Modelagem de tópicos

Fundamentos de Sistemas Inteligentes

Rafael Silva de Alencar

Visão geral

- É uma técnica de classificação não supervisionado utilizada para encontrar padrões em dados de um conjunto de documentos de textos (corpus), através da análise e determinação de *clusters*.
- Objetivo, encontrar temas latentes em um corpus.
- Documentos => Artigos de jornais, tweets, textos em geral.
- Ideia, encontrar uma vasta quantidade de temas em grupos de textos (*Amplified reading*).

Visão geral

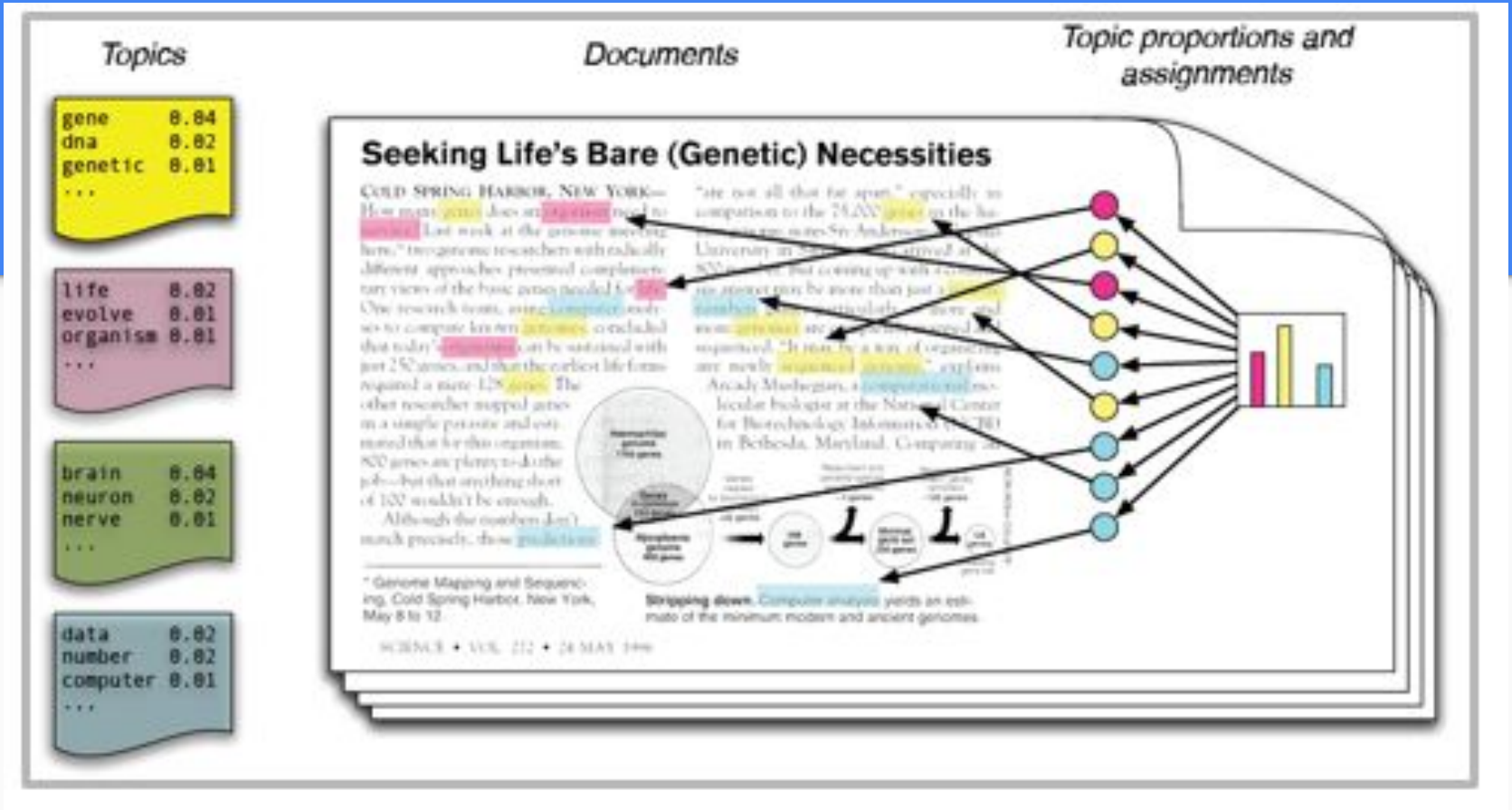
- O que é um tópico?
    - um padrão recorrente de palavras co-ocorrentes (*a recurring pattern of co-occurring words*).

- Exemplo:
    - “health”, “doctor”, “patient”, “hospital” - > Healthcare
    - “farm”, “crops”, “wheat” -> Farming

LDA (*Latent Dirichlet allocation*)

- É a técnica mais popular de modelagem de tópicos.
- Aplicado pela primeira vez por David Blei.
  - Detectar temas em (abstracts) de jornais científicos

- Não diz quantos tópicos existem.
- Não nomeia os tópicos
- Retorna uma lista de palavras associadas com cada tópico.
- *Mixed membership model*



fonte: Blei, David M. "Probabilistic topic models." *Communications of the ACM* 55.4 (2012): 77-84.

Referências

- http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/#topic-modeling-a-basic-introduction-by-megan-r-brett-n-1
- https://monkeylearn.com/blog/introduction-to-topic-modeling/
- https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/?
- https://www.tidytextmining.com/topicmodeling.html
- Blei, David M. "Probabilistic topic models." *Communications of the ACM* 55.4 (2012): 77-84.