Hi Professor,

This is Issiah from your machine learning class. I hope you are doing well and not too stressed. I just wanted to give you my project proposal.

For my project I am thinking of making a basic search engine with PCA. I am thinking of either showing a demo of it working in class or a scatter plot of the web pages along the principle components or maybe lists of the most relevant words that the different components represent. To do this I plan on using lots of web pages as input data(I have ~5000 I will probably try to get more.)

Another improvement that I am thinking about is using another algorithm to do unsupervised parts of speech(grammar tags like preposition,verb,noun, etc.) tagging in order to increase the amount of information associated with each word. The word and the tag will be treated as a single word. My hypothesis is that will be beneficial since unlike PCA it captures the words in context meaning, especially how the word is actually used inside the sentence. Furthermore, my hope is that unsupervised parts of speech tagging, will capture more than just the strict grammatical significance of the words but also some of the more fluid subtleties of the word usages since it is unsupervised and creates its own parts of speech.

To do the unsupervised parts of speech I will use the EM algorithm to train a Hidden Markov Model with web page data. I already have the algorithm written and it is working. The code is actually from another class that you were teaching earlier this year, I recently fixed some issues and the algorithm seems to be converging a lot better.

To do the PCA step I plan on using tf-idf to preprocess the document texts and then use truncated PCA with a Sparse Matrix, which I will do instead of normal PCA to save computer resources and leave open the option to using lots of data. The sci kit learn packages seems to have all of this available(tf-idf, truncated PCA for Sparse matrices).

I have built my own web crawlers but I might change to using pyspider as it seems to be faster and more robust.

Thanks,
Issiah Cantrell