1. What progress you have made towards your proposed goal? (just data collection is not an option)

* Preprocessing of the data set into a single coherent file.
* Basic visualization of the data via histograms
* Methods for transforming content into vectors

2. If you tried some basic approaches: what worked well and what did not?

* We’ve attempted to use the elbow method to determine that the correct number of clusters should be 20. This doesn’t seem like it should be the case however and need to validate our results.
* We plan on using k-means++ or some similar algorithm to perform clustering on the data.

3. What could be done to improve the basic approaches?

* We can use cross validation to determine the correct k-gram parameter to use as well as the correct number of clusters. Currently our method to determine the ideal cluster size runs on a random 5% of the data and so cross validation could be used to improve that accuracy.

4. What experiments have you run and are you planning to run to demonstrate the effectiveness?

* Aside from visualization and the method to determine the ideal cluster size we haven’t ran any experiments yet. We plan on running a clustering algorithm on the data and plot the high-density clusters as histograms to see if these clusters correlate at all with time.
* We may also be able to use logistic regression with the cluster labels and the account and publish date features to fit a more robust model.