Peter Jefferson – peterwj2 - peterwj2@illinois.edu
Technology Review – CS410 – Fall 2022

Gensim

Overview, Analysis, Comparison

Overview

Today, in context of the Natural Language Processing (NLP), there are a plethora of toolkits, libraries, and approaches that one can take to solve an NLP problem. This can be overwhelming for newcomers due to the vast amounts of options they have. It can also be resource intensive for individuals, academic researchers, and companies to quickly utilize the benefits of NLP to meet their specific use case.

To exacerbate this, some toolkits are better for certain applications than others. This technology review will mainly be reviewing Gensim, what it is, what it offers, what its strengths and weaknesses are, and how it compares to other NLP tools available today.

Introduction

As described by its own creator, Radim Řehůřek, Ph.D.:

> "Gensim is a free open-source Python library for representing documents as semantic vectors, as efficiently (computer-wise) and painlessly (human-wise) as possible. Gensim is designed to process raw, unstructured digital texts ("plain text") using unsupervised machine learning algorithms.
>
> The algorithms in Gensim, such as Word2Vec, FastText, Latent Semantic Indexing (LSI, LSA, LsiModel), Latent Dirichlet Allocation (LDA, LdaModel) etc, automatically discover the semantic structure of documents by examining statistical co-occurrence patterns within a corpus of training documents."[1]

To generalize, Gensim is designed to be an intuitive Python library that can black-box most major unsupervised NLP algorithms. This brings immense value to newcomers, those who want to prove out a quick proof of concept, or if one has a large number of documents that needs to be streamed rather than read into memory.

Analysis

To analyze Gensim, the following metrics will be used to measure its strengths and weaknesses: ease of use, major NLP algorithms provided, documentation, performance, OS limitations, and support.

Ease of use

Given that Gensim is a Python library, its syntax is Pythonic, making it easily readable. Gensim's documentation is also among the best I have seen for any Python library, with a comprehensive

tutorial for each core concept that includes explanations as well as Gensim specific code snippets.
For users that need a quick proof of concept or do not have data for a model, they can quickly leverage Gensim's *Gensim-data Project* to use specific datasets such as an industry like legal or healthcare. [2] This saves users from the time and hassle of finding a clean, usable dataset to train their models.

Major NLP Algorithms Provided
Gensim supports the following vector space model algorithms, Term Frequency – Inverse Document Frequency (TF-IDF), Latent Semantic Indexing (LSI), Random Projections (RP), Latent Dirichlet Allocation (LDA), and Hierarchical Dirichlet Process (HDP). [3]

Documentation
Gensim is *extremely* well documented, especially for those who are new to NLP concepts, as it has tutorials explaining the core concepts such as Corpus, Vectors, and Models. The tutorials also show you how to implement these topics in Gensim. [4]

Performance
Gensim is known for its speed. It has been highly optimized on Python, which is valuable because it combines Python's ease-of-use without its famous speed concerns. Its core algorithm runs in parallel C routines. [5]

In terms of memory, Gensim is also known for its intuitive solution for training models without needing the whole training data (corpus). One can *stream* large documents from a storage solution such as an AWS S3 bucket instead of reading the whole corpus into RAM. This is essential for users with massive corpora like large companies.

OS Limitations
There are no OS limitations to worry about for Gensim, as long as the platform can support Python 3.6 or higher and NumPy, it will work. This is great for those specific odd use cases. [5]

Support
Gensim is one of the top NLP packages used today. Many large companies, academic researchers, and individual users use Gensim. The ongoing support and maintenance of this package will not be a problem for the foreseeable future. There are three types of support, Open-Source support, Commercial support, and Developer support.

Comparison – Gensim, NLTK, SpaCy

It is important to pick the right tool(s) for your specific use-case; Gensim, NLTK, SpaCy, and other NLP packages have their unique pros for different problems. Below is a chart to demonstrate the pros and cons of each of these three NLP packages from 1 -10, lowest to highest respectively.

Package	Ease of Use	NLP Algorithms Provided	Documentation	Performance	OS Compatibility	Support
Gensim[1]	10	6	10	8	9	8
NLTK[6]	5	10	7	5	7	4
SpaCy[7]	7	8	8	10	7	8

Of course, your actual use case and NLP algorithms needed will vary, but if you are just looking for a quick-to-learn, easy to use package, Gensim and SpaCy are both great options as they are documented well, perform well, and there is ongoing support for them. If your use case is more complex, or you require more advanced algorithms, NLTK shines over Gensim and SpaCy in that aspect.

Conclusion

While there are still many other unmentioned NLP packages out there, the majority of NLP algorithms are covered between Gensim, NLTK, and SpaCy. As stated earlier, if your goal is to get something done quickly, you are a newcomer, or performance is a top priority, go with Gensim or SpaCy. If your priorities align more with using more complex or less popular NLP algorithms, utilize NLTK. Gensim is also the clear winner for companies needing to use NLP. The support is good, they can bring employees up to speed quickly via the great documentation, and the performance is superior on streaming large corpuses instead of reading everything into RAM at once.

Citations

1. - Řehůřek, Radim. “Gensim: Topic Modelling for Humans.” What Is Gensim? - Gensim, 6 May 2022, https://radimrehurek.com/gensim/intro.html#what-is-gensim.

2. - Sharaf, Ibrahim, et al. “New Download API for Pretrained NLP Models and Datasets in Gensim.” Pragmatic Machine Learning, 27 Nov. 2017, https://rare-technologies.com/new-download-api-for-pretrained-nlp-models-and-datasets-in-gensim/.

3. - “Gensim: Topic Modelling for Humans.” Topics and Transformations - Gensim, 6 May 2022, https://radimrehurek.com/gensim/auto_examples/core/run_topics_and_transformations.html#sphx-glr-auto-examples-core-run-topics-and-transformations-py.

4. - “Gensim: Topic Modelling for Humans.” Core Concepts - Gensim, 6 May 2022, https://radimrehurek.com/gensim/auto_examples/core/run_core_concepts.html#sphx-glr-auto-examples-core-run-core-concepts-py.

5. - “Gensim: Topic Modelling for Humans.” Radim Řehůřek: Machine Learning Consulting, 6 May 2022, https://radimrehurek.com/gensim/index.html.

6. - “Spacy · Industrial-Strength Natural Language Processing in Python.” · Industrial-Strength Natural Language Processing in Python, https://spacy.io/.

7. - “NLTK - Natural Language ToolKit.” NLTK, https://www.nltk.org/.