

# Advanced Probabilistic Machine Learning and Applications

Martina Contisciani and Caterina De Bacco

July 16, 2021

## 1 Tutorial 12: Topic Modeling with LDA

### Exercise 1: processing bag of words representation and analyze results

We will start by learning how to process standard text data and doing some basic analysis on the output of LDA.

- (a) Tokenize the document. This is needed to process the dataset assigning ids to words and a matrix  $D \times V$  for the corpus, where  $D$  is the number of documents and  $V$  the length of the vocabulary. We will use the *CountVectorizer* function of *sklearn*. This allows to perform preprocessing tasks such as:
  - Remove punctuation.
  - Remove "stop words".
  - Remove low/high-frequency words.
  - Create the dictionary.
  - Create the bag-of-words representation.
- (b) Run LDA
- (c) Analyze the resulting parameters.
- (d) Apply to new documents.
- (e) Visualize results.

### Exercise 2: analyze real dataset of NY Times articles

Repeat the same analysis for a real dataset.

- (a) Download the dataset from <https://archive.ics.uci.edu/ml/machine-learning-databases/bag-of-words/>. You need the files *docword.nytimes.txt.gz* and *vocab.nytimes.txt*.
- (b) Run bash script:

```
tail -n +4 docword.nytimes.txt > nytimes.txt
```

This will remove the first 3 lines, which are not part of the dataset and output data inside the file *nytimes.txt*.
- (c) Import data into the proper format.
- (d) Run LDA.
- (e) Analyze results.