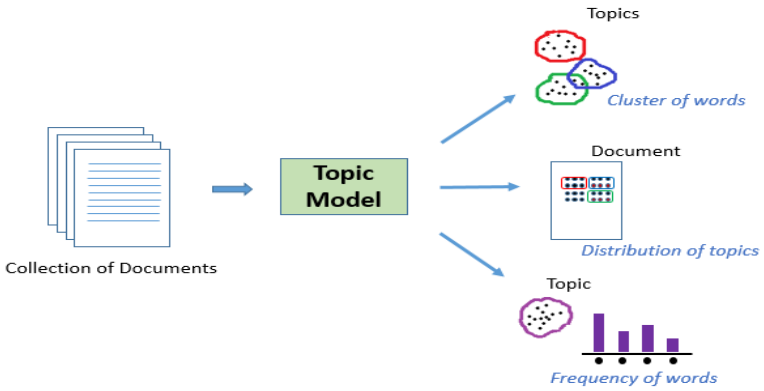
AMAZON_FOOD_REVIEWS
TOPIC_MODELLING

MVP

Overview of work completed so far:

1. Data preprocessing and cleaning:
   - Removed non-alphabetic characters and converted text to lowercase.
   - Removed stop words and short words (length < 3).
2. Tokenizing and lemmatizing the reviews:
   - Tokenized the reviews using spacy.
   - Lemmatized the tokenized reviews using spacy.
3. Creating the term dictionary and document-term matrix:
   - Created a term dictionary of the corpus, where every unique term is assigned an index.
   - Converted the list of reviews into a document-term matrix using the dictionary prepared above.
4. Building the LDA model:
   - Built an LDA model using the genism library.



Here are some issues I've run into so far, and what I plan to do to mitigate them:

1. Determining the optimal number of topics:

This is a challenging task, as there is no one-size-fits-all answer. I plan to evaluate the LDA model for different numbers of topics and select the number of topics that produces the best results on a held-out test set.

2. Interpreting the LDA model:

LDA models can be difficult to interpret, as the topics are represented as distributions of words. I plan to use a variety of techniques to interpret the LDA model, such as examining the top-scoring words for each topic and identifying the relationships between topics.

3. Evaluating the LDA model:

There is no standard metric for evaluating the performance of LDA models. I plan to use a variety of metrics, such as perplexity and coherence, to evaluate the performance of the LDA model.

In addition to these issues, I am also aware of the following general limitations of LDA models:

LDA models are sensitive to the order of the training data. This can lead to different results if the training data is shuffled in different ways. I plan to mitigate this issue by using a variety of techniques, such as random shuffling and averaging the results of multiple models.

Next steps:

Evaluate the LDA model:

- Use a variety of metrics to evaluate the performance of the LDA model, such as perplexity and coherence.
- Interpret the LDA model:
- Identify the key topics in the corpus and the relationships between them.

Use the LDA model for downstream tasks:

- Should build the interface for smooth interaction I.e Upon selecting the keyword, user should be presented with the related topic.

Overall, the work completed so far provides a good foundation for building and evaluating an LDA model for topic modeling.