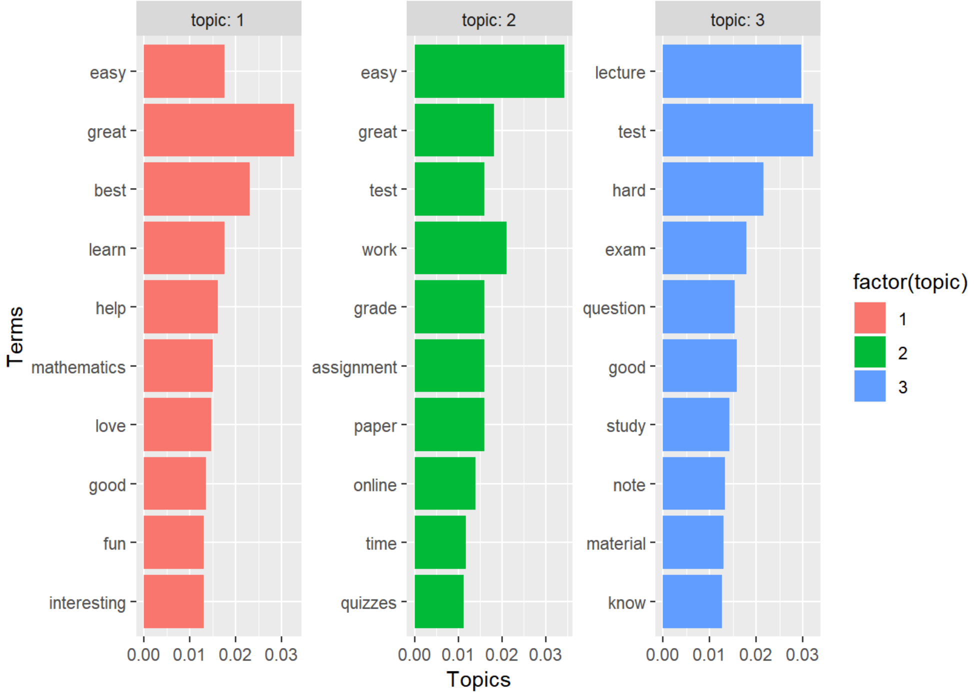
University of
South Wales
Prifysgol
De Cymru

MS4S09

Topic Modelling



https://learn.datacamp.com/courses/topic-modeling-in-r

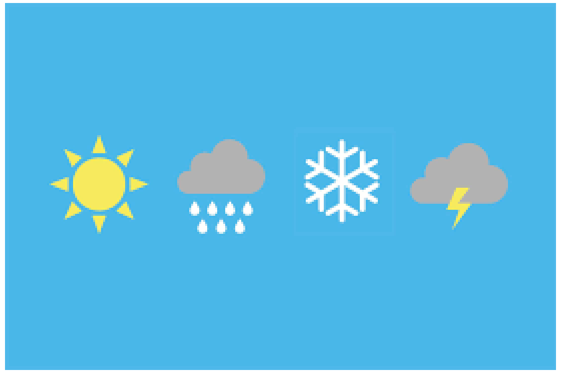
University of South Wales
Prifysgol De Cymru

Topic Modelling

Topics give us an idea what text is about quickly.

A topic is a label for a collection of words that often occur together.

What words would you associate with the **weather**?



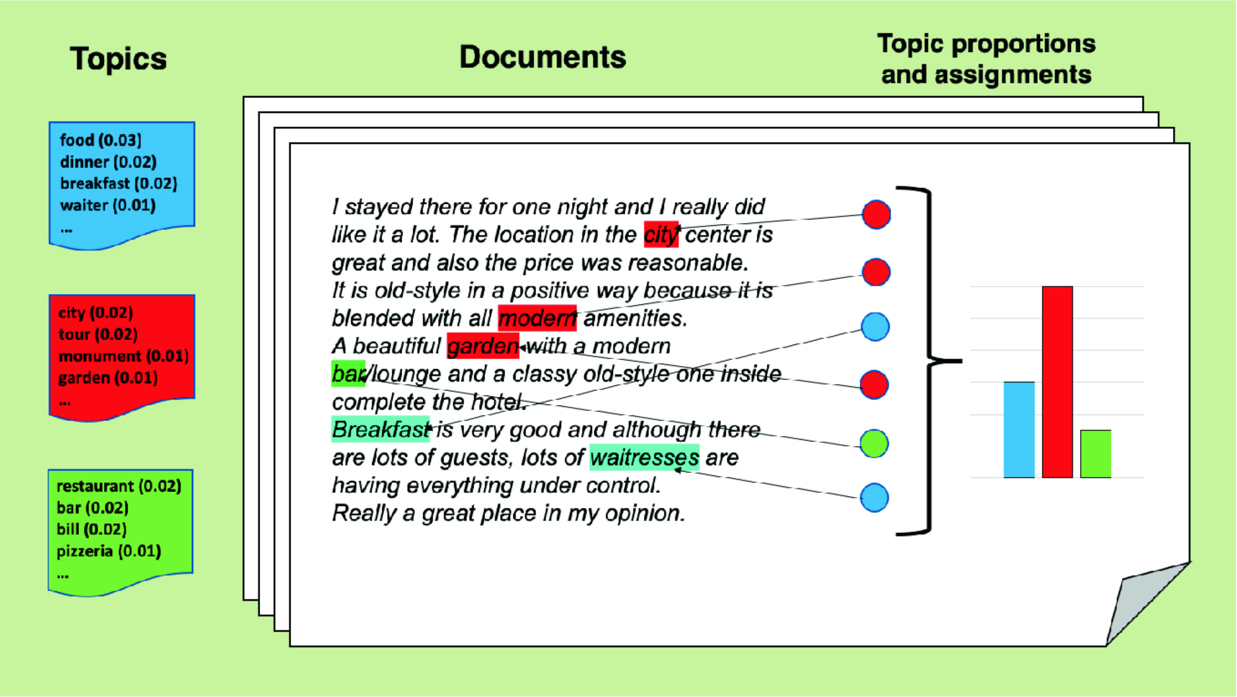
Algorithms for Topic Models

The most common algorithm for topic modelling is called the Latent Dirichlet Allocation (LDA).

LDA takes a document-term matrix as its input.

LDA returns two matrices:

- Prevalence of topics in documents
- Probability of words belonging to topics



Topics

food (0.03)
dinner (0.02)
breakfast (0.02)
waiter (0.01)
...

city (0.02)
tour (0.02)
monument (0.01)
garden (0.01)
...

restaurant (0.02)
bar (0.02)
bill (0.02)
pizzeria (0.01)
...

Documents

I stayed there for one night and I really did like it a lot. The location in the city center is great and also the price was reasonable. It is old-style in a positive way because it is blended with all modern amenities. A beautiful garden with a modern bar/lounge and a classy old-style one inside complete the hotel. Breakfast is very good and although there are lots of guests, lots of waitresses are having everything under control. Really a great place in my opinion.

Topic proportions and assignments

University of South Wales
Prifysgol De Cymru

Topic Modelling

Every document is a mixture of topics.
For example, in a two-topic model we could say “Document 1 is 90% topic A and 10% topic B, while Document 2 is 30% topic A and 70% topic B.”

Every topic is a mixture of words.
For example, we could imagine a two-topic model of the news, with one topic for “politics” and one for “entertainment.”

University of South Wales
Prifysgol De Cymru

Topic Modelling

LDA is a mathematical method for estimating both at the same time: finding the mixture of words that is associated with each topic, while also determining the mixture of topics that describes each document.

Need to decide what words to keep.

Remove common words, stop words etc.

We need to know beforehand how many topics we want. ‘k’ is pre-decided.

This can be an iterative process.

University of South Wales
Prifysgol De Cymru

Topic Modelling Example

The AssociatedPress dataset provided by the topicmodels package, is an example of a DocumentTermMatrix.

This is a collection of 2246 news articles from an American news agency, mostly published around 1988.

We will use the LDA() function from the topicmodels package, setting k = 2, to create a two-topic LDA model.