**DS5559 Exploratory Text Analysis Project:**

**Victorian England in Charles Dickens’ Works**

Hanyu Li

1. **Introduction**

Charles Dickens is considered the greatest English novelist of the Victorian era. As the first mainstream 19th-century writer to reach out to hundreds of thousands of lower class semiliterate readers, he enjoyed a wide popularity during his lifetime. The most abundantly comic of English authors, he was much more than a great entertainer. The range, compassion, and intelligence of his apprehension of his society and its shortcomings enriched his novels, making him both one of the great forces in literature and an influential spokesman of the conscience of his age.

In this project, the corpus used for analysis is a collection of Dickens’ works, including about 20 novels. Varied machine learning methods are applied to accomplish the goal of exploratory text analysis on Dickens’ works. Dickens’ works are taken as both historical and cultural references for analyzing the characteristics of Victorian England society.

1. **Data**

The corpus contains following works of Dickens: *A Christmas Carol*, *A Tale of Two Cities*, *Barnaby Rudge*, *Bleak House*, *David Copperfield*, *Dombey and Son*, *Great Expectations*, *Hard Times*, *Little Dorrit*, *Martin Chuzzlewit*, *Master Humphrey’s Clock*, *Nicholas Nickleby*, *Oliver Twist*, *Our Mutual Friend*, *The Chimes*, *The Old Curiosity Shop*, *The Pickwick Papers*, with 60000 or more paragraphs as documents.

1. **Product**
2. Topic Model Results

Result 1:



Result 2:



Result 3:



The three results above are picked as typical representatives for three social classes in Victorian England from Dickens’ point of view.

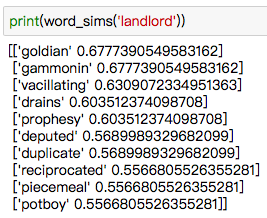
The first model which starts with “prison” includes negative places such as prison, jail, yard, cell, and identities that often involved in violent activities like prisoner, soldiers, rioters, hangman, cap, spy. This is perhaps the darkest side of the society, which is always filled with riots and fights, crimes and tricks, blood and violence. One interesting thing is that “French” also appeared in this model, and it could be an indication of England’s attitude towards France back then.

The second model contains the majority, which are the people who make a living by labor. Some of them might be among the most pitiful group, and this is probably the reason that “orphan” appears here.

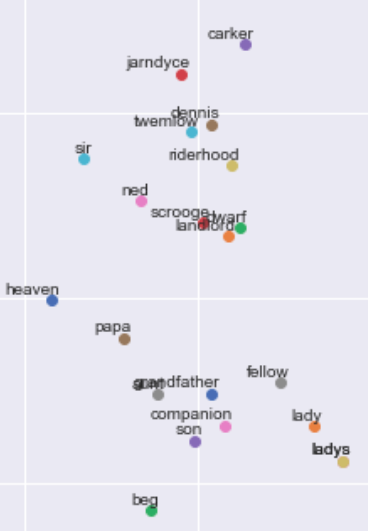
The third model could be regarded as a reflection of the so-called upper class. The words appear in this model turn to have more positive elements such as honour, distinguished, admiration, precious, etc.

1. Word Embedding
2. PMI & SVD

The word embedding results through PMI and SVD provide us with words’ most similar companies and some experiments show meaningful results.

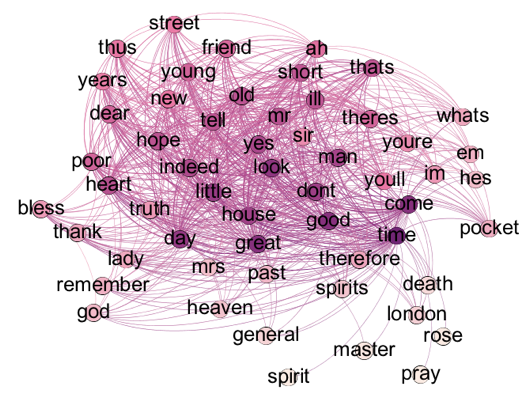
According to the analysis product, for the corpus used for this project, the word “child” has following similars: unborn, nominating, beggarman, impressionable, bewails, theeking, grubby, quelling, queller, reverenced. From the words such as unborn, beggarman, bewails and grubby, it is not hard to imagine what children’s life was like in the Victorian England. Just as Dickens described in his works, kids were experiencing hardship and ugliness starting from their childhood. For the word “landlord”, the result is: goldian, gammonin, vacillating, drains, prophesy, deputed, duplicate, reciprocated, piecemeal, potboy.

The word similar report results are shown as follows. For “woman”, it gives TINY, LOVELY, GROOMED, GIVETH, OLD, BIGODD, SEDUCES, RIDDED, TRIPARTITE.

1. word2vec

The method of word2vec also reveals the relationship between words. It should be noticed that the word “beg” appears very closely with words like grandfather, son, papa. And even though, in sense of POS, characters like Scrooge and Dennis are the same as the words mentioned above, they obviously are divided into two different groups in the plotted image.

1. Word Networks

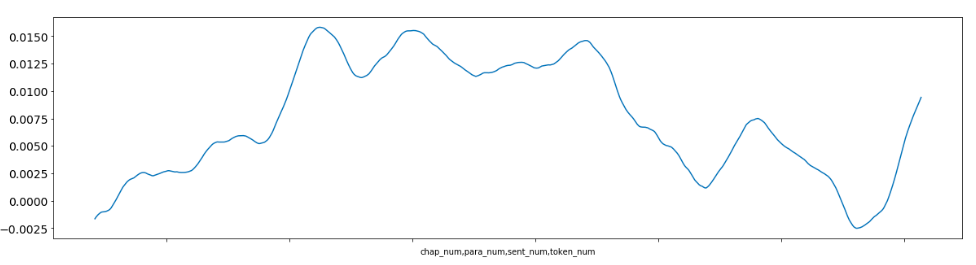
In this project, word network is built with only words that annotated as nouns in the corpus. The central part of the network contains following words: yes, sir, man, mr, etc. On the edge of the network are words including heaven, god, death, London, master, street, spirits. It could be assumed that religion had noticeable effects on Victorian England people’s daily life that words like heaven and god play similar roles as London, master and street.

1. Sentiment Analysis

Great Expectations is selected for sentiment analysis for the reason that it is believed to be one of the best among his works, and the leading character Pip experiences life in different class of society, from an orphan dreaming of becoming a blacksmith to a gentleman.

The result of multinominal emotions shows the NRC lexicon terms in following order: trust, joy, fear, sadness, anticipation, anger, disgust, surprise, etc.

As the analysis reveals, there’s a sudden deterioration in the narrative before the end of story, which is very likely the part that Pip found out the truth of his sponsor and his lover Estella married someone else.



1. **Conclusions**

According to the product of exploratory text analysis, it is reasonable to infer that, in most of Dickens’ works describing Victorian England, the socially vulnerable groups were living a miserable life in an unstable society. However, they still kept