Justin Melde
Data Science III
05/12/2015

Assignment #3 : LDA & Jargon Distance

Toy Data Set

Document #1
I like to read science fiction and fantasy books.

Document #2
I read a book about science and robots yesterday.

Document #3
Ancient Greek and Latin are fun languages.

Document #4
My friend learned Latin last year.

Document #5
Try reading a fantasy book in Ancient Greek someday.

LDA Results
Number of requested latent topics to be extracted from the corpus: 2
Number of iterations (passes): 20

Sample Set
Topic 1: 0.004*assembl + 0.004*member + 0.004*protector + 0.004*support + 0.004*bathtub + 0.004*wall + 0.004*seat + 0.004*heel + 0.004*strap + 0.004*toe

Topic 2: 0.004*edg + 0.004*bag + 0.004*sleep + 0.004*semiconductor + 0.003*perineum + 0.003*first + 0.003*roller + 0.003*bodi + 0.003*abras + 0.003*scrubber

Toy Set
Topic 1: 0.063*languag + 0.063*fun + 0.063*Latin + 0.062*robot + 0.062*yesterday + 0.054*year + 0.054*learn + 0.054*last + 0.054*friend + 0.054*My

Topic 2: 0.072*fantasi + 0.067*fiction + 0.067*like + 0.067*someday + 0.067*Tri + 0.057*read + 0.057*book + 0.056*I + 0.056*scienc + 0.053*Greek

Jargon Distance Results
Calculated on Toy Data Set
Alpha parameter of 0.01 used to teleport disciplinary terms into a common corpus

Writer (‘Books’ Topic): Documents 1, 2, and 5
Reader(‘Languages’ Topic): Documents 3 and 4

Shannon Entropy of Writer Set: $H = 4.3404$

Cross Entropy of Writer with Reader Set: $Q = 12.7821$

Efficiency of Communication between two Sets: $E = 0.3396$

Cultural Hole/Jargon Distance between two Sets: $C = 0.6604$

Discussion

Using the NLTK package in Python, I preprocessed both the sample and toy data sets by a) removing stop-words, b) removing punctuation, and c) stemming remaining words using a function based on the Porter Stemming algorithm.

I then tokenized the stemmed words and generated a dictionary of terms with frequency counts using *tf-idf* reduction. I then applied the LDA model to this corpus with the intent of finding two latent topics, using a total of 20 iterative passes to generate the topics as word probabilities shifted and settled.

Reviewing the results of LDA on the Sample Set (based on the set of 10 documents provided in Assignment #2), no clear topics suggest themselves. Word probabilities are low in all cases, and no single term or pattern of related terms stands out. This is somewhat expected as the documents themselves are not particularly related to one another, and thus no strong topic themes appear in the results.

The Toy Set provides more interesting output. I purposefully engineered the documents to show two clear topics: books/reading, and languages. Documents 1 and 2 contain more of the former, with 3 and 4 having more to do with the latter; document 5 is an intended mixture of the two topics, with a slightly stronger leaning towards the book topic.

Running LDA on the Toy Set produces the two intended topics, with ‘Topic 1’ containing high probabilities for words like ‘language’ and ‘Latin’ while ‘Topic 2’ contains strong probability for ‘fantasy,’ ‘fiction,’ ‘read,’ and ‘book.’ An interesting outlier is ‘Greek’ appearing with low probability in the topic dominated more by terms related to books than language, though this also reflects the mixed nature of some of these documents (particularly document 5).

Calculating the Shannon Entropy on the Writer Set (books) produces a result of 4.34; this is noticeably less than the Cross Entropy between the Writer and Reader Sets of 12.78. Intuitively, we take this to mean that the amount of effort required to exchange information within the book set is around a quarter of the amount required for communicating information across the two sets. Because of a lack of shared terms between the two topics, there is much higher required energy expenditure to communicate terms across the topic barrier, leading to low efficiency of communication and a high jargon distance for Readers.

Overall, the LDA and Jargon Distance methods produce two similar, but different results: the former draws out the underlying latent topics within a corpus of documents through term probability distributions, while the latter calculates the efficiency of communication *between* the topics by comparing their respective disciplinary corpora. This assignment shows the two work well in tangent as LDA can suggest topics which may then be used as Writer and Reader frameworks for calculating Jargon Distance.