Matthew Miller

BIA686

“I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.”

Project Proposal:

Gendered Ambiguous Pronoun Kaggle Competition

Research Question:

Can NLP and Machine Learning be used to create a model which will successfully classify pronouns as being associated with a specified subject.

Introduction and Problem Background:

Pronoun resolution is an important task in natural language processing to better make sense of sentences from the viewpoint of a computer. For example, take the sentence “*Kara*, after running into her next door neighbor, *Susan*, began to walk down the block with **her** dog.” On the surface, the word “her” could refer to Kara or Susan. In context, it makes more sense that the dog belongs to Kara. When processing sentences, situations like this continue to occur, and recently Google has revealed that success in this field has been marred by the uneven distribution of gendered examples, typically skewed in the direction of masculine pronouns. This led to the creation of their Gendered Ambiguous Pronouns (GAP) dataset, which has an even distribution of masculine and feminine pronouns to classify.

Results from Exploratory Analysis:

From the exploratory data analysis, I believe there is some proof that the type of pronoun and type of sentence can be important in identifying which subject a pronoun refers to. Pronoun types include subject pronouns (he/she), object pronouns (him, her), and possessive pronouns (his, her, hers), while sentence types are based on whether the pronoun appears before either subject, between them, or after them in the sentence.

Data Mining and Analysis Methodology:

Going forward, I plan on adding more features using word vectors, lemmatization, and other NLP tools to better prepare the data for data mining methodologies. Some specific ideas include finding the offset between the pronoun and subject in words rather than characters, finding how often each subject appears in the excerpt, and generally using a Count Vectorizer to see if certain words are associated with certain subjects. These features, as well as those explored in the initial analysis, will be used for data mining techniques such as SVM and RF. I also plan to explore stacking to get better predictions overall.

Schedule:  
The Kaggle schedule itself says to have a model and submission file submitted for stage 1 by 4/15, at which point stage 2 begins, which requires a final submission by 4/22.

Moving forward, I’d like to have the feature creation and first stage of data mining done by April 8, the deadline for the data modeling assignment. This leaves another week to do the stacking portion, as well as ready the model for the Kaggle submission.