

Text Analytics

Introduction to Topic Modelling

By
Kathirmani Sukumar

Definition

*“Topic modelling is a type of statistical model for discovering the abstract **topics** that occur in a collection of documents” - Wikipedia*

“Topic modelling helps to discover and annotate large documents with hidden topics” – Lin Liu et al

Example

▶ Amazon has great customer support. Very quick response

Customer Support

▶ Delivery speed is good. Great packing

Delivery

▶ Got a damaged packing from Amazon. Need a replacement

Delivery

▶ Customer support promptly responded back to my queries

Customer Support

▶ Amazon prime offers quick delivery with free of cost

Delivery

▶ Deals are quite good. Good offers

Deals

Terms define topics

Customer Support

- Support
- Service
- Call center
- Response
- Polite

Delivery

- Speed
- Packing
- Damaged
- Protection

Deals

- Offers
- Sales
- Promotions
- Coupons
- Offer

Amazon reviews



- The first group contains words related to tablet (ipad, android, app, device etc). Hence we can label it as **Tablets**
- The second group contains words related to e-reader (nook, kindly, books, touch, page etc). Hence we can label it as **E-Reader**
- The third group contains words related to battery problems. We can name it as **"Batteries"**
- The fourth group contain words related to wall mounting. We can name it as **"Wall mounting"**

Latent Topics

Topics

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

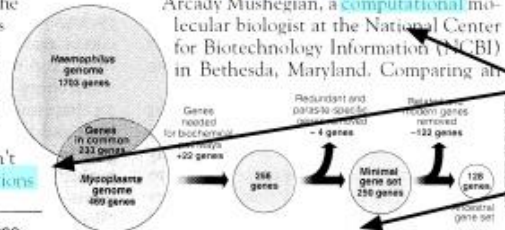
Documents

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 genome 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 250 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, notes Siv Andersson at Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers** game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

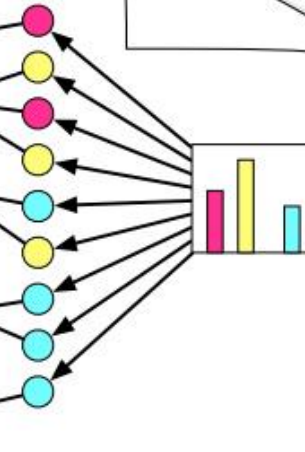


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

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

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments



Popular Methods

- ▶ Latent Semantic Analysis (LSA)
- ▶ Latent Dirichlet Allocation (LDA)



Applications

▶ Clustering

- ▶ Cluster documents in to different groups
- ▶ Creating catalogue automatically for products, books based on description and content

▶ Multi-Tagging

- ▶ Tag files, emails, SMS automatically

▶ Reduce Dimension

- ▶ Group words together to reduce dimension in Document Term Matrix



THANK YOU

All product details and company names used or referred in this work are copyright and trademarks or registered trademarks of their respective holders. Use of them in this work does not imply any affiliation with or endorsement by them.

This work contains a variety of intellectual property rights including trademark and copyrighted material. Unless stated otherwise, Manipal Global Education Services Pvt Ltd ("Company"), owns the intellectual property for all the information provided on this work, and some material is owned by others which is clearly indicated, and other material may be in the public domain. Except for material which is unambiguously and unarguably in the public domain, permission is not given for any commercial use or sale of this work or any portion or component hereof. You may view or download information for personal use only. Any unauthorized access to, review, publish, adapt, copy, share, reproduction, dissemination or other use of the information contained herein is strictly prohibited.

All material on this site is subject to copyright under Indian law and through international treaties, and applicable law in other countries. Company respects the intellectual property rights of others. If you believe your copyright has been violated in such a way that it constitutes a copyright infringement or a breach of a contract or license, we request you to notify our designated representative on the contact column of the website.