Ed Costantini October 1, 2018

Topic Modeling Writeup for Red Marlin

Text Preprocessing:

1. Read in json and split into uid and relevant text data for analysis
2. Create Document-Term matrix by removing stopwords and stacking term vectors with scikit-learn
3. Build vocabulary for corpus
4. Apply TF-IDF term weighting using sklearn

Parameter Selection

1. Pre-specify range of values for number of topics (k) and apply NMF for each of the values
2. Build word embedding using Gensim/Word2Vec
   1. Tokenize documents for Gensim input
   2. Build Skipgram model from all documents
3. Calculate coherence for each of the topics for each value in suggested k range
4. Plot coherence scores in order to select k value (k=13)

Topic Modeling:

1. Using k=13, as determined by parameter selection
2. Create model and examine factor matrix W (document membership weights relative to each of the topics)
3. Sort by largest weight values, and output top 5 topics for each document, in conjunction with original id in desired format.

Analyzing Results:

1. Use get\_descriptor to find the terms corresponding to top-ranked indices, to get a better understand of what each topic is about.
   1. The most interesting of these results is topic 2, which is in Spanish. I considered a few options for handling this, such as including the common Spanish words into the stopwords and re-creating the model, but decided that having all of the Spanish documents in one topic would be better than spreading them to create additional topics. A separate model for handling Spanish pages would be ideal, and require additional data.
2. Use matplotlib to analyze the weightings of each of these terms, for any particular topic.

Links and References:

|  |  |
| --- | --- |
|  | Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature. http://www.columbia.edu/~jwp2128/Teaching/E4903/papers/nmf\_nature.pdf |
|  | Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4). https://cacm.acm.org/magazines/2012/4/147361-probabilistic-topic-models/fulltext |
|  | O’Callaghan, D., Greene, D., Carthy, J., & Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. Expert Systems with Applications. http://derekgreene.com/papers/ocallaghan15eswa.pdf |
|  | https://github.com/derekgreene/topic-model-tutorial |