Computational Literary Analysis of Shakespeare

Mark Martinez Thomas Schaffner

Introduction

Some subjects in the humanities traditionally avoid computational methods. We believe modern probabilistic modeling techniques can be effectively applied to historically non-computational fields. Specifically, we will use the topic modeling technique of Latent Dirichlet Allocation (LDA)[1] to analyze the themes of a coherent corpus of literary works. We hypothesize that probabilistic models can identify topics associated with the main themes of Shakespeare's plays. By breaking plays down into their act- and scene-based structures, we examine the relationship between individual characters and topics.

The complete works of Shakespeare can be found on a website hosted by MIT.[2] After removing all non-play works from the text file, we are left with 773 individual scenes which we will treat as separate documents for the topic modeling process. To separate the file by scenes, we manually identified extraneous information (such as cast and copyright information) and labeled scene headings.

Methods

We aim to analyze the effectiveness of Latent Dirichlet Allocation topic modeling as a means to identify themes in literary works. More specifically, we will apply LDA to the plays of Shakespeare. While Shakespeare was a prolific author, the total corpus size is relatively small and manageable from a computational perspective. After basic preprocessing, we will break down the plays into individual "documents" for topic modeling analysis in two ways. We will examine the more than 750 scenes appearing in Shakespeares plays. The raw text for these breakdowns was obtained from a single Etext file from Project Gutenberg and World Library, Inc.[2] Once we identify topics and their representations across each document, we then hope to compare the most highly represented words of the most highly represented topics with common themes in Shakespeares plays to evaluate whether the topic model accurately extracted textual themes. Using Natural Language Processing to gauge things such as sentiment is regularly done in todays world of hashtags and tweets. K-CAP '03 Proceedings[3] of the 2nd international conference on Knowledge capture dealt with using full text documents to gather intricate sentiment details from specific parts of text instead of classifying an entire document as positive or negative. More specific projects using NLP on literature such as "Did Shakespeare Write Double Falsehood? Identifying Individuals by Creating Psychological Signatures With Text Analysis, by Ryan L. Boyd, James W.Pennebaker" try to validate authorship of certain Shakespeares texts. Similarly to these works we will be using NLP to see the cohesive connections that make Shakespeares texts unique amongst his corpus of works.

Model. We use Latent Dirichlet Allocation to create our topic model. LDA considers documents as bags-of-words, and produces unlabeled "topics" represented by distributions over all words in the corpus. For example, a single topic may have the majority of its probability mass distributed over specific and similar words like "love", "marriage", etc. LDA also associates each document with different topics in different proportions. Because the topics are unlabeled, they will require some user interpretation in the context of literary analysis. While a topic like this example seem easily interpretable, topics in general have no guarantee that they will be nearly this easy to interpret. The model also takes as a parameter the number of topics to identify in the corpus. There is often no obvious "good" number of topics, so the selection of this parameter will likely require some trial and error to arrive at topics of decent detail and meaning.

Data and Preprocessing. The text file obtained from Project Gutenberg[2] is a plain text file that requires preprocessing before it is usable as a topic modeling data source. To begin with, we manually annotated section breaks to account for differences in formatting throughout the document. We also denoted extraneous information for removal (such as copyright information) and removed all non-play works of Shakespeare. Next, we pro grammatically broke the text file into individual scenes (using the manual annotations) and performed standard natural language preprocessing techniques; we tokenized and stemmed all words using Python's NLTK package and the Porter Stemmer. Finally, we identified a cast of character names throughout the plays programmatically, which we then validated manually. These names were also removed from the text before fitting the LDA model.

Evaluation. As topics themselves are somewhat subjective, we provide graphs showing the probability mass distribution across the top 10 words for each topic. Some of these topics are more clearly interpretable (**TODO: Examples**) than others. These topics by themselves, while interesting, have no inherent notion of their importance or relevance to the literary structure of the plays. We therefore leverage the structure of the plays to compare against metadata. Specifically, we calculate the Pearson correlation coefficient between each character and topic across all scenes. We then use Benjamini-Hochberg correction[4] to calculate the false discovery rate for each coefficient. Here we report some of the most highly correlated pairs. In conjunction with

Results

As a preliminary result, we are able to fit a simple LDA topic model to the corpus of Shakespeares plays. We break down each play by scene into multiple documents, then apply basic preprocessing and text-cleaning techniques before fitting our model. We have tried several different settings for the number of topics. Before we removed all character names in the preprocessing step, individual topics considered play-specific character names to be highly weighted. One topic might highly weight characters from Romeo and Juliet, for example. Reported here are the probability mass distributions across the top 10 terms in each of several topics of a 25-topic model. Figure 1 shows the probability mass distributions for several topics. Topic 12 seems to refer to English royalty or some combination of England and royalty. Topic 13 seems to refer to Rome. Topic 14 fairly clearly refers to "love", which is encouraging to see given how many of Shakespeare's plays have to do with love. However, topic 23 seems to be fairly associated with murder and violence ("sword", "hurt"). Note also that "enter" is extremely prominent in topic 24. We believe this is because we have not yet fully removed all stage direction. Therefore, "enter" appears very frequently across all plays and characters enter each scene.

Additionally, we include here a graphic of the correlation between certain characters and topics. We will improve the graphic (isolating highly correlated pairs) and address explicitly the corrected false discovery rates. However, at this time, we have not been able to generate a clean graphic addressing these. Therefore, we have subsets of the full heatmap of Pearson correlation coefficients. In figure 2, we see in particular a high correlation between "capulet" and Topic 23, associated with murder and violence. "Capulet" is a family name in *Romeo and Juliet*, a play that certainly involves plenty of death and violence. "Chiron" is also highly correlated with Topic 13, associated with Rome. "Chiron" is a character in the play *Titus Andronicus*, which is set in Rome and includes characters in influential Roman positions. We see in Figure 3 that "Romeo" is also highly correlated with Topic 23. These examples show significant promise for the topic model, although we still want to perform more formal validations (see Discussion).

Discussion

Direction. There are several directions in which we could move forward. The input data themselves are subject to interpretation and opinion when it comes to preprocessing. Currently, we remove stop words and represent each document as a simple bag-of-words. However, these decisions should be examined and tested, as we have no reason to believe this setup might be optimal. Additionally, we are initially comparing all of Shakespeares plays as a single corpus. Splitting the plays by genre may yield more accurate or interesting results.

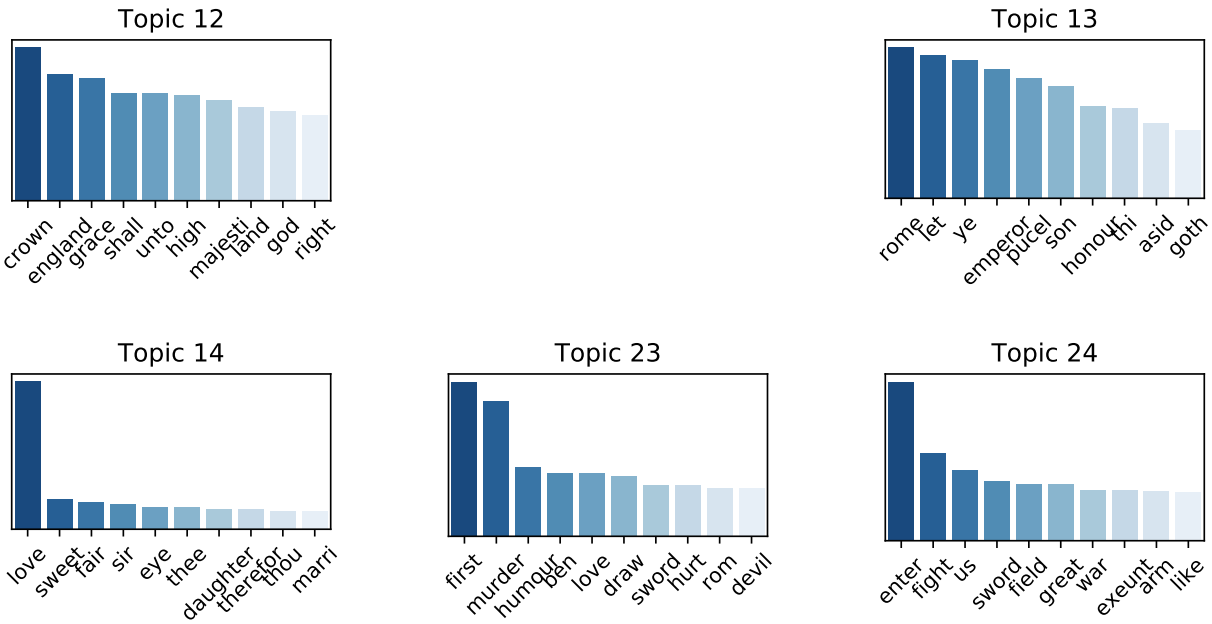

Figure 1: Probability mass distribution for top 10 words of several topics

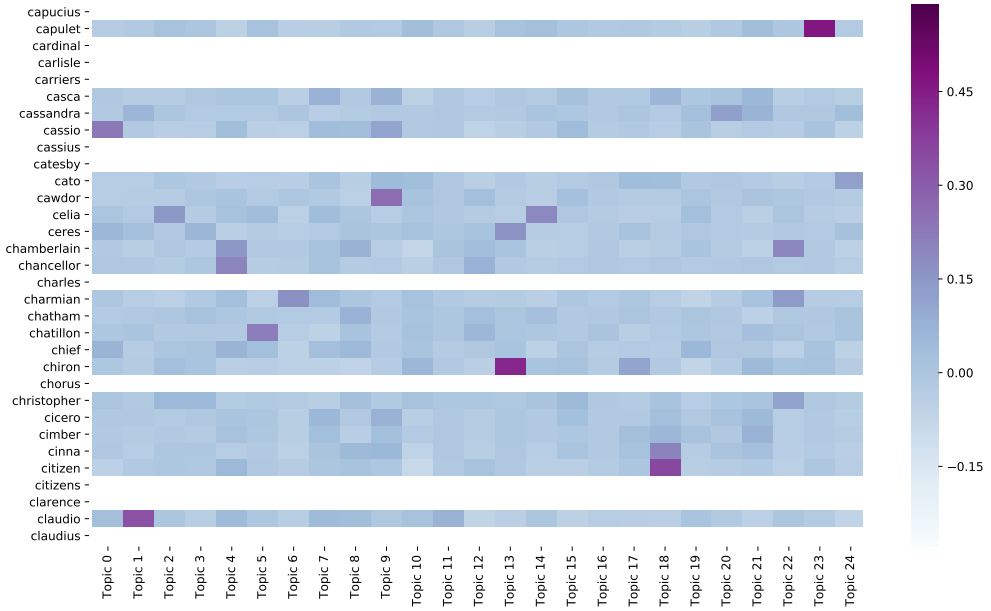

Figure 2: Heatmap showing the correlation between topics and characters. Note in particular “chiron” and “capulet”.

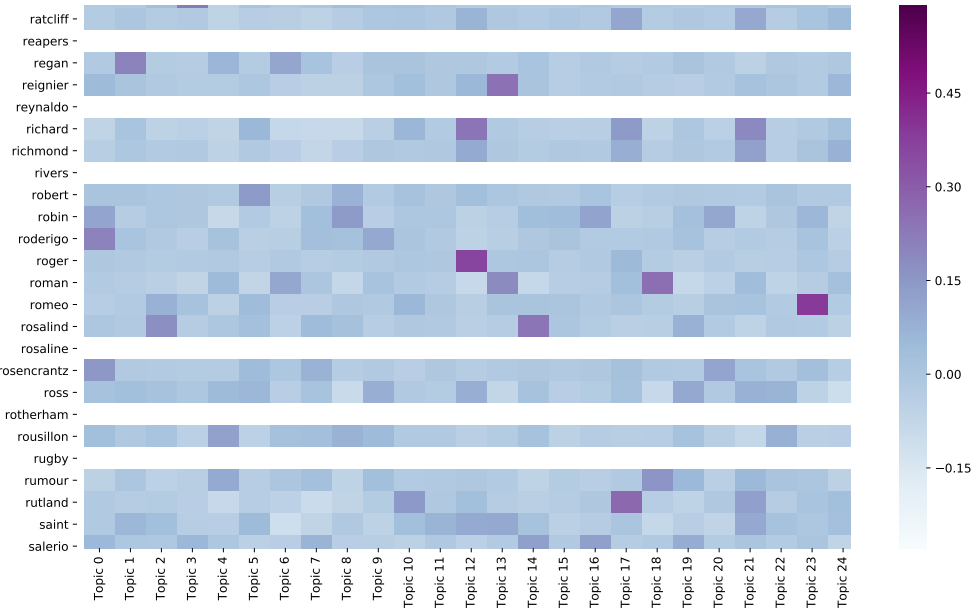

Figure 3: Heatmap showing the correlation between topics and characters. Note in particular "romeo"

Stemming. Common word stemming methods, including the Porter Stemmer provided by Pythons natural language toolkit, are geared towards modern English. However, many conjugations and terms in Shakespearean language follow now-archaic rules (e.g. thou and hath).

Evaluation. Evaluating a topic model in general can be difficult, and evaluating the meaning of specific topics can become subjective and sensitive to the number of topics chosen. In order to approach the model with some amount of ground truth for comparison, we will use the play metadata available to us. As a preliminary test, we will fit a 3-topic model to the corpus. We can then test the correlation between the topic distribution for each scene and the genre of that scenes play (comedy, tragedy, or history). Moving to more complex models, we plan to compare character line count with topic distribution for individual scenes in order to identify topics or themes associated with characters.

At the moment we have several ideas to extend the project from where its currently at. Some of these ideas include:

- Calculation of correlation between topics and whole plays
- Seeing the biggest contributors to a specific topic. If royalty are the only contributors to topics like Greed it would be interesting to see how this could fit in with Shakespeares commentary on his own time
- Adding a play from a completely different author and seeing if its possible to detect this new play using LDA
- Using stage directions to compare the themes in the play to see what is going on compared to what we find in the topics for certain plays
- Adding a time component so that we can see how one topic folds into another one and if there are common pairings, such as love followed by jealousy.
- Evaluate the stability of topics across multiple LDA fittings

References

[1]Blei, D. M., Ng, A. Y., and Jordan, M. I., "Latent Dirichlet allocation," *Journal of Machine Learning Research*, Vol. 3, No. Jan, 2003, pp. 993–1022.

[2]Gutenberg, P., "The Complete Works of Shakespeare," `http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm`, 1994.

[3]Rector, A., "K-CAP03: Proceedings of the 2nd international conference on Knowledge capture," 2003.

[4]Benjamini, Y. and Hochberg, Y., "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, 1995, pp. 289–300.