1. A description of your problem and motivations.

If the goal of Natural Language Processing is to find ways of making computers understand human language, then the use of human language as art will probably be its hardest challenge as well as greatest achievement. Because of the multidisciplinary nature of the field, I’m interested in looking at NLP’s intersection with the artistic application of human language in the literary realm and how NLP can contribute to the literary analysis realm. Being a member of the Wellesley College Shakespeare Society, I naturally found myself drawn to Shakespeare as a literary source. The common availability of resources on the topic and structured format of plays makes Shakespeare’s works an ideal candidate for this cross disciplinary project. For the rest of the semester, I’ll be investigating and implementing NLP and data science techniques to analyze the primary characters of Shakespeare’s plays. The goal is to develop a computational method that categorizes Shakespeare’s characters into personae sets based on characters’ similarities.

1. A brief survey of existing work.

In the intersection between literature and NLP, there have already been projects done which focus on author identification. That is, given a sample text, a prediction is made about which author would have written it. A couple of recent projects have pushed this idea further and implemented a personality prediction of each author given the author’s works[[1]](#footnote-2). With the topic of Shakespeare’s works specifically, there are a lot of general, high level projects that have been funded in the past, including the 1994 lexicographic “Shakespeare Database Project”[[2]](#footnote-3) and the Open Source Shakespeare[[3]](#footnote-4). Both of these should prove to be good data resources for me during the semester. I also came across a specific project focusing on Shakespeare’s works which analyzes sentiment dynamics between four different pairs of major characters (Hamlet and Gertrude, Othello and Desdemona, Romeo and Juliet, as well as Petruchio and Katherine)[[4]](#footnote-5). The specific niche of investigating the personae of Shakespeare’s characters is a yet untouched topic of interest. The closest work related to my topic of interest would be two projects out of Carnegie Mellon University, the latter in collaboration with University of Illinois. In the first work, the team used NLP and statistical techniques to categorize 100 different character personae that exist in films across various genres[[5]](#footnote-6). The second is an expansion on the first project, in which character personae identification is run on literary works published in the 18th and 19th century, but in which each identification is purposely made independent of authorial diction so that character similarities across various authors can be identified.

1. A description of your proposed solution(s), including the data and tools you will be using.

Although I’ve only briefly investigated the resource, Open Source Shakespeare (OSS) does have a database that the public can download and use for non-commercial purposes. There are also a couple of Github repositories that are the users’ personal extension work on the OSS database, including a promising looking script that can parse play formatted texts. If this does not pan out, I plan on pull straight from the Gutenburg Project and make use of some of the resources cited in both Bamman et al.’s works to do some basic tagging, parsing, and processing.

In terms of developing the specific analysis techniques, I intend to approach this project from a Data Science angle -- focusing on iterative development and accuracy improvement. The accuracy will come from a combination of existing seminal works in the Shakespearean scholastic field as well as Shakespeare’s own continuity of character (Falstaff shows up in 4 plays). In Bamman, O’Connor and Smith’s work on film character analysis, they proposed an interesting approach of categorizing characters based on the latent topics associated with each character’s verb governors. From there, each character is clustered and normalized with other characters that shared the same latent topics. I’ll have to spend a bit more time understanding their procedure better, but I think that this, paired with many of the topics we’ve covered in class so far would be a good starting point for a first iteration.

1. Description of work you have already completed. This section can be short but not empty. Plan to start reading the background literature and brainstorming before you submit the proposal.

At this moment, I’ve read through the six related papers and projects mentioned above. I’ve also reached out to a couple of Shakespearean sources for advice on seminal works in their fields. Folger Shakespeare Library has responded with a couple of books as well as a link to the OSS, of which its duplication at least proves the project’s legitimacy if nothing else. Professor Yu Jin Ko from the English Department has also suggested several related seminal works. Of the resources I’ve been able to get my hands on, most have proved to be quite useful in the traditional, literary analysis sense. I am currently investigating the details of the various sources of preprocessed Shakespeare work on OSS and on Github.

1. Three milestones. Be realistic about the milestones, and yet ambitious.
   * What to complete by April 4th when the progress update is due

By April 4th, I would like to have the data fully processed and some initial, basic statistical description done on the major characters. Worst case scenario, if the preprocessed databases don’t pull through, I still want to have the data mostly processed and in a usable format. The processing will involve play separation, character separation and tagging (since unrelated characters from two plays can share the same name), as well as minor character removal.

* + The minimum desired outcome of your project by the final submission on May 12th

By May 12th, I expect to have a program that can identify similar characters. The specific technique used to associate may depend. It could be the case that multiple answers are fed to the program and the closest character or topic should be returned, or it could be that two characters are inputted and a binary or probability answer on their similarity is returned.

* + The ideal final outcome of your project

Ideally, the program will not need prompting -- having done the analysis already -- and can output the various sets of characters that are most similar. These sets would somewhat resemble reasonable categories, as evaluated based on what Shakespearean scholars have written. For example, I would expect Hamlet and King Lear to be grouped together.

1. Kim Luyckx and Walter Daelemans. 2008. Personae: a corpus for author and personality prediction from text. In *International Conference on Language Resources and Evaluation*; Marc Pickett, Chris Tar, and Brian Strope. Feb.24, 2016. On the Personalities of Dead Authors. Google Research Blog. [accessed Mar. 1], 2016. http://googleresearch.blogspot.com/2016/02/on-personalities-of-dead-authors.html [↑](#footnote-ref-2)
2. http://www.shkspr.uni-muenster.de/index.php [↑](#footnote-ref-3)
3. http://www.opensourceshakespeare.org/ [↑](#footnote-ref-4)
4. Eric Nalisknick and Henry Baird. 2013. Character-to-Character Sentiment Analysis in Shakespeare’s Plays. In [*Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*](http://aclanthology.info/volumes/proceedings-of-the-51st-annual-meeting-of-the-association-for-computational-linguistics-volume-2-short-papers) [↑](#footnote-ref-5)
5. Bamman, O’Connor, Smith. 2013. Learning Latent PErsonas of Film Characters. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)* [↑](#footnote-ref-6)