Natural Language Processing:

Assignment 7: Topic Modeling

Jordan Boyd-Graber

Out: **3. November 2014**
Due: **14. November 2014**

Introduction

As always, check out the Github repository with the course homework templates:

`git://github.com/ezubaric/cl1-hw.git`

The code for this homework is in the `hw7` directory.

This homework is about unsupervised clustering. You'll complete an implementation of latent Dirichlet allocation and run it on real data, discovering clusters of documents in Wikipedia. While there's not much programming you have to do for this assignment (probably less than 10 lines of code), it depends on understanding the rest of the code around it. Don't leave it until the last minute.

1 Implementing a Gibbs Sampler (20 points)

1.1 Changing topic assignments / counts

Finish implementing the `change_topic` function so that it keeps track of the necessary counts so that you can remember the association of terms to topics, documents to topics, and the specific assignments (i.e., you should update three things).

1.2 Computing the sampling distribution

Finish implementing the `sample_probs` function so that it returns a dictionary that provides the conditional distribution of a token. This should be an unnormalized distribution (this will make it easier to debug).

In addition to using the unit tests (as usual), there's also a directory of toy data you can try the code out on.

After you've done these things, turn in your completed `lda.py` file on Moodle.

2 Running on Real Data (10 Points)

In the Github repo, there's a directory called `wiki`. Run you topic model on this set of 400 random wikipedia pages and examine the topics. Upload your results as `topics.txt`. You'll run to run it at least for 1000 iterations.

3 Extra Credit (5 points)

Run Mallet on the same data and take a note of the comparative runtime. Upload the mallet file as `mallet.txt`.