WE1S's use of Keyphrase Extraction to supplement topic modeling.

WE1S has supplemented topic modeling (see card M-2) with keyphrase extraction for extra insight into its data. Keyphrase extraction generates a list of the most significant words or phrases within individual documents. WE1S takes the top ten keyphrases in each document and ranks them according to their frequency across the collection. These keyphrases are grouped thematically into topic-like lists which can then be compared to those generated by topic modeling the collection.

WE1S implements the SGRank algorithm (Danesh, et al., 2015), a hybrid statistical-graphical approach to extracting keyphrases from individual documents. Statistical approaches examine the frequency of occurrence of terms in the document or rareness in the corpus, whereas graphical approaches construct graph-based representations of documents with words as nodes and edges reflecting co-occurrence relations, which are then used to rank nodes. Keyphrase extraction thus supplements our topic modeling method by providing insights based on (a) the prominence of keyphrases within documents, rather than just the collection as a whole, and (b) the salience of keyphrases within the context of surrounding language in the documents. Although numerous algorithms are available, WE1S uses SGRank because it produces intelligible results and is conceptually a good complement to topic modeling.

WE1S uses Python Textacy and spaCy libraries to perform keyphrase extraction. , although Textacy provides a number of algorithms for keyphrase extraction. SGRanks is computationally intensive, so in our experiments we have improved the efficiency of the algorithm by including only nouns and proper nouns and by setting a window of 70 words to either side of each candidate keyphrase (which may be 1-6 words). Most keyphrases occur under 15 times, creating a long tail of less significant terms. We found that the top 1500 keyphrases are most useful for interpretation.

Keyphrase extraction often captures significant phrases like "Black History Month" more accurately than our topic models, while our topic models provide insights that incorporate single-word adjectives, unlike our keyphrases. By and large, our keyphrase lists confirm observations we make based on our topic models, providing us with added confidence in our topic modeling results. Where our collections have metadata tags, we are able to compare prominent keyphrases in subsets of our collections.based on metadata categories. Research using this method is ongoing.

Resources

Recommended article

Danesh, Sumner, et al. "SGRank: Combining Statistical and Graphical Methods to Improve the State of the Art in Unsupervised Keyphrase Extraction" (2015), doi: 10.18653/v1/S15-1013.

Tools

Textacy
spaCy