Social Media Analytics: Topic Modelling and Ethical Considerations

COMP61332: Text Mining
Week 5
Riza Batista-Navarro

Topic Modelling

Finding hidden topics--represented as a group of words--in a corpus of documents

Based on **Latent Dirichlet Allocation (LDA)** by Blei and Jordan, 2003

Latent Dirichlet Allocation (LDA)

an **unsupervised** approach

a **generative statistical model** that produces words associated with a topic

Given that observed data = words in documents

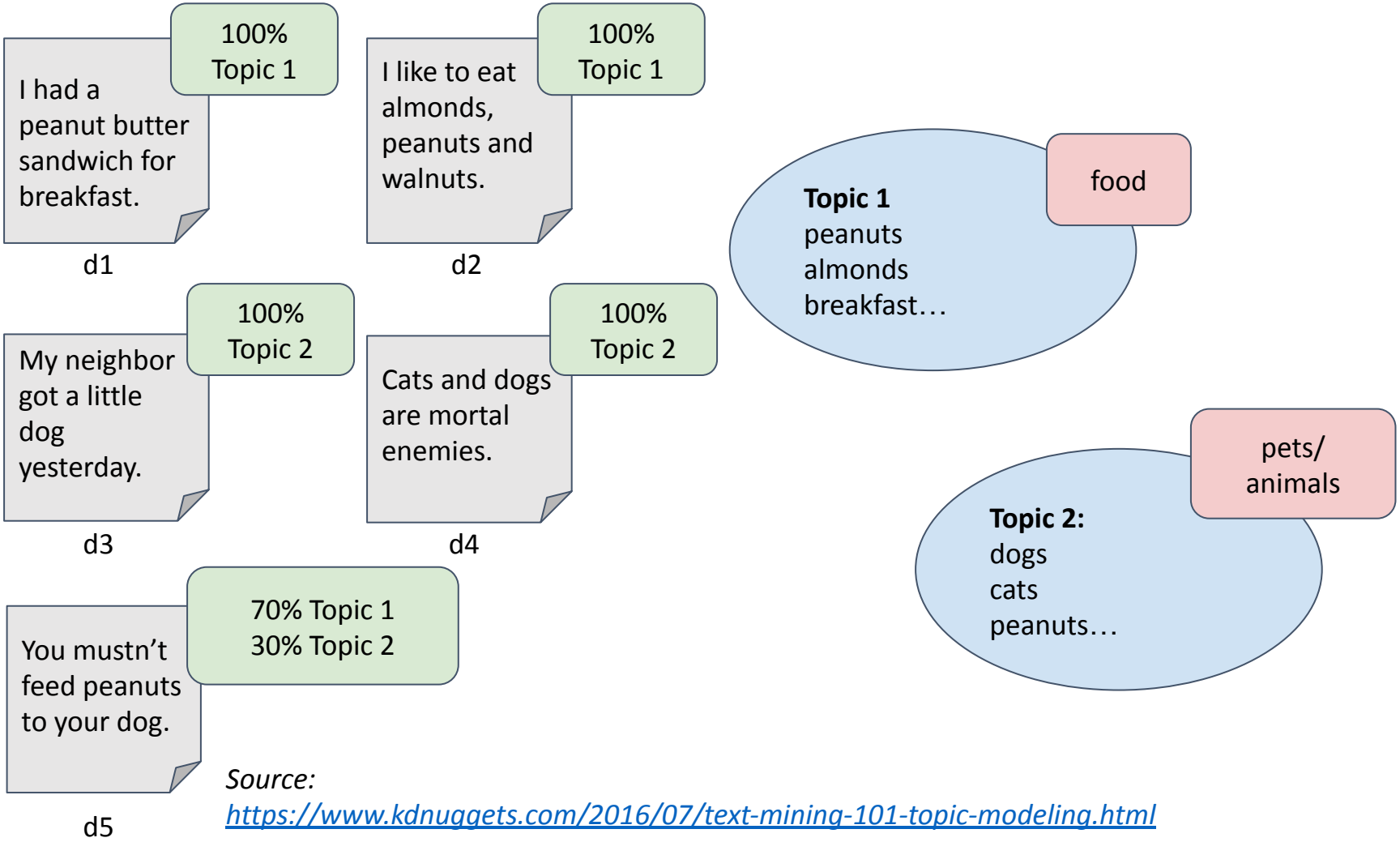
- each document is a mixture of topics
- the presence of each word is attributable to at least one of the topics

number of topics *k* often needs to be pre-defined

LDA Model

Can be interrogated:

- to retrieve a **list of words** (ranked by probability) associated with a topic of interest
- to determine the **probability that a word** of interest is associated with a topic
- to determine the **proportion of a document** that pertains to a topic



Source:
https://www.kdnuggets.com/2016/07/text-mining-101-topic-modeling.html

Ethical Considerations

Just because you can do something does not mean that you should (do it)!

At different stages

- Data collection
- Data analysis
- Data dissemination/publication

Ethical Considerations: Data Collection

Do you need to collect personal information (age, address, contact details, gender)?

How will you store it?

Do you need to share this information? How do you ensure it does not fall into the wrong hands?

Ethical Considerations: Data Analysis

If you need people to manually label/analyse the data, do you use crowdsourcing?

How will you account for bias?

Ethical Considerations: Data Dissemination

If you are writing/publishing a report/paper, how do you make sure you are not giving away identifiable information (e.g., usernames)?

Do you need to give real examples?

Aggregation is key!