Comparison of latent dirichlet allocation (LDA) and cultural holes methods

Both LDA and cultural hole or jargon distance methods provide natural language processing analysis of text documents. There are similar in that they use probabilities, offer flexibility and provide meaning to their users. But they are different in their methodology, output and the meaning they provide.

LDA calculates the topics in a corpus of documents based on the probability of a word occurring in the topic and on the basis of their likelihood of co-occurrence in documents. Jargon distance is calculated based on the differing probabilities of words in the codebooks or dictionary of the writer and reader.

Both methods offer flexibility to their users. With LDA, the user can determine the number of topics to be returned by an analysis. Experimenting with differing counts of topics changes the output. Jargon distance provides flexibility in determining what is defined as a word for comparison. Choices are unigrams, bigrams, trigrams etc. But the user can also compare based on the length of the word. This would capture long words that typically take more effort to understand if unfamiliar than short words.

Users of both methods receive meaning from the respective method. But how that meaning is presented is different. LDA's output is a list of words that comprise each topic. The topic gives context and meaning to the underlying documents. This is especially useful for a large corpus that would take a long time to read the individual documents. The topics provide a quick reference point of the meaning of the content. Jargon distance returns a numerical measure of the complexity differences between the documents in the comparison. A high complexity means you will need to pack your dictionary or have internet access available to google all the terms you as a reader do not understand.

The methodology of the two methods differs. LDA is a generative model. Words are randomly assigned to a topic and then each word is compared by probability to the other words and their probabilities and locations. If the word needs to be moved to a different topic it is. This continues word by word and over the number of iterations the user requests. A sufficient number of iterations get all the words assigned to a good fit of the topic. It has a black box almost magic feel to the results. The voila at the end is the meaning of the topics - seeing it on a document you are familiar with provides confidence of its accuracy on documents that you are not familiar with.

Jargon distance uses formulas from information theory: Shannon entropy and cross entropy in its calculations. Word probabilities within the writer's field/topic are compared to words in the reader's field/topic. Readers and writers within the same field/topic use common words. The resulting calculation, the cultural hole is smaller than when the reader and writer are from different fields/topics. The magnitude of the cultural hole represents how different the language is between the reader and writer.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993-1022.

DA Vilhena, JG Foster, M Rosvall, JD West, J Evans, CT Bergstrom, 2014. Finding cultural holes: how structure and culture diverge in networks of scholarly communication Sociological Science 1, 221-238