CE807 Text Analytics Assignment 2 Name: Ken Suzuki Student id: 2004473

I have identified five topic models (TM) based on previous researches reviewing prominent models in the study of TM. [1] Below are brief explanations on how each model determines the number of topics from the corpus.

Latent semantic analysis (LSA): This model combines document-term matrix (DTM) which calculates the frequency of word 'w' appearing in each document, and singular value decomposition (SVD) to determine the weight of each topic. [2]

Probabilistic latent semantic analysis (PLSA): PLSA utilizes probabilistic methods instead of SVD utilized in LSA. The formed clusters by the pLSA algorithm represent the number of topics in the corpus. [3]

Latent Dirichlet allocation (LDA): LDA adapts Baysian statistical aspects and regards documents as a bag-of-words entity explaining the distribution of words or topics. [4] LDA would produce several topics assigned with stochastic numbers which add up to 1. [5] [6] The number of topics was determined through the log-likelihood function by finding the number of topics with peak scores.

Correlated topic model (CTM): CTM is a modified version of the LDA model by newly considering the correlation between topics. [7] It enables respective documents to represent multiple topics with different weights and optimal topic numbers were calculated by the held out log probability (HOLP) scheme.

lda2vec model: This model combines the essence of LDA and word2vec by building a foundation on LDA topic modeling and adding the context of words. The algorithm would learn the interrelation of a word to its surrounding words, hence measuring an optimal number of topics based on accuracy of topics using topic-based clusters. [8]

References:

[1] Alghamdi, R., & Alfalqi, K. (2015). A survey of topic modeling in text mining. Int. J. Adv. Comput. Sci. Appl.(IJACSA), 6(1).

[2] Valdez, D., Pickett, A. C., & Goodson, P. (2018). Topic modeling: latent semantic analysis for the social sciences. Social Science Quarterly, 99(5), 1665-1679.

[3] Wang, X., Chang, M. C., Wang, L., & Lyu, S. (2019). Efficient algorithms for graph regularized PLSA for probabilistic topic modeling. Pattern Recognition, 86, 236-247.

[4] Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl 1), 5228-5235.

[5] Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. Journal of the american statistical association, 101(476), 1566-1581.

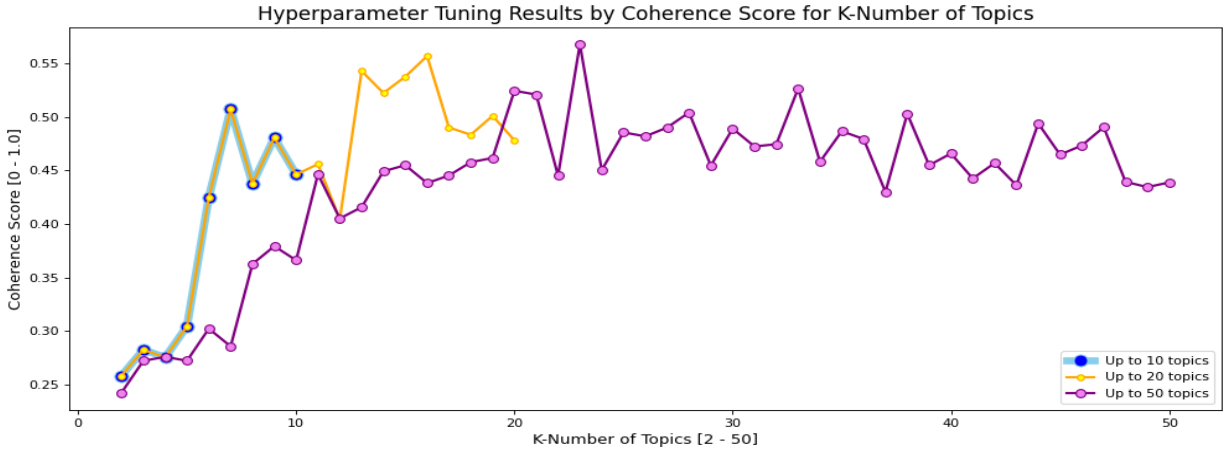
[6] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

[7] Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. The annals of applied statistics, 1(1), 17-35.

[8] Hasan, M., Hossain, M. M., Ahmed, A., & Rahman, M. S. (2019, September). Topic Modelling: A Comparison of The Performance of Latent Dirichlet Allocation and LDA2vec Model on Bangla Newspaper. In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 1-5). IEEE.

Dataset: 2005 - 2006 BBC News Dataset (Original Source: http://mlg.ucd.ie/datasets/bbc.html / Kaggle page)
Pre-Processing: 1) remove punctuation 2) convert to text to lowercase 3) tokenization 4) remove stop-words 5) lemmatization
Model codes: gensim (Coherence Model)
Experimental setting: Clustering
Hyperparameter tuning: alpha = 0.91, beta = 0.91 (After hyperparameter tuning)
My Experiments: [Train/Tune/Test split = 75/10/15]

1. Download the BBC News Dataset from the source page created by Science Foundation Ireland (SFI) or from the Kaggle page. The dataset contains 2225 text (.txt) files with each file containing texts of online articles categorized into 5 topics: business, entertainment, politics, sport, and technology, and extract the title text.
2. We used regular expressions to remove punctuations and unnecessary spaces from the text for preprocessing, then convert texts to lower cases. Next, we used gensim 'simple_preprocess' to tokenize each sentence to words excluding remaining unnecessary punctuations and words all at once. Lastly, remove stop words utilizing 'nltk' and conduct lemmatization to translate words to simpler forms.
3. We used gensim to construct an LDA topic modeling model and we set the number of topics as 10 for the base model with default alpha and beta parameter values. The base trained model resulted in a 0.277 coherence score.
4. We created a for-loop function that automatically displays the coherence score for each number of topics by changing alpha and beta parameter values from 0.01 up to 1 in 0.3 increments while making either of the parameter constant until every pair-wise hyperparameters are tested. We ran the hyperparameter tuning for-loop three times setting the maximum number of topics to 10, 20, and 50 on the tuning set. The below chart displays the result the best coherence score for each number of topics from each hyperparameter tuning:



5. The line chart suggests that 13, 16, and 23 topics marked the top three coherence scores. Then we test the top three highest-performing number of topics with optimized hyperparameters again to determine the final model on the test set. The best model was ***'16 topics'*** [alpha = 0.91, beta = 0.91] with a coherence score of 0.483, a 74 percent increase in performance from the base-model: 0.205 coherence score improvement.