

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*

Topics

gene 0.84
dna 0.82
genetic 0.81
...

life 0.82
evolve 0.81
organism 0.81
...

brain 0.84
neuron 0.82
nerve 0.81
...

data 0.82
number 0.82
computer 0.81
...

Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an organism need to survive? Last week at the genome meeting here,* two genomic researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 252 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, says Svend Andersson, a senior University of Stockholm researcher who arrived at the 800 number. But coming up with a minimum number may be more than just a **science** exercise. "Especially in these and some **genomes** are being sequenced and sequenced. It may be a way of organizing, are newly **sequenced genomes**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information, 10101 in Bethesda, Maryland. Comparing an

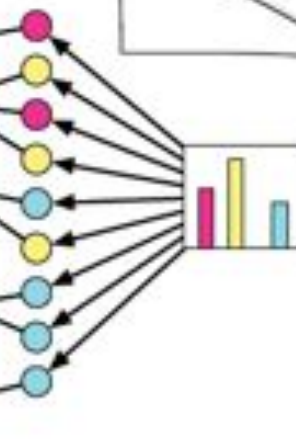


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 212 • 24 MAY 1996

Topic proportions and assignments



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.tidyttextmining.com/topicmodeling.html>
- Blei, David M. "Probabilistic topic models." *Communications of the ACM* 55.4 (2012): 77-84.