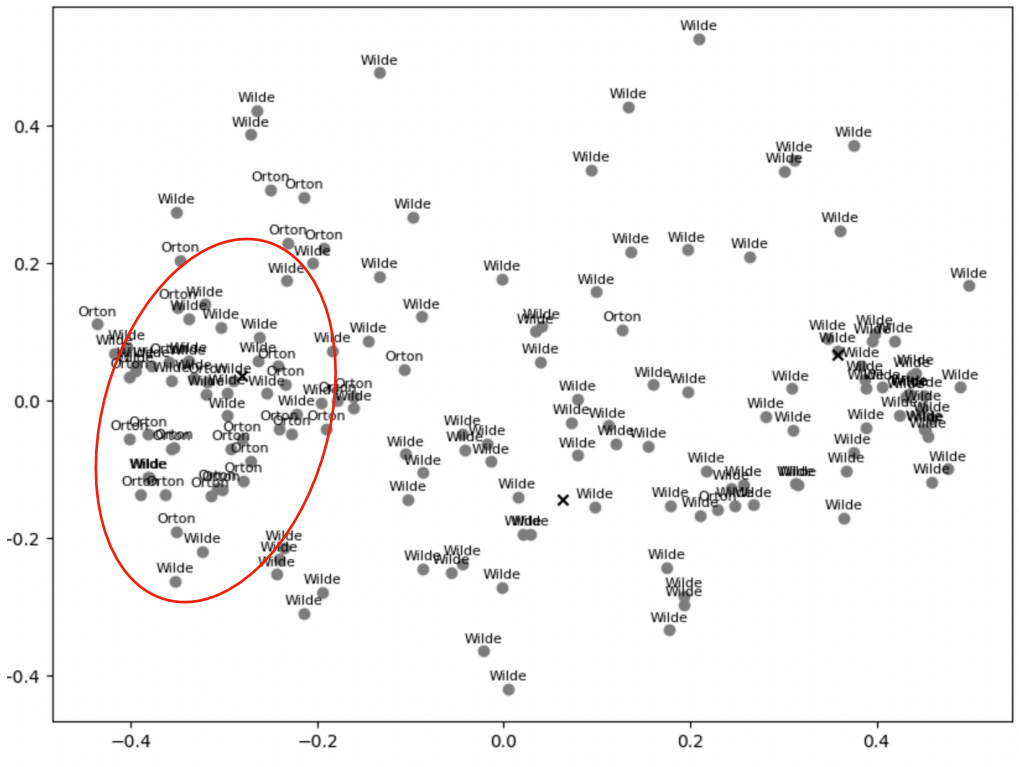
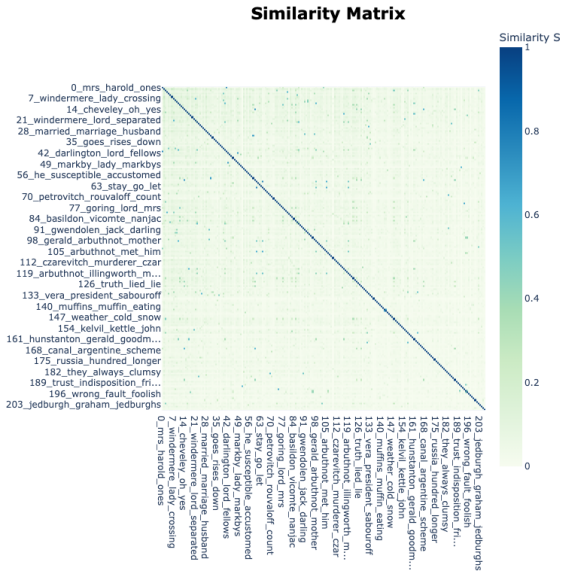


Figure 1.1.3 Wilde & Orton Characters Corpus PCA stylometric analysis result

Similarity Matrix of Topic Analysis - BERTopic

According to the result of BERTopic similarity matrix of topics analysis, in Wilde's corpus, the topic among characters are differentiated and has low similarities; however, in Orton's corpus, the similarities among characters' topic is higher. Hence, Comparing to Wilde, Orton has more "Auteur" or signature on topics of storytelling than Wilde.



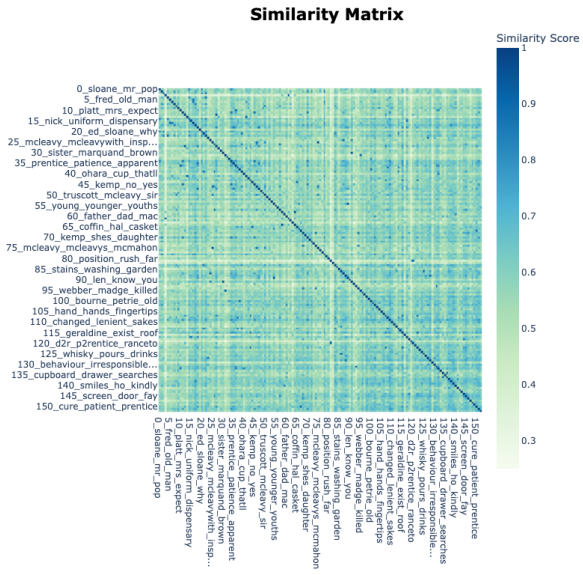


Figure 1.2.1 (On left) Wilde Characters Topic Similarity Matrix
Figure 1.2.2 (On right) Orton Characters Topic Similarity Matrix

HCA clustering result - BERTopic

Based on the clustering result, Figure 1.2.1 and 1.2.2, Orton's topics also presents stronger internal correlation than Wilde's.

Gaining form the stylometric analysis clustering result, Figure 1.2.3 and 1.2.4, I set the total desired cluster number as 8, Figure 1.2.5, which I think is the number that maximizing the differentiation between each cluster and also gaining enough details from the characters.

I trim the characters, due to Wilde play tend to include more small supporting characters, I exclude the supporting characters mainly for Wilde's play by two main standard: 1) Has more than 10 speech lines. 2) Has a name (i.e first solider or General is not an actual name). At this point, I will receive 8 new clustered older that maintain the most similar, accounting to HCA BERTopic clustering, on Wilde and Orton's character's speech line corpus. For future more meaningful result seeking purposes, I discard the cluster#2 and cluster#3 due to it only contains Wilde's play in those two clusters.

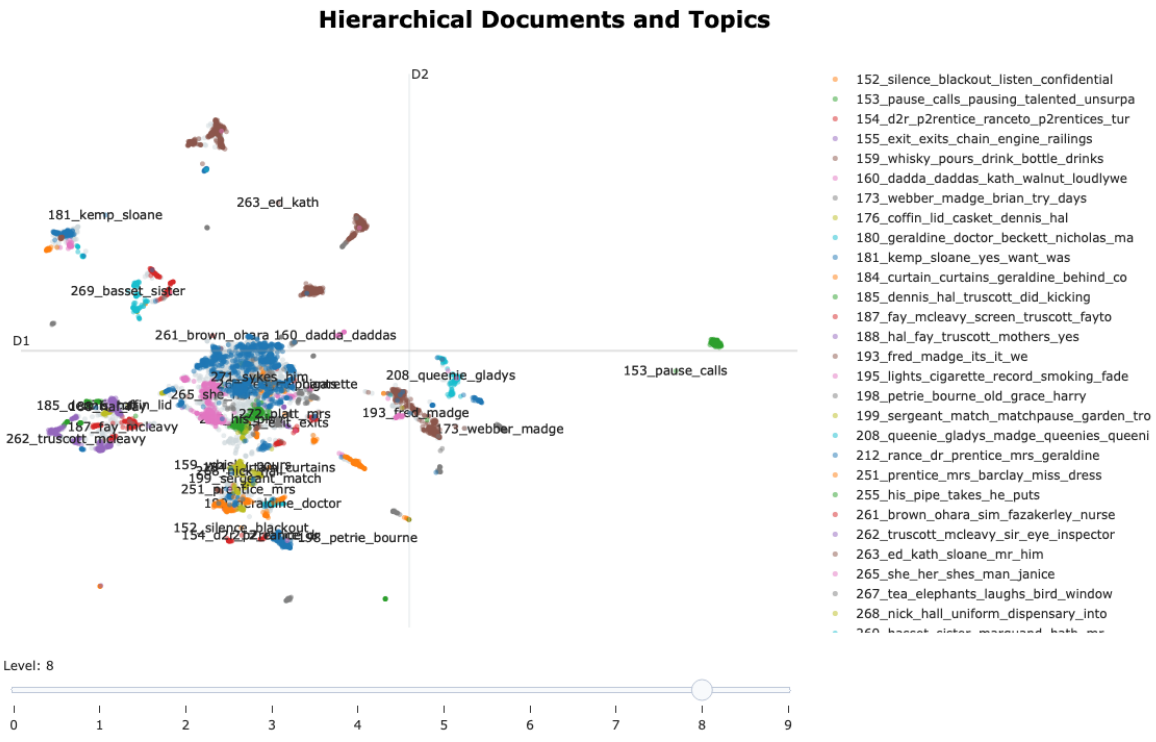


Figure 1.2.3
Orton HCA clustering result - L8

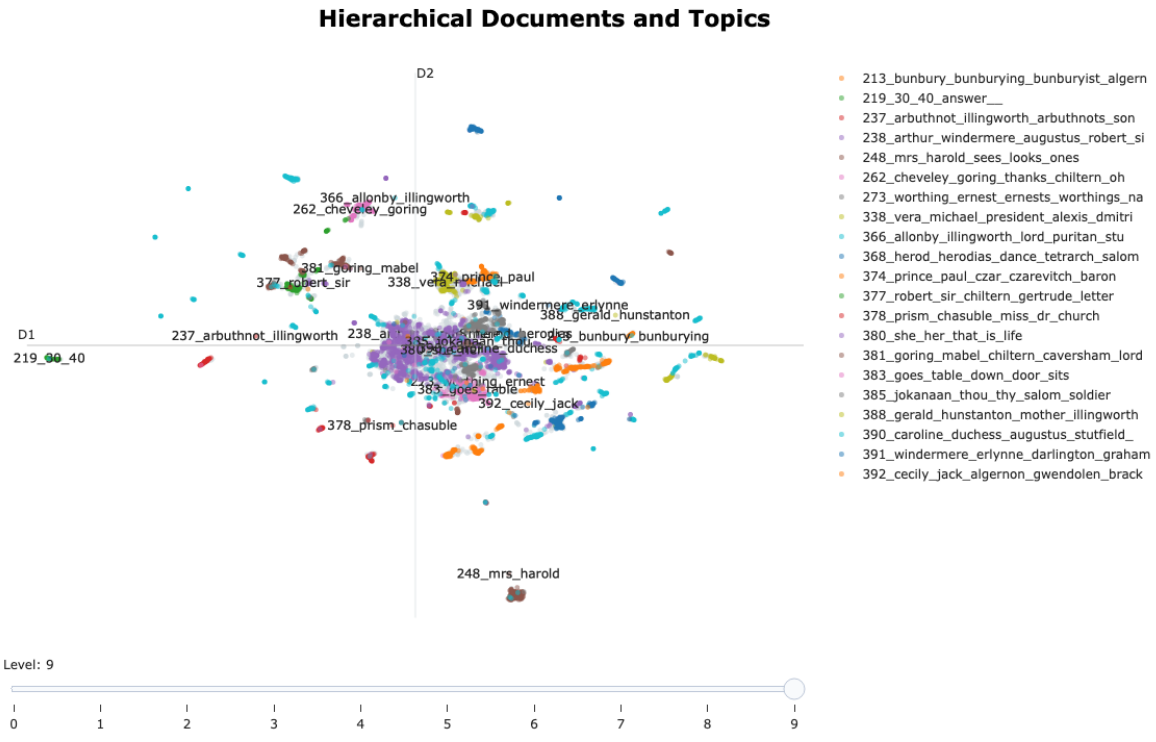


Figure 1.2.4
Wilde HCA clustering result - L9

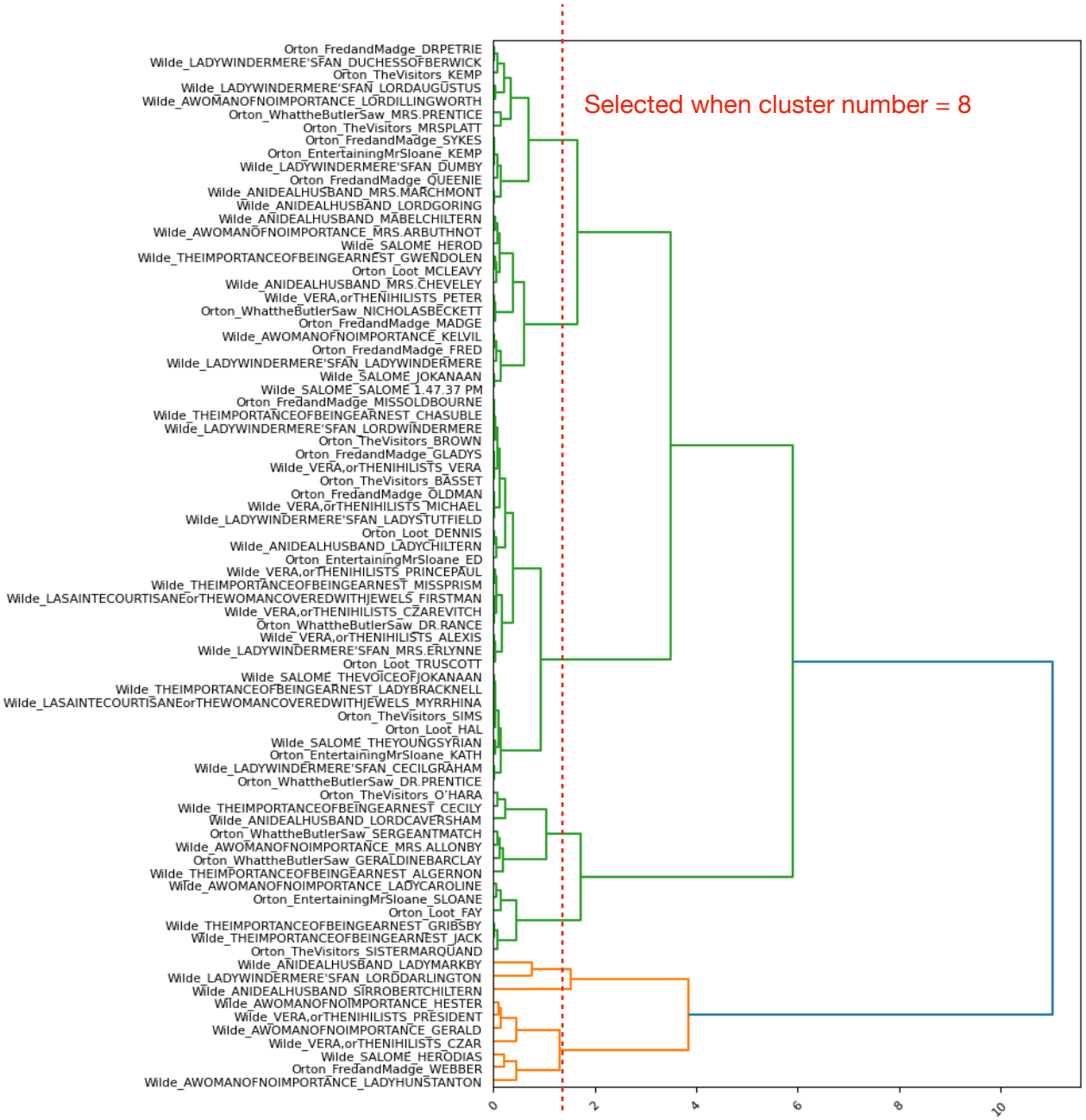


Figure 1.2.5
Main characters Wilde and Orton HCA clustering result

NRC Emotion Lexicon

Based on the teach cluster, I applied NRC Emotion Lexicon, which is a list of words and their associations with eight basic emotions: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. It is used in natural language processing (NLP) and sentiment analysis to identify the emotions expressed in clustered folder to analysis the emotion that has been expressed that Orton and Wilde characters towards the same topic.

NRC Emotion results are virtualized by radar chart (Figure 2.2), Orton's characters are colored in red, and Wilde's in black. The emotions exhibited by different character clusters towards the

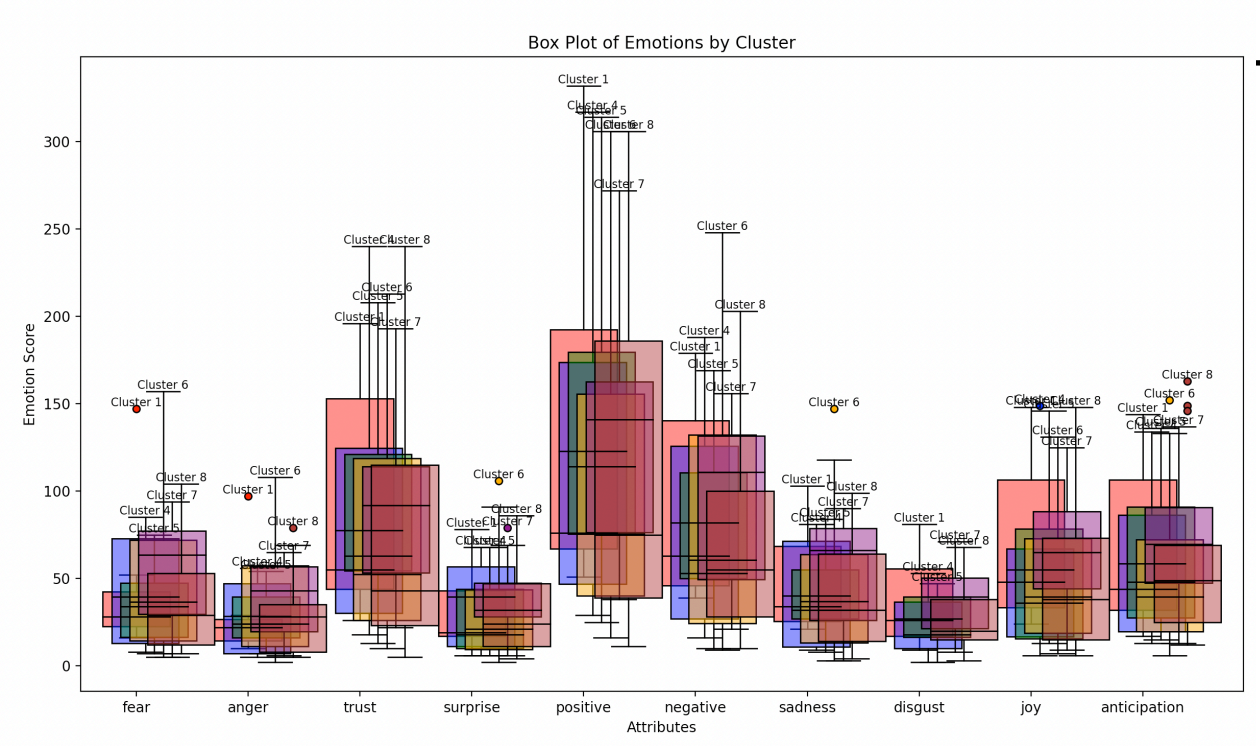


Figure 2.1
Main characters
Wilde and Orton
HCA clustering result

Cluster	Fear	Anger	Trust	Surprise	Sadness	Disgust	Joy	Anticip	Postive	Negative	STD
1	Low						High		High		High
4			High	High	High						
5											Low
6			Low					Low			
7	High	High			High	High	High		High	High	
8	Low				Low	Low				Low	

Table 1
Main characters
Wilde and Orton
HCA clustering result

works of Oscar Wilde and Joe Orton significantly. Cluster #1 reflects lack of fear and high joy that turned into overall positive, also noticed that cluster 1 has extremely high std compare to others. In Cluster #4, high levels of trust and the presence of surprise, sadness are observed, with Orton's works demonstrating more joy. Cluster #5 no obvious distinguishable emotions observed, but the standard deviation significantly smaller than other clusters. Cluster #6 experiences low trust and anticipation, highest std on disgust. Cluster #7 is characterized by high levels of joy, fear, anger, sadness, disgust and anticipation turned this cluster into the most negative and also the most positive character cluster, this brings me confuse. Cluster #8 is marked by low levels of negative: low fear, sadness and disgust. These distinctions highlight the varied emotional construction of different character clusters towards the works of Wilde and Orton. Draw a chart to virtualized the result better.

To explore the potential reason of differentiation on standard deviation I virtualized each cluster by radar chart to see which character is the outlier. Color different by author. Red is Orton, Black/grey is Wilde. Results display that high std in cluster#1, might mainly generated by the

gap of difference between Wilde and Orton's character corpus, Low std for cluster#5 is not relate to the similar emotion expressions between Wilde and Orton but the consistency of Wilde. In general, Cluster#6,#7 and #8 has similar character emotions from both Wilde and Orton. Hence, of the next step, future analysis focus on Cluster#6,#7 and #8.
Cluster#6 can be characterized by a sense of caution and skepticism as "**guarded"**.
Cluster #7 can be described as emotionally complex or emotionally "**volatile"**.
Cluster#8 is characterized by a positive and resilient demeanor, described as "**clam**".
On this level, the characters in Wilde's and Orton's plays that can be featured as "guarded," "volatile," and "clam" often have more similarities than others.

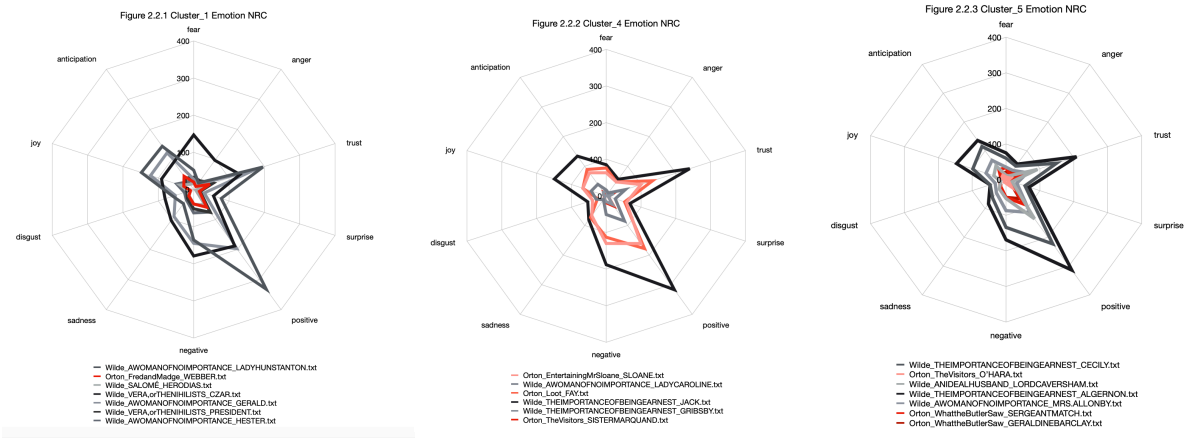


Figure 2.2
Emotion NRC Radar charts by characters' clusters

	Topic#1	Topic#2	Topic#3	Topic#4
Cluster#6 Guarded	'worthing', 0.39913696, 'cecily', 0.36841357, 'sermon', 0.31917334, 'beg', 0.28794765, 'desire', 0.27991834, 'diary', 0.25695163, 'say', 0.2510131, 'mean', 0.24350807, 'primitive', 0.23820981, 'suppose', 0.23589206	'erlynne', 0.60621345, 'margaret', 0.5956734, 'mrs', 0.5036453, 'woman', 0.34906012, 'wife', 0.3439616, 'mother', 0.32384753, 'sorrow', 0.31876013, 'cecil', 0.28798407, 'moralises', 0.27861825, 'gossip', 0.27649608	'worthing', 0.5329516, 'sermon', 0.3459756, 'cecily', 0.33535695, 'mr', 0.33071285, 'beg', 0.3287043, 'charity', 0.32352632, 'dear', 0.30618346, 'desire', 0.29274946, 'mean', 0.2786731, 'occasion', 0.27463138	'sloane', 0.24641445, 'word', 0.2229948, 'prentice', 0.21702458, 'shall', 0.21575631, 'put', 0.18507384, 'mr', 0.17613357, 'get', 0.16995364, 'father', 0.16962764, 'robert', 0.16881354, 'think', 0.16733491
Cluster #7 Volatile	'shall', 0.24411505, 'lady', 0.21831216, 'hope', 0.20762467, 'janice', 0.20507793, 'illingworth', 0.19967556, 'mr', 0.19833925, 'wish', 0.19771956, 'think', 0.19482401, 'give', 0.18866614, 'must', 0.18108451	'mamma', 0.4478253, 'marry', 0.35542756, 'propose', 0.33986413, 'mr', 0.32424462, 'worthing', 0.29125857, 'speak', 0.29120117, 'married', 0.2659316, 'cecily', 0.24328811, 'say', 0.24312328, 'ernest', 0.2369959	'erlynne', 0.5270393, 'margaret', 0.37162858, 'mrs', 0.34171164, 'shall', 0.32305574, 'arthur', 0.31634587, 'name', 0.30092448, 'duchess', 0.30054384, 'selby', 0.2950972, 'windermere', 0.29341352, 'rose', 0.28099972	/
Cluster #8 Clam	'lady', 0.2515624, 'shall', 0.22417705, 'mrs', 0.22290239, 'miss', 0.21872208, 'letter', 0.20769674, 'chiltern', 0.20477453, 'mean', 0.20022875, 'must', 0.19813599, 'mr', 0.19694263, 'think', 0.19675523	'agatha', 0.69123876, 'margaret', 0.38655493, 'hopper', 0.37126616, 'kangaroo', 0.31273484, 'mr', 0.28397572, 'dear', 0.27694896, 'niece', 0.27508593, 'lady', 0.2723523, 'scandal', 0.24716687, 'waltz', 0.24280477	'erlynne', 0.42135555, 'dear', 0.41752583, 'windermere', 0.40028447, 'mrs', 0.39713883, 'lady', 0.3606015, 'wife', 0.35951442, 'evening', 0.3323633, 'manner', 0.3285527, 'demmed', 0.32439405, 'mr', 0.31819344	/

Table 2
Keywords among the BERTopic

Large Language Model - ChatGPT3.5

Now I applied ChatGPT3.5, Large language model to merge these close reading topic details into more meaningful and descriptive result. My prompts to GPT is: “Give me a summary description based this”, this refers to the above chart information, also I already trained and tuned this conversation model by previous code generating so GPT would gain enough perspectives on my current project.

Here is an example of my result obtained from GPT for Cluster#6:

- “Topic#1: This topic is characterized by words such as 'worthing', 'cecily', 'sermon', 'beg', 'desire', 'diary', 'say', 'mean', 'primitive', 'suppose'. These words suggest a theme related to personal reflections, desires, and introspection.

Final result from ChatGPT 3.5 as follow Table

	Topic#1	Topic#2	Topic#3	Topic#4
Cluster#6 Guarded Trust(L), Anticipation(L)	personal reflections, desires, and introspection	interpersonal relationships, moral judgments, and societal gossip	religious or philosophical discussions, personal values, and occasional events	communication, actions, relationships, and thoughts
Cluster#7 Volatile Joy(H), Fear(H), Anger(H), Sadness(H), Disgust(H), Anticipation(H), Trust(L), Anticipation(L)	desires, thoughts, decisions, and actions	marriage proposals, relationships, and communication	interpersonal relationships, family dynamics, and character names	/
Cluster#8 Clam Fear(L), Sadness(L), Disgust(L)	conversations, thoughts, letters, and social interactions	character names, relationships, and potentially social events	interpersonal relationships, marital dynamics, and social situations	/

Table 3
Emotions of the characters clusters and the topic they engaged most

Conclusion

In terms of overall writing style, Orton is more stylistically focused than Wilde, while Wilde's stylometric analysis based on character classification results are more dispersed, by comparing the position of characters within specific play corpus, all characters of the same corpus are similarly positioned in terms of results, so Wilde's stylometric dispersion is not due to role diversity, but more likely due to the large differences between different plays.

But in contrast Orton's stylometric results are more concentrated, and can even be roughly divided into two categories in terms of style from the PCA results. Mixing and overlapping PCA stylometric analysis of Orton's and Wilde's character corpus, it can be seen that in the narrative interval of a certain character, the stylometric similarity between Orton's and Wilde's works is high, so the next step of research can be conducted.

From the perspective of similarity of textual diction, I used TF/IDF and the results were very satisfactory, learning that O and W have 77.44% similarity, which says that from the perspective of diction, they are very similar. In addition to style, the similarity between the characters' topic degrees was also further compared. Among them, the overlap of topics among Orton characters is obviously very much higher than Wilde's, which further proves that Orton's writing topics and style are highly focused compared to Wilde's, and even the narrative topics overlap between different plays. This proves the first point that Orton's Auteur is clearly present, but Wilde would be more dispersed. I personally believe this is due to Wilde's play scrpit having more characters and his expertise in depicting rich social topics.
Cluster analysis was performed based on the similarity of topics. I divided all the character collections of O and W into 8 clusters and removed the cluster with only a single author character. Six were obtained and Emotion NRC analysis was performed to see if the characters showed the same emotion in each text set when faced with similar topics. The results were that only clusters #6, #7, and #8 had smaller cluster standard deviations, meaning that the Orton and Wilde characters showed similar strengths and types of emotional expressions on this topic. Further close reading to see the specific topics contained in these clusters results in that when character personalities are displayed as low trust and high suspicion, here is my character profiling:

Guarded Character:

This person keeps their personal reflections and desires hidden, often masking their true feelings behind a guarded facade. They engage in introspection and may have hidden depths of emotions and desires.

Judicious Character:

This person approaches interpersonal relationships with caution, making thoughtful moral judgments. They are aware of societal gossip and are careful in navigating social dynamics.

Contemplative Character:

This person engages in deep contemplation and reflection, particularly in matters of religion, philosophy, and personal values. They may keep their beliefs private and only express them on select occasions.

Articulate Character:

This person communicates with wit and chooses their words carefully. They are mindful of the impact of their actions on relationships and tend to guard their thoughts until they feel comfortable sharing them.

Among Cluster #7 - Volatile: These characters exhibits high levels of joy, fear, anger, sadness, disgust, and anticipation. They also have low levels of trust and anticipation. The character can be described as:

Bold Character:

This person is driven by their intense desires and emotions, often making impulsive decisions and taking actions based on their immediate impulses. They may have a volatile nature and can be unpredictable in their behavior.

Romantic/Marriage Character:

This person approaches relationships with passion and intensity, expressing their emotions openly and making grand gestures in romantic pursuits. They are enthusiastic communicators and are not afraid to express their feelings.

Dynamic Character:

This person thrives in interpersonal relationships and is actively involved in various family dynamics. They pay attention to the roles and dynamics within their social circles and are interested in the names and identities of individuals around them.

The character profile for Cluster #8 describes a reserved and observant individual who is attentive to interpersonal relationships, including marital dynamics and social situations. They prefer to communicate through thoughtful conversations and written means, and they value maintaining social harmony.

Reserved Person:

This person tends to be reserved in their conversations, thoughts, and interactions with others. They may carefully choose their words and be selective in expressing their emotions and opinions. They prefer to communicate through written means such as letters and may value their privacy in social interactions.

Observant Person:

This person pays close attention to the names and identities of individuals around them. They are observant of the relationships between people and may find significance in social events. They may prefer to observe rather than actively participate in social gatherings.

Harmonious Person:

This person values harmonious interpersonal relationships and seeks to maintain balance and peace in their interactions. They are mindful of marital dynamics and the dynamics within social situations. They may prioritize maintaining social harmony over expressing negative emotions.

Overall I was very please with the final characterization, I believe I saw a lot of references to the characters in Orton and Wilde's play, for example, in Orton's play, What the butler saw, the story is about to unveil the dark secret of Dr. Prentice, which fits in well with the cluster#6 The Guarded Character character type in cluster#6. In Wilde's work, the many elegant and beautiful male characters are very suitable for the Harmonious Person in cluster #8.

Limitation

In particular, due to the patriarchy, sex-related violent scenes or topics were discarded throughout the data processing. I assume this is caused by the low frequency of related emotional expressions or token occurrences, which does not imply that these topics are unimportant, but on the opposite, they are probably the climax of the drama, but are eliminated as outlier or noisy by the machine learning algorithm. This is unfortunate to me because sexuality, violence, madness, and rebellion against the patriarchy have always been identified as a central themes toward Orton and Wilde's plays.

Reference

Bull, John. "'What the Butler Did See': Joe Orton and Oscar Wilde." *Joe Orton: A Casebook*, edited by Robert V. Green, Routledge, 2003, pp. 45-60.