Technology Review: Gensim

Scott Downey (scottmd3)
CS410: Text Information Systems
October 30, 2021

At the beginning of 2020, it was estimated that there would be 44 zettabytes of data on the internet, much of this written in natural language (text made for humans). It would be impossible for a human to sift through, analyze and consume this amount of data in a lifetime. As the amount of data in the world continues to grow faster and faster, it becomes imperative to use machines to implement text mining and analytic techniques to help advance our knowledge and better understand the world we live in. One of these techniques is topic modeling - the process of discovering latent topics in a body of text. There are many motivations for this type of analysis. For example, what are users discussing on social media platforms? How do people feel about specific products? How have debate issues in the presidential elections changed over time? These questions can be answered by machines using topic modeling. However, natural language processing by computers can be difficult to do. Not only is the amount of data enormous and time consuming, natural language for humans can have ambiguity and require common sense knowledge. The toolkit Gensim provides a way to semantically analyze these human-oriented, plain-text documents and give meaningful topic modeling with as little human intervention as possible.

Gensim (short for “Generate Similar”) is an open-source Python library that translates unstructured, plain-text documents into semantic vectors to use for analysis and modeling. It is platform independent (can run on any platform that supports Python and NumPy), only requires to have Python (tested on versions 3.6+), NumPy and smart_open libraries, and is free for both commercial and personal use, although users cannot modify the license itself. One highlight of the Gensim toolkit is the algorithms are unsupervised, so there is no human input required for tagging or annotating the text documents. Gensim is also highly scalable being “memory independent”, as the entire training corpus does not need to reside in Random Access Memory and can use large scale data streaming (for example, from the internet). Gensim also boasts highly optimized modeling algorithms, making computations fast and efficient.

Gensim provides a variety of tools to translate plain-text documents to a fully functioning model for analysis. The process begins by defining the corpus, a collection of document objects (a sequence of text). In Gensim, the corpus is the input for training a desired model. It also helps to serve as a form of organization for the documents and used to discover latent topics for new documents. Gensim is capable of working with a corpus fully in memory or streaming millions of documents in one at a time via an iterable function. In fact, Gensim will take any object that is iterable and produce documents in succession (and in turn, produce a document vector each iteration). Once the corpus is established, Gensim can help preprocess the data in many different ways to make it more meaningful to model. 'Gensim.utils.simple_preprocess' will convert a document to a list of lowercase tokens and ignore tokens too short or too long (user defined arguments). The 'gensim.parsing.preprocessing' library also contains a variety of functions, such as converting to unicode, removing stop words, splitting lines by spaces, stemming words, removing html tags, etc.. The 'preprocess_string' function is particularly helpful, as it combines multiple string preprocessing techniques into one function. Gensim will also tokenize the documents while preprocessing, turning them into a list of words. After preprocessing, Gensim's 'Dictionary' class can be used to map each word to a unique ID number. The 'Dictionary' class also has several useful metadata available, such as collection frequencies, document frequencies, number of docs processed, and various other helpful data on the preprocessed corpus.

Once tokenization of documents and dictionary establishment is complete, Gensim functions, such as 'doc2bow', can be used to represent the corpus as a bag-of-words vector. After the corpus has been translated into vectors, we can start to transform them utilizing the various models Gensim has available. One such model is 'tfidf' (term frequency - inverse document frequency). Once trained with this model, users can start performing similarity queries for text retrieval. A variety of topic modeling functionality is available as well. Gensim can perform Latent Semantic Analysis/Indexing (LSA/I) using the 'LsiModel' function. It also has the capability of creating Latent Dirichlet Allocation (LDA) models, which is a probabilistic generative model, using the built in 'LdaModel' function. Both of these models can be used to find latent topics from the documents in the corpus. Gensim is also capable of performing many

transformations of these vector spaces, such as Hierarchical Dirichlet Process, Log Entropy and Author-Topic models, to name a few. It even has functionality to load models from Facebook's fastText I/O. All of these trained models in Gensim can be saved to disk and loaded back at the user's discretion.

These models and analysis derived from Gensim can be applied to a multitude of potential real-world applications. Bioinformatics has an increasing amount of biological datasets. Utilizing topic modeling, such as Gensim's LSA and LDA models, can help discover latent topic information and improve the interpretability of biological information for researchers. Using Gensim's 'Word2Vec' and 'Doc2Vec' models, users can create sentiment analysis for a variety of topics, such as analyzing movie reviews (Rotten Tomatoes, IMDB) or how people feel about a specific product (from social media posts, blogs, or online article reviews). Gensim topic modeling capabilities can also be applied to marketing and discovering customer interests, finding specific topics being discussed (via transcripts or other text communication) and tailoring advertisements as necessary. There are currently several companies using Gensim in their operations, including DynAdmic (online marketing), Sports Authority (social media and customer surveys) and Issuu (digital publishing).

From Bioinfomratics, to sentiment analysis to customer marketing, topic modeling is widely applicable to help drive insights to a variety of subjects. The Gensim toolkit develops a path to unlocking this. With a robust library of functionality, Gensim aids in loading and preprocessing documents. It translates these documents into vectors that can be manipulated and transformed mathematically. It has several topic models at its disposal for analyzing and finding insights to text data through LDA, LSA and other modeling functions. The amount of text data people produce will continue to grow, and Gensim provides an excellent solution to topic modeling, helping us better understand the information and world around us.

References:

Czerny, Michael. “Modern Methods for Sentiment Analysis.” *Medium*, District Data Labs, 21 Dec. 2017, https://medium.com/district-data-labs/modern-methods-for-sentiment-analysis-694eaf725244.

Fatma, Fatma. “Industrial Applications of Topic Model.” *Medium*, Medium, 4 Apr. 2019, https://medium.com/@fatmafatma/industrial-applications-of-topic-model-100e48a15ce4.

Liu, Lin, et al. “An Overview of Topic Modeling and Its Current Applications in Bioinformatics.” *SpringerPlus*, Springer International Publishing, 20 Sept. 2016, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5028368.

Vulveta, Branka. “How Much Data Is Created Every Day? [27 Powerful Stats].” *SeedScientific*, 28 Jan. 2021, https://seedscientific.com/how-much-data-is-created-every-day/.

Zhai, ChengXiang, and Sean Massung. *Text Data Management and Analysis a Practical Introduction to Information Retrieval and Text Mining*. ACM, 2016.

Řehůřek, Radim. “Gensim: Topic Modelling for Humans.” *What Is Gensim? - Gensim*, 30 Aug. 2021, https://radimrehurek.com/gensim/intro.html#what-is-gensim.