Anupama Garimella

Assignment 3 (Part I) – Write-up

1) LDA: I used the genism package available in python to perform LDA on the ten documents provided by Professor West. On performing LDA, I found that it assigns different words to different topics with probabilities assigned to each word (i.e. given a certain word, what is the probability we are talking about a certain topic?)
2) I then created a toy data set using the five sentences at: http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/
3) I took the following steps to implement the method described in the paper assigned:
    - Load the documents
    - Cluster them into two "fields", in this case, animals and food.
    - Tokenize the documents and remove stop words.
    - Compute the probability distributions for each word within each of the clusters and within the entire corpus.
    - Compute the Shannon entropy for communicating within the same field
    - Compute the cross entropy for communicating between fields
    - Compute the efficiency of communication and the jargon distance

Discussion:

While LDA is a great tool to employ unsupervised learning to lump documents / words into meaningful topics, the jargon distance method takes clusters (possibly created through LDA) and finds the cost of communication between these topics, a.k.a. fields. They both go hand in hand and can be combined together to build powerful models for clustering and measuring similarities / dissimilarities between documents.