Tracy Sandifer

INFX 575

Assignment 3

May 12, 2015

Domain-specific language comparison between technical and biomedical publication databases by Jargon distance per Vilhena et al and by Latent Dirichlet Allocation

A toy data set for beginning this work was assembled by searching for the terms “influenza” and “model” in a journal publication title in two domain-specific publication databases: dblp (technical focus) and PubMed (biomedical focus). I downloaded the first four abstracts from each database that were indexed in one database but not the other.

Using the jargon distance methods described in the Vilhena et al. paper presented in class, repeated using unigrams, bigrams, and trigrams from each group of domain-specific abstracts, I calculated the following “cultural hole” estimates of communicating the concepts found in the technical domain to a participant of the biomedical domain:

Using unigrams, the calculated cultural hole from the tech domain to the biomedical domain was 0.62

Using bigrams, the calculated cultural hole from the tech domain to the biomedical domain was 0.55

Using trigrams, the calculated cultural hole from the tech domain to the biomedical domain was 0.54

I did not feel the bigram and trigram results were very trustworthy because in this small set of documents, there were few bigrams in common between the tech-domain-associated and biomedical-domain-associated lists, and even fewer when using trigrams.

For comparison, I also applied the Latent Dirichlet Allocation method to the two groups of abstracts using the gensim package for Python. I initially modeled 6 topics per corpus, but found most topics to be near-repeats of the first two, I repeated the LDA modeling with just two topics per corpus. These topic sets are:

Tech corpus topics:

(1) [u'0.052\*patient + 0.040\*hospit + 0.033\*inuenza + 0.033\*outbreak + 0.027\*model + 0.027\*resourc + 0.027\*alloc + 0.027\*pandem + 0.020\*consid + 0.020\*addit', u'0.033\*system + 0.033\*bird + 0.033\*stabil + 0.023\*global + 0.023\*diseasefre + 0.023\*equilibrium + 0.023\*model + 0.023\*inuenza + 0.012\*individu + 0.012\*analyz', u'0.004\*studi + 0.004\*use + 0.004\*public + 0.004\*health + 0.004\*method + 0.004\*epidem + 0.004\*analyz + 0.004\*individu + 0.004\*limit + 0.004\*network', u'0.004\*studi + 0.004\*use + 0.004\*public + 0.004\*health + 0.004\*method + 0.004\*epidem + 0.004\*analyz + 0.004\*individu + 0.004\*limit + 0.004\*network', u'0.044\*search + 0.033\*health + 0.033\*optim + 0.033\*otc + 0.023\*correl + 0.023\*product + 0.017\*use + 0.017\*identifi + 0.017\*process + 0.017\*select', u'0.051\*event + 0.051\*peak + 0.026\*properti + 0.026\*influenza + 0.018\*studi + 0.018\*propos + 0.018\*surveil + 0.018\*record + 0.018\*time + 0.018\*counti']

(2) [u'0.004\*studi + 0.004\*use + 0.004\*public + 0.004\*health + 0.004\*method + 0.004\*epidem + 0.004\*analyz + 0.004\*individu + 0.004\*limit + 0.004\*network']

Biomedical corpuc topics:

(1)[u'0.004\*increas + 0.004\*base + 0.004\*pattern + 0.004\*provid + 0.004\*use + 0.004\*gener + 0.004\*specif + 0.004\*ill + 0.004\*influenzalik + 0.004\*model', u'0.038\*model + 0.038\*epidem + 0.027\*observ + 0.022\*influenza + 0.022\*serfl + 0.022\*exclud + 0.022\*tradit + 0.017\*use + 0.017\*period + 0.017\*perform', u'0.059\*surveil + 0.022\*outbreak + 0.022\*model + 0.019\*system + 0.016\*diseas + 0.016\*environ + 0.016\*inform + 0.016\*perform + 0.014\*provid + 0.014\*syndrom', u'0.043\*behavior + 0.031\*model + 0.031\*control + 0.019\*result + 0.019\*health + 0.019\*influenza + 0.019\*effect + 0.019\*educ + 0.019\*rate + 0.013\*pattern', u'0.004\*increas + 0.004\*base + 0.004\*pattern + 0.004\*provid + 0.004\*use + 0.004\*gener + 0.004\*specif + 0.004\*ill + 0.004\*influenzalik + 0.004\*model', u'0.004\*increas + 0.004\*base + 0.004\*pattern + 0.004\*provid + 0.004\*use + 0.004\*gener + 0.004\*specif + 0.004\*ill + 0.004\*influenzalik + 0.004\*model']

(2)[u'0.038\*model + 0.038\*epidem + 0.027\*observ + 0.022\*influenza + 0.022\*serfl + 0.022\*exclud + 0.022\*tradit + 0.017\*use + 0.017\*period + 0.017\*perform']

Given that the subject matter of the publications was chosen the to be roughly the same for both groups, and generally biomedically relevant, I was curious whether the papers published in the technical domain but not the biomedical one may have ended up there because of being particularly heavy on technical jargon not generally used in the biomedical domain. The LDA topics modeled from the abstracts do not show evidence of this. I will be interested to repeat the process with the full text of these and additional publications.