

# 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!

## The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

15



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

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  $w_1 \dots w_{i-1}$

Can you please come **here** ?

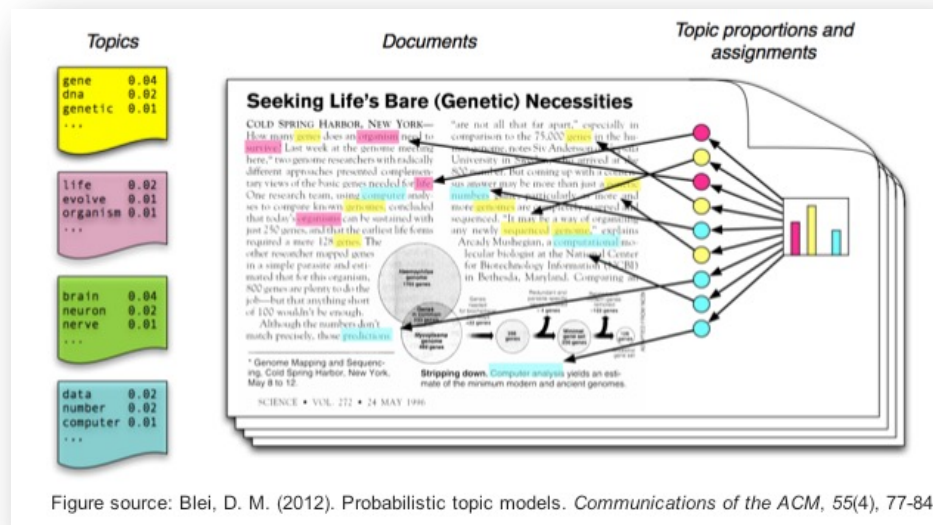
The diagram illustrates the concept of a language model. It shows the sentence "Can you please come here ?". A bracket under the words "Can you please come" is labeled "History". An arrow points from the word "here" to the label "Word being predicted".

<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



- ➔ 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!)