Cheatography

Gensim CheatSheet Cheat Sheet

by Sonali Gadpandey via cheatography.com/137364/cs/28829/

About

- Open-source python library.
- Used in unsupervised topic modeling.
- Designed to extract conceptual concepts from documents.

Corpora and Vector Spaces

From string to vector -

```
>>> from gensim import corpora
>>> doc = [#put any document]
>>> dictionary = corpora.Dictionary(texts)
>>> dictionary.save('/tmp/deerwester.dict')
>>> print(dictionary)
>>> print(dictionary.token2id)
```

Corpus Format -

Market Matrxi Format:

```
>>> corpus = [[(1, 0.5)], []]
>>> corpora.MmCorpus.serialize('/tmp/corpus.mm', corpus)
```

Other formats include Joachim's SVMlight format, Blei's LDA-C format and GibbsLDA++ format.

API References

matutils - Math helper functions.

- class gensim.matutils.Dense2Corpus(dense, documents_columns=True)
- class gensim.matutils.MmWriter(fname)
- class gensim.matutils.Scipy2Corpus(vecs)
- class gensim.matutils.Sparse2Corpus(sparse, documents_columns=True)

API References

Models -

- models.ldamodel
- models.lsimodel
- models.tfidfmodel
- models.hdpmodel
- models.word2vec
- models.doc2vec

API References (cont)

- models.fasttext

Features

1. Scalability
2. Robust
3. Platform Agnostic
4. Open-source
5. Community Support

Topics and Transformations

```
#initialize a model
from gensim import models
tfidf = models.TfidfModel(-
corpus)
#use the model to transform
vectors
doc_bow = [(0, 1), (1, 1)]
print(tfidf[doc_bow])
```

API References

utils - contains various general utility functions.

- class gensim.utils.ClippedCorpus(corpus, max_docs=None)
- class gensim.utils.FakeDict(num_terms)
- class gensim.utils.InputQueue(q, corpus, chunksize, maxsize, as_numpy)
- class gensim.utils.RepeatCorpus(corpus, reps)
- class gensim.utils.SaveLoad

API References

Corpora -

- corpora.bleicorpus - corpus is Blei's LDA-C format
- corpora.dictionary - construct word <-> id mappings
- corpora.lowcoprus - corpus in list-of-words format
- corpora.mmcorpus - corpus in matrix market format

API References (cont)

- corpora.svmlightcorpus - corpus in SVMlight format
- corpora.wikicorpus - corpus in Wikipedia dump
- corpora.textcorpus - building corpora with dictionaries

Core concepts

1. Document: any text

```
>>> doc = "Gensim is open-source python library. "
```

2. Corpus: a collection of documents.

```
>>> corpus = ["Gensim is an open-source library", "Used in unsupervised topic modelling"]
```

3. Vector: a document that can be represented in a mathematically useful way.

```
>>> pprint.pprint(dictionary.token2id)
```

4. Model: an algorithm to transform vector.

```
>>> tfidf = models.TfidfModel(BoW_corpus)
```

API References

interfaces - realized as abstract base classes.

- class gensim.interfaces.CorpusABC

```
>>> for doc in corpus:
#do something with the doc...
>>> for attr_id, attr_value in doc:
#do something with the attribute
```

- class gensim.interfaces.SimilarityABC(corpus)

```
>>> index = MatrixSimilarity(common_corpus)
>>> similarities = index.get_similarities(common_corpus[1])
```

- class gensim.interfaces.TransformationABC

```
>>> model = LsiModel(common_corpus, id2word=common_dictionary)
>>> bow_vector = model[common_corpus[0]]
>>> bow_corpus = model[common_corpus]
```

By **Sonali Gadpandey**
cheatography.com/sonali-gadpandey/

Published 11th August, 2021.
Last updated 11th August, 2021.