Marcus Blaisdell
Cpt_S 315
Course Project Proposal
March 30, 2018
Professor Doppa

1) Data Mining Task: What is your data mining task? This task could be a series of exploratory questions that you want to investigate or analyze. What is your motivation behind choosing this problem for your project?

I am choosing to work on the Kaggle project: “Spooky Author Identification”. That is a challenge to be able to distinguish sentences as being from one of three notable horror authors.
I am intrigued by the idea of being able to use machine learning to distinguish writing styles. I have long been interested in writing an AI capable of recognizing natural language, including nuance, and this would be a great first step towards that.

2) Dataset: What is the source of your data?

Kaggle provides one set of training data that includes labeled sentences from each of three authors and one set of testing data that I can use to submit to Kaggle for evaluation.
Each of the three authors also have works in the public domain so I was able to download the text versions of one of each of their works from Project Gutenberg that I can parse into sentences to build my own additional training and testing sets.

3) Methodology: How will you solve the data mining task? You should have some idea of the algorithms or software tools you plan to investigate.

An initial scan of the sentences suggests that use of punctuation is highly indicative of at least one particular author so the analysis will have be more than a bag of words but needs to include that as well. The bag of words is a logical starting place though.
While reading through some of the Kaggle users comments, it seems that Kernels will be the key to success so I will be experimenting with them as well.
I have a list of nouns, verbs, and adverbs from:
https://www.worldclasslearning.com/english/list-of-verbs-nouns-adjectives-adverbs.html
that I can use to label individual words from the texts to further distinguish them, that is, to add another feature to the classifier to help distinguish the author.
While trying to compare the tools mentioned, xgboost seems to be preferred by most Kaggle users so it seems a good choice.

4) Final product: What will be the outcome of this project? How will you measure the success of your course project? Will this project help you explore or learn something new?

The final product is expected to be a function that can distinguish among three specific authors with a high degree of success and a generalizable concept that can be applied to other authors of other works.