CPSC340: Midterm Answers

Exercise 1.

(a) Average the bag-of-words vectors for all the papers written by that author. (You could just sum them as well, especially if you normalize the vectors in the next step.)

(b) Take all the author vectors, and remove the authors of the paper from the set. Normalize all the remaining author vectors, and the vector for the paper in question, by their Euclidean lengths. Compute the cosine/dot-product similarities between the paper and all the authors (it would be a good idea to include inverse document frequency weights here). Take the three authors most similar to the paper. Note that this is completely equivalent to perform k-NN using the Euclidean distance on *normalized* bag-of-words vectors.

(c) The bag-of-words vectors will have as many dimensions as there are words in the dictionary. This will be very large (tens of thousands). Do PCA on the document bag-of-words vectors and take the first few principal components, say 100. Represent each document only by its projection on these PCs. Now repeat everything as before. This will involve much less storage and computation, and handle correlations among words. This is "latent semantic indexing".

Exercise 2.

(a) For `states.pca.1`, PC1 is almost exactly the state's area, and PC2 almost exactly its population. For `states.pca.2,` PC1 is high for states with high murder, high illiteracy, low education, and little frost. It separates warm, high crime and less educated states from cold, low crime and higher educated ones. PC2 is high for large, populous, rich states, especially if they tend to be more violent and more educated than the norm. It separates big, rich states from small, poor ones.

(b) I would rather use `states.pca.2`. The features are not on the same scale, so `states.pca.1` is very weird. PC1 is almost exactly the state's area, just because the numbers there are immensely bigger than all the other features. (Look at the table of summary statistics.) PC2 is basically population, for the same reason. If I was going to do that, I could just discard all the features except for area and population. `states.pca.2`, on the other hand, is scaled to make the features comparable in size, and shows more interesting patterns.

(c) The biggest labels are in the south-eastern quarter of the USA, with Florida being an exception.

Exercise 3.

(a) Lecture notes.

(b) For T_1: no as, for example, if $\boldsymbol{\pi}_0 = (1\ 0)^{\mathrm{T}}$ then $\boldsymbol{\pi}_{2k} = (1\ 0)^{\mathrm{T}}$, $\boldsymbol{\pi}_{2k+1} = (0\ 1)^{\mathrm{T}}$.
For T_2, yes as 2 is an absorbing state so $\boldsymbol{\pi} = (0\ 1\ 0)^{\mathrm{T}}$.
For T_3, no as the states $\{1,3\}$ do not communicate with $\{2,4\}$.
For T_4, yes as all the elements of this matrix are stricly positive (Perron-Frobenius theorem).

(c) Lecture notes.