Gensim is a free Python library designed to automatically extract semantic topics from documents, as efficiently (computer-wise) and painlessly (human-wise) as possible.

Gensim is designed to process raw, unstructured digital texts ("plain text").

Features:

- Memory independence - no need for the whole training corpus to reside fully in RAM at any one time (can process large, web-scale corpora)
- Memory sharing - trained models can be persisted to disk and loaded back via mmap. Multiple processes can share the same data, cutting down RAM footprint.
- Efficient implementations for several popular vector space algorithms, including Word2Vec, Doc2Vec, FastText, TF-IDF, Latent Semantic Analysis, Latent Dirichlet Allocation or Random Projection.
- I/O wrappers and readers from several popular data formats.
- Fast similarity queries for documents in their semantic representation

Core Concepts:

Corpus:

A collection of digital documents. Corpora serve two roles in Gensim:

- Input for model training. The corpus is used to automatically train a machine learning model, such as LsiModel or LdaModel. The models use this training corpus to look for common themes and topics, initializing their internal model parameters. Gensim focuses on unsupervised models so that no human intervention, such as costly annotations or tagging documents by hand, is required.
- Documents to organize. After training, a topic model can be used to extract topics from new documents (documents not seen in the training corpus). Such corpora can be indexed, queried by semantic similarity, clustered etc.