Sameer Phadnis

Professor ChengXiang Zhai

CS410

11/15/2020

Tech Review: Gensim Toolkit

Introduction

Gensim is a free and open-source Python library for text analytics. It was developed by RARE technologies, originally as a research\academic project for Czech Digital Mathematics. The initial release was in 2009, and the first stable release was in 2019.

Gensim is a popular and robust library used both for commercial and academic applications by thousands of organizations. It provides powerful Topic Modelling capabilities based on:

- Training large scale semantic NLP models
- Representing text as semantic vectors
- Finding semantically related documents

These capabilities are based on implementations of popular algorithms such as *Latent Semantic Analysis* and *Latent Dirichlet Allocation.* Gensim uses unsupervised models so human intervention is not needed.

Gensim is highly scable and efficient because it does not require loading of the entire corpus in memory, and its implementation is optimized to take advantage of multi-core clusters.

Details

Dictionary and Corpus

It is necessary to create the Dictionary and Corpus before doing text processing with GenSim. A Dictionary is a mapping between each word and its unique id. The corpus is the entire set of available documents. Gensim has APIs to create the Dictionary object by processing the set of documents. The documents can be passed as a Python list or a set of text files.

The Corpus is a bag-of-words representation of all the documents, and can be considered GenSim's equivalent of a Document-term matrix. It has the word id and frequency in each document. Gensim supports reading files over the network, also it is possible to populate the corpus by reading one ine at a time, without having to read the entire file in memory. Thus it is possible to process very large files and data sets. The Dictionary and Corpus are persisted on the disk. It is possible to use the TF-IDF model to create the Corpus, which weighs down words that occur frequently across documents.

Topic modelling with Gensim

Topic modelling is the process of extracting underlying topics from a set of documents, given a pre-defined number (count)of topics as input. On real data sets, number of topics of 200 to 500 is recommended. Gensim implements the Latest Dirichlet Association (LSA) and Latest Semantic Indexing (LSA) algorithms for topic modelling.

Similarity Queries

Gensim supports the querying of a corpus to find similar documents. This can apply to determining similarity between pairs of documents, or similarity between a specific document

and a set of documents (such as a user query vs indexed documents). This is implemented based on Latest Semantic Indexing transformation of the corpus, and using Cosine similarity to determine the similarity of two vectors. Gensim API returns the set of documents that are most similar to the input query, and a similarity score for each document. Thus the results can be presented to the user in decreasing order of similarity.

Word Embedding with Gensim

Word Embedding is a dense vector representation for text where words having the same meaning have similar representation. Word embedding methods learn a real valued vector representation from a corpus of text. Gensim provides implementation of the following algorithms:

- Word2Vec algorithm by Google.

  This is a widely used algorithm based on neural networks. Using large amount of unannotated plain text, Word2Vec learns relationship between words automatically.

- Doc2Vec model

  This represents each document as a paragraph vector. Gensim provides an implementation of this model for finding similar documents.

- fastText

  This allows training word embeddings from a training corpus with the additional ability to obtain word vectors for out-of-vocabulary words. Gensim provides an implementation of this model for finding similar documents.

Conclusion

Gensim is a powerful library having a rich set of capabilities for text analytics. Learning this library can equip the user with powerful set of techniques, as well as serve as an excellent means to learn about text processing techniques in a practical context.

Works Cited

https://radimrehurek.com/gensim/auto_examples/index.html#core-tutorials-new-users-start-here

https://www.tutorialspoint.com/gensim/index.htm

https://www.machinelearningplus.com/nlp/gensim-tutorial