**Final Report Prompt**

Intro: 1 – 2 paragraphs, **1 – 1.5 pages**

Feature creation and filtering: 4-5 paragraphs, 2 – 3 pages writing, 1 page visuals **3 – 4 pages**

Random Forest: 2 paragraphs, 1 – 1.5 pages, 1 page visuals **2 – 2.5 pages**

SVM: 2 paragraphs, 1 – 1.5 pages, 1 page visuals **2 – 2.5 pages**

K–means: 1-2 paragraphs, 1 page, ½ page visuals **1.5 – 2 pages**

? Selection of model: 1 – 2 paragraph, **1 – 1.5 pages**

Conclusion: 1-2 paragraphs, **1 page**

(1 to 1.5) + (3 to 4) + (2 to 2.5) + (2 to 2.5) + (1.5 to 2) + (1 to 1.5) + (1) = **11.5 – 15 pages**

------------------------------------------------------------------------------------------------------------------------------------------

Intro (purpose):

* Purpose/relevance/significance

Intro (methods):

* Which methods from class were excluded (Linear, LASSO, etc) and **WHY**
* Which methods from class were used (RF, SVM, etc) and **WHY**

------------------------------------------------------------------------------------------------------------------------------------------

Parsing:

* Combined list of stop-words to remove (remove words)
* Removed punctuation, numbers, single/two letter words, and white spaces (remove extra)
* Removed words closely tied to subject line (exclude redundancy)

Word selection:

* Corpus functions utilized
* Using trial and error (**GRAPHS AND TABLES NEEDED HERE)** we decided on **X** for sparsity
* Compared results from different samples and different sparsity levels, as well as with the test set to exclude uncommon words **(EXAMPLE NEEDED)**

Problems:

* Some bugs in code with word removal but easily fixed
* **CREATED FUNCTION NAME??** so that the new data could be turned into a matrix with the exact columns required by the model (i.e., the same ones as the train set)

Power Features:

* Utilized non-parsed data
* Used features such as punctuation/number usage, email length, number of words, average length of sentences, etc **(HELP NEEDED FOR THIS, KUANG?)**
* Combined two matrices

Problems:

* ???????

------------------------------------------------------------------------------------------------------------------------------------------

Selecting Forest:

* Cross validation used/train, OOB, individual rates **(PLOTS AND GRAPHS NEEDED)**
* Used “importance” feature in rf, (also needed for K means later on
* **TOP TEN FEATURES????**
* ?????

Problems:

* Had to adjust matrices (see above)
* Had to control for column names
* ?????

------------------------------------------------------------------------------------------------------------------------------------------

Selecting SVM:

* Kernel chosen based on trial and error **(TABLES/EXAMPLES NEEDED)**, clearly not polynomial or sigmoidal
* Used CV/train, OOB, individual rates **(PLOTS AND GRAPHS NEEDED)**
* **TOP TEN FEATURES????**
* ?????

Problems:

* Less accurate than rf overall **(SEE PLOTS AND GRAPHS)**
* Difficulty determining linear vs radial **(I.E., linear less biased, radial less variance)**
* ?????

------------------------------------------------------------------------------------------------------------------------------------------

K means:

* CV used for nstart, OOB, individual rates **(VISUALS NEEDED)**
* Lower accuracies **(expected?)**
* ?????

Selection of model:

* Summary of which models were the least biased (pure train)
* Summary of which models were the least varied (CV and validation)
* Explanation of why the models that were the least biased were
* Explanation of why the models that were the least varied were
* Why we decided on the model we decided on
* Why we believe the model we chose will perform well

------------------------------------------------------------------------------------------------------------------------------------------

Conclusion (Analysis):

* Potential setbacks we could have controlled for (ex: over and under-fitting, better power features)
* Potential areas of improvement we could not have controlled for (ex: Gradient Boosting, other models?)

Conclusion (Conclusion):

* Potential implications of our code
* How the code can be implemented in other areas
* Final thoughts?