Daniel Oliveros

CS 4320

**K-means**

**In order to implement K-means I did the following:**

1. Created an inverted index to determine which documents had given words.
2. Made an incidence matrix based on the inverted index
3. Calculated the tf-idf values for each entry in that matrix
4. Randomly selected a group of k vectors to be the centroids for each cluster
5. Determined which centroid each given node was closest to
6. Grouped all nodes into new clusters and calculated the new centroids as well as RSS
7. Went back to step 5 until RSS converged
8. Returned all the final clusters

**What is your R value?**

I choose r to be the number of documents, so that I could reduce the odds of finding myself stuck with bad seeds. This value is the same as the number of documents so it should allow for good certainty.

**What are your stopping criteria for K-means?**

Once the difference between the next RSS and the previous one was less than 10% of the previous RSS I decided to terminate the process. This would mean that they have started converging onto clusters where there isn’t much of an improvement over time.