**Assignment Ten**

Jacob Berlin

CS432 – Spring 2016

1. Using the data from A8:

- Consider each row in the blog-term matrix as a 500 dimension vector, corresponding to a blog.

- From chapter 8, replace numpredict.euclidean() with cosine as the distance metric. In other words, you'll be computing the cosine between vectors of 500 dimensions.

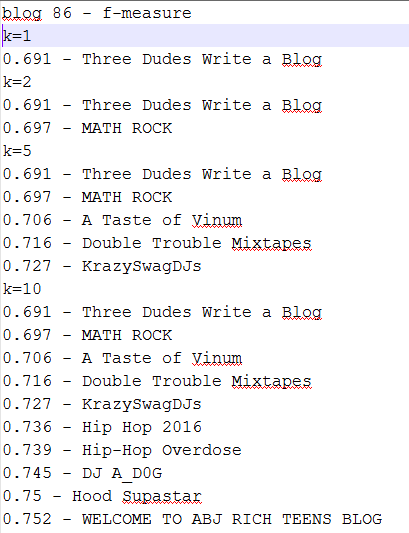
- Use knnestimate() to compute the nearest neighbors for both:

http://f-measure.blogspot.com/

http://ws-dl.blogspot.com/

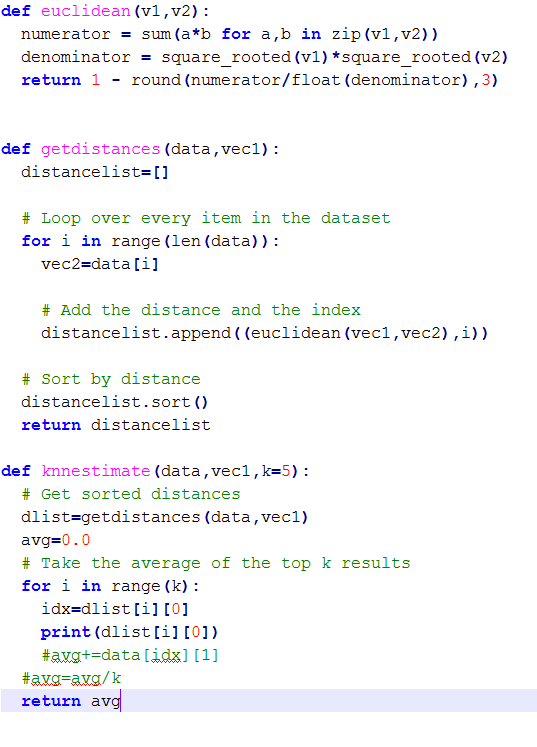
for k={1,2,5,10,20}.

To start this problem, I modified Programming Collective Intelligence’s version of ‘numpredict.py’ and placed that code into ‘clusters.py’. To make this work, I modified the Euclidean function to calculate the cosine distance and then modified the knnestimate to allow blog entries to be calculated. This is the output and code that I received/made:









2. Rerun A9, Q2 but this time using LIBSVM. If you have n categories, you'll have to run it n times. For example, if you're classifying music and have the categories: metal, electronic, ambient, folk, hip-hop, pop you'll have to classify things as: metal / not-metal electronic / not-electronic ambient / not-ambient etc.

Use the 500 term vectors describing each blog as the features, and your manually assigned classifications as the true values. Use 10-fold cross-validation (as per slide 46, which shows 4-fold cross-validation) and report the percentage correct for each of your categories.

For this question, I ran the four categories that I had through LIBSVM –

Percent Cross-Validation Percent Correct

1. HowTo – 48% 32%

2. Competitions – 83% 85%

3. Sales – 64% 58%

4. Information – 76% 52%