Modelling Text

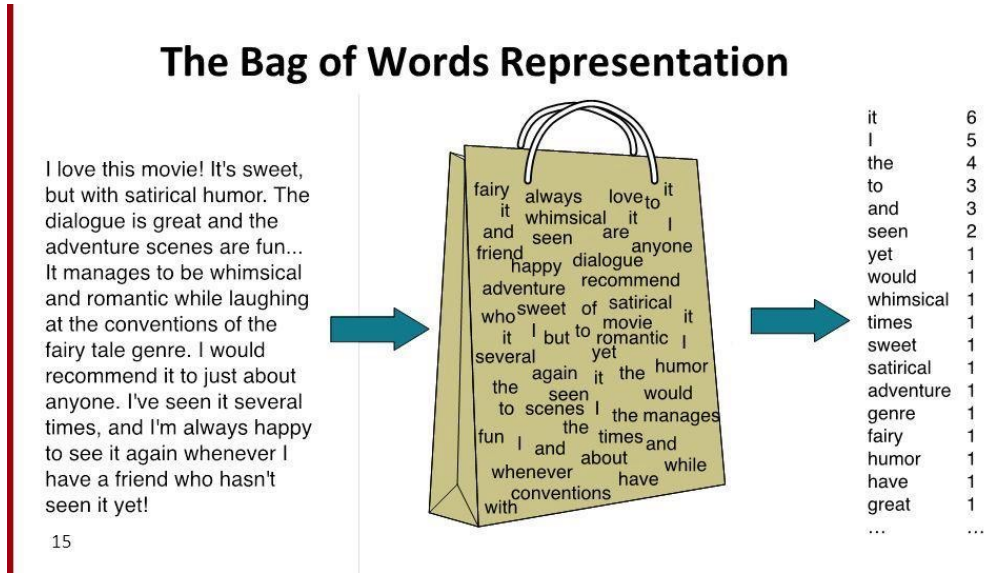
Introduction

Different ways to look at a text

- So far we have focused on specific information in the text
- From now on, we will focus on the text as an entirety
- To be more precise, we will look at three different scenarios:
  1. Text as a bag of words
  2. Text as a sequence of words
  3. Texts as a bag of topics

Text as a Bag of words

- This approach will model a text as an unordered set of words appearing in it
  - But we are asked to model our words appropriately!



Taken from: https://www.programmersought.com/article/4304366575/

➔This yields computer interpretable models of a text, based on its words

➔And computer interpretable models for individual words, which we call „embeddings“

Text as a sequence of words

- We are now facing a different challenge

What is the probability to find word w_i, given the previous context $w_1 \dots w_{i-1}$

https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-language-model-nlp-python-code/

➔This allows us to generate new text!

➔Usually called a Language Model

Text as a bag of topics

- We are now reducing a text to the topics it is dealing with

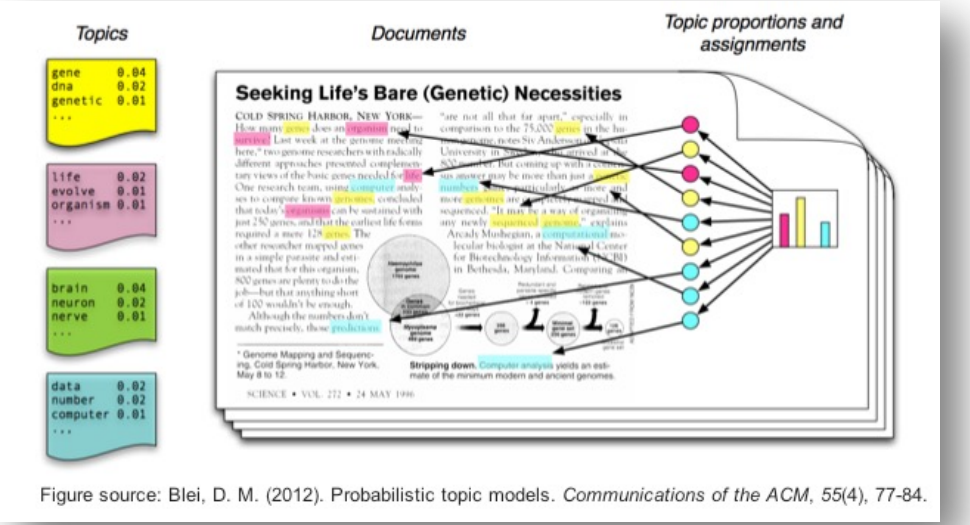


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

➔Yields computer interpretable models of a text, based on its topic distribution!

➔Allows us to generate new texts (however yet again only as a bag of topics!)