Text Mining for the Social Sciences
Assignment 2

Please work in same groups as formed for the first two problem sets. Please upload code for Q1 into your group repositories by 4 May, and turn in written solution to Q2 at the beginning of that day's TA session.

1. Generate the tf-idf-weighted document-term matrix X for a corpus of your choice. Perform a singular value decomposition on it using numpy, and retain a "reasonable" number of singular values (no more than a few hundred) to form the approximate matrix $\hat{X}$.

   Now compare the cosine similarity of documents using both X and $\hat{X}$ in an example of your choosing. Does latent semantic analysis appear to outperform the standard analysis? For example, in the state of the union dataset, one could look at the average cosine similarity within and across speeches made by Republicans and Democrats, and assess whether LSA provided a sharper distinction between political parties.

2. Consider the following variant of the multinomial mixture model. First, each observation i in the data is allocated a latent variable $z_i \in \{1, \dots, K\}$, where $\Pr[\, z_i = k \,] = \rho_k$. Then two separate features are drawn independently. The first is drawn from some $\beta^1_{z_i} \in \Delta^{V_1 - 1}$, and the second is drawn from some $\beta^2_{z_i} \in \Delta^{V_2 - 1}$.

   For example, we could think of consumers as having demand for both food and drinks. z_i could then be interpreted as consumer i's demand type, which defines two probability distributions. $\beta^1_{z_i}$ would be a distribution over V_1 different food products, and $\beta^2_{z_i}$ a distribution over V_2 different drink products. Moreover, conditional on a demand type, the choice of food and drink is made independently.

   Suppose for each observation i we observe the (V_1-length) vector of counts arising from the first distribution. Denote the vth element of this vector x^1_{iv}. Similarly, define x^2_{iv} as the observed count of the vth feature in the second distribution.

   (a) In this model, what are the parameters, what are the latent variables, and what is the observed data?

   (b) Write down the complete data log-likelihood function. Recall that this expresses the log of the joint probability of the latent variables and the observed data.

   (c) Compute the expected value of the above log-likelihood function given fixed values for the parameters. Recall this is the E-step in the EM algorithm. Denote this function Q, and note that it will depend on the parameter values.

   (d) Maximize Q with respect to the parameter values. Show ALL steps.

   (e) Use your answers to the above to write pseudo-code for implementing the EM algorithm for this model.