

Tutorial 8

In this tutorial you will learn how to take advantage of topic modeling and LDA to find the major topics discussed in the text. Moreover, you will be able to extract the terms associated to those topics in text.

- 1- [Prelab] Load the recipes dataset and topicmodels library.
- 2- Randomize the dataset and select the first 10000 rows and your dataset.
- 3- Build the document term matrix with term frequency over your data.
- 4- Filter the term with the frequency of less than 10 from your matrix
- 5- Plot the frequency of ingredients in your recipes (top 25 ingredients) .
- 6- Remove the top 3 ingredients from the corpus and remove the empty documents.
- 7- Build the LDA model over the corpus with 20 topics.
- 8- Observe the most frequent 5 terms in each topic.

Hand in: How can we improve the quality of our topic model? Which parameters can affect the quality of our topic model?