CS769 Advanced NLP Homework 5

Due 3/24/2009

This homework is light to give you time to prepare your project proposal.

1 PageRank

Let us consider a tiny WWW world with 40 websites. The hyperlinks among websites are as follows:

- Website i points to websites 11, 21, 31, for $i = 1 \dots 10$.
- In addition, website i points to websites $i - 1$ and $i + 1$, for $i = 1 \dots 40$, whenever these neighbors are in $[1, 40]$.

Q1. **[10]** Create the transition matrix P. Note $P_{uv} = P(u|v)$, the probability of going from v to u. Do not consider teleporting yet. You do not need to show P. If a random walker starts from website 1, where will the walker be after one step?

Q2. **[10]** Now assume the walker at any step has probability 0.1 of being teleported to a random website (uniformly in $1 \dots 40$). If a random walker starts from website 1, where will the walker be after one step?

Q3. **[20]** Compute the PageRank of the 40 websites (scale them so that they are the stationary distribution of the random walk).

Q4. **[10]** Plot the PageRanks. Do websites 11, 21, 31 have the same page rank? Explain why.

2 Latent Semantic Indexing

We will use the data at `http://pages.cs.wisc.edu/~jerryzhu/cs769/dataset/wisc/cs.dat`. Use spconvert() in Matlab to turn this into a matrix. Make sure each column is a document – you will need to transpose the matrix. You should have 8218 rows and 1090 columns. Perform latent semantic indexing on this dataset, keeping only 2 columns in the $\hat{U}$ and $\hat{V}$ matrices (hint: use `svds()`). Use $\hat{S}\hat{V}^\top$ as the new document representation, which should now be in two dimensions.

Q5. [**20**] Plot all documents as dots in 2D space using the new representation.

3 Latent Dirichlet Allocation

Download the LDA code from `http://www.cs.princeton.edu/~blei/lda-c/index.html`. Study the readme file. Our department user dataset is at `http://pages.cs.wisc.edu/~jerryzhu/cs769/dataset/wisc/`. The file `lda.cs` contains the documents in LDA format. I have removed the top 50 most frequent word types (i.e. stopwords). The file `user.txt` lists the documents' corresponding user names. The file `vocab.txt` lists the vocabulary.

Q6. [**10**] Run LDA with 50 topics. Use default settings ($\alpha = 1$, seeded). Look for `likelihood.dat` in the model directory. The first column is the log likelihood of the dataset that the LDA model tries to maximize. Plot log likelihood vs. iteration.

Q7. [**10**] The file `final.beta` contains the 50 learned topics, each row is a multinomial distribution over the vocabulary, the values are log probabilities. (Note: The LDA code adds a dummy word type to the vocabulary. Therefore, each row in `final.beta` has $|V| + 1$ columns. The first word type in the original vocabulary is the 2nd column, the second word type is the 3rd column, and so on.) Print the top 10 word types in each topic, sorted by probability. Please print the actual word strings, not the indices. Comment on the topics as you wish.

Q8. [**10**] The file `final.gamma` contains the posterior Dirichlet parameters of those 50 topics for each document (user). Think of them as the mixing weights of the topics for each user. For each of the following users, list the topic numbers that most significantly contribute to their document (i.e., with relatively large weights): jerryzhu, shavlik, miron, sohi, pb.