Latent Dirichlet Analysis

Latent Dirichlet Analysis

- Current workhorse for topic modelling
- Each document belongs to multiple topics, i.e. has a share of each topic
  - Mixed-membership model
  - "Distribution of Distributions"
- Topic is collection of words that belong together, *not* a topic in semantic sense: rose * 0.05 | chocolate * 0.02 | potato * 0.25

Semantic Flexibility

Doc. 1 unemployment

Doc. 2 inflation

? Which topic/document does "rate" belong to?

Probabilistic approach

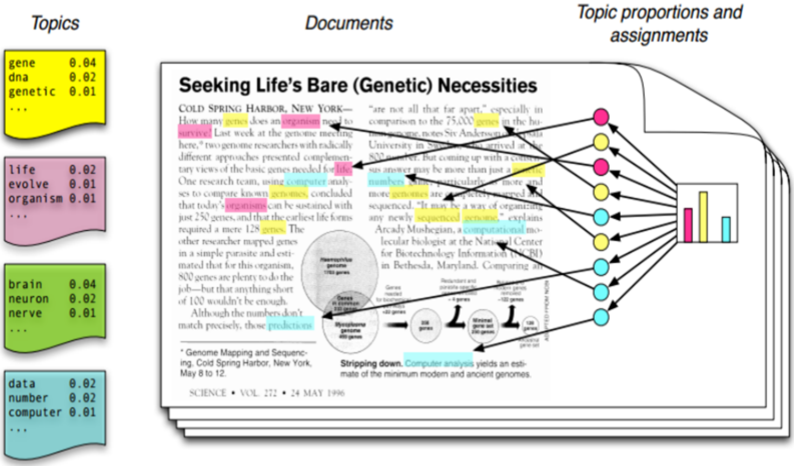
	Topic 1	Topic 2
Document A	10 min	5 min
Document B	5 min	10 min

Probabilistic approach

	Topic 1	Topic 2
Document A	10 min	5 min
Document B	5 min	10 min

- For each word, A draws a topic, which is topic 1 with $p_1 = 10/15 = 2/3$
- Then the word is drawn from the probability distribution associated with topic 1
- Document B draws from the same

Functioning



from: Félix Revert (2018): "An overview of topics extraction in Python with LDA"

LDA

Three important parameters

1. Number of topics
2. Prior of document topic distribution α
3. Prior of topic word distribution β

LDA

Three important parameters

1. Number of topics
2. Prior of document topic distribution α
3. Prior of topic word distribution β

$$\mathcal{L}(w) = \log p(w|\Phi, \alpha) = \sum_d \log p(wd|\Phi, \alpha)$$

$$\text{perplexity}(w) = \exp(-1 \times \mathcal{L}(w))$$

for <u>unseen</u> documents w and topics Φ