LDA Class Documentation

1. Load_data (train data)

- Loads training data of type list of strings (documents).
- Converts loaded data to class objects to be used by internal class methods.

2. Preprocess (documents)

- splits list of documents so each document is processed individually.

- passes each single document to tokenize method.

- appends processed documents in a list.

- yields list of processed (tokenized + stemmed + lemmatized) documents.

3. Tokenize (document)

-preforms tokenization on each individual document.

- splits document into sentences
- sentences into words.
- remove punctuation.
- Words that have fewer than 3 characters are removed.
- All stop words are removed.

- passes each tokenized word(token) to lemmatize method.

- passes each lemmatized word(token) to stemmed method.

- appends processed words in a new document.

- yields a document of processed (tokenized + stemmed + lemmatized) words.

4. Lemmatize (word)

-preforms lemmatization on each passed token(word).

- words in third person are changed to first person.
- verbs in past and future tenses are changed into present sentences into words.

- yields lemmatized token.

5. Stemm (word)

-preforms stemming on each passed token(word).

- words are reduced to their root form

- yields stemmed token.

6. doc2bow_dict (processed_docs)

-generates dictionary (mapping of words with ids) from processed documents.

-generates bag of words (id of each word + occurrence frequency in each document).

- saves generated dictionary.

- yields dictionary + bag of words.

7. filter_docs (min. reps, doc_occur)

- Applies another filtering step to generated dictionaries by removing
  - Words repeated less than min. reps
  - Words appear more than doc_occur(fraction of documents).

8. LDA_model (num of topics, dictionary, epochs, cores)

- Initialize LDA mode for training by setting:
- Sets Number of cores in case of multicore.
- Sets Number of output topics (singlecore + multicore).
- Sets the generated dictionary.
- Sets the generated bag of words.
- Sets number of epochs (full passes over training data).

9. Get_topics ( )

- Loads generated topics from trained LDA model.
- Each topic is composed of a distribution of specific words.

10. train (save_path, num_cores , batch size , epochs , topics)

- Passes training data to preprocess method-> preprocessing documents
- Passes processed documents to doc2bow_dict method->
  - Dictionary + bag of words generation.
- Calles filter_docs if required.
- Starts training of LDA model on training data(documents).
- Saves trained model.