1. Problem Overview: Describe the problem you are solving; what makes it interesting?

The proposed project is a genre and authorship prediction of the document. Our solution can be used in many areas like an email filter, plagiarism detection and to check the originality of the content.

1. Data: Which data will you use? How will you collect it? What problems do you anticipate?

We planned to scrap news articles from Chicago sun times and The New York Times. Following are the problems we will face during data collection first identifying correct markups to filter in the html document, identifying author for each article, additional delays enforced by websites for data collection.

1. Method: What method or algorithm will you use? Will you use an existing library to do so? Do you plan to modify the code at all?

We will use one model for genre detection and another for authorship detection. We plan to use the POS tagger and discriminant analysis method for genre detection. Additionally, we will use a multiple discriminant analysis method for authorship detection and we would be implementing those algorithms.

1. Related Work: List at least 5 references (with links) to research papers that are related to your project (use Google Scholar to search).

* <http://www.jait.us/uploadfile/2014/1223/20141223050800532.pdf>
* <http://www.mitpressjournals.org/doi/pdf/10.1162/089120100750105920>
* <http://www.aclweb.org/anthology/C94-2174?CFID=897164713&CFTOKEN=27774739>
* <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.67.6480&rep=rep1&type=pdf>
* <http://ceur-ws.org/Vol-502/paper11.pdf>

1. Evaluation: How will you evaluate your results? What baseline method will you compare against? What are the key plots or tables you will produce? What performance metrics will you use? What descriptive evaluation will you do (e.g., look at specific predictions made by your system; visualizations)?

After data collection, we will keep aside some articles and run through our proposed solution. We will compare the results and provide the error rate, this metric will provide the effectiveness of our approach. Additionally, we can test our proposed solution to articles of different length/words count to see the change behavior, this would give us the Text length related to accuracy for the author identification and genre detection plots.