

Hierarchical Topic Modeling

Ziv Epstein

`ziv.epstein@pomona.edu`

Pomona College

October 28, 2016

- The world wide web has given us access to large quantities of text data, and often times there is too much text to read manually.
- The goal is to find automated techniques for understanding these large corpora.

Definition

Given a corpus of documents $\mathcal{C} = \{d_1, d_2, \dots, d_n\}$ with vocabulary set \mathcal{V} , a **topic** t_i is a frequency vector $\{f_1, f_2, \dots, f_m\}$ where $|\mathcal{V}| = m$.

An example of Topic Modeling

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 analysis** 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 Sir Andersson, a geneticist at the 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 **sequenced** 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

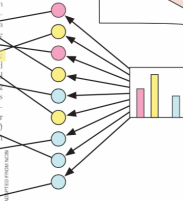


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

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

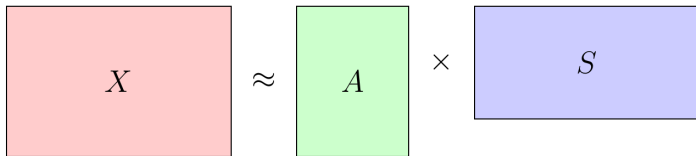
SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments



Non-negative matrix factorization

Given n documents with $|\mathcal{V}| = m$, let $X \in \mathbb{R}^{n \times m}$ be the word/document matrix.

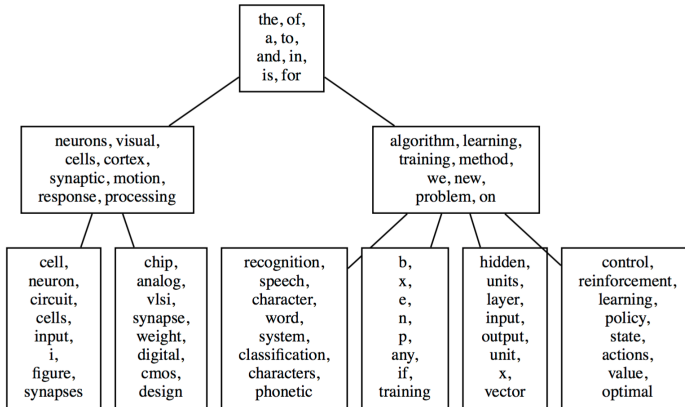

$$X \approx A \times S$$

Let r be the inner dimension, the number of topics, such that $A \in \mathbb{R}^{n \times r}$ is the document/topic matrix and $S \in \mathbb{R}^{r \times m}$ is the topic/word matrix.

- Nice linear algebraic intuition
- Fast implementations

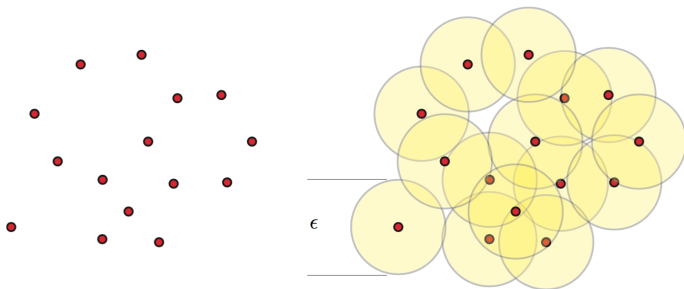
Hierarchical Topic Modeling

What if you want to impose some structure onto the topics, such as a hierarchy?



An Algorithm for Building Hierarchical Topic Modeling

Insight: View the rows of S (word embeddings of the topics) as points in \mathbb{R}^m .

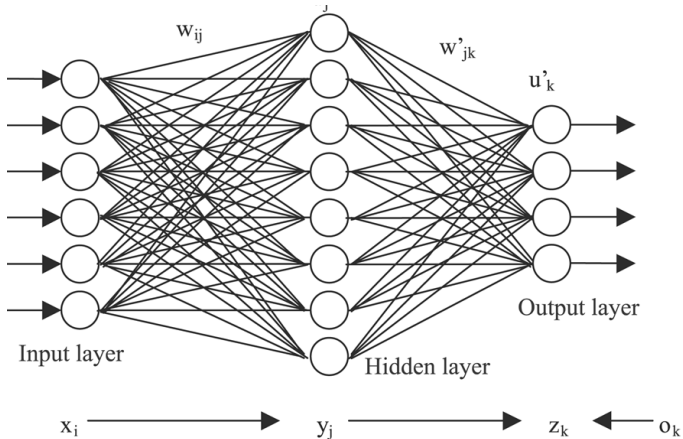


Grow ϵ -balls around each point, and merge leaves when ϵ -balls overlap.

<http://www1.cmc.edu/pages/faculty/BHunter/ziv.html>

Deep Semi Supervised NMF

Quick Review of Neural Networks



Can be thought of as a matrix equation

$$\overrightarrow{output} = \sigma_2(W'(\sigma_1(W(\overrightarrow{input}))))$$

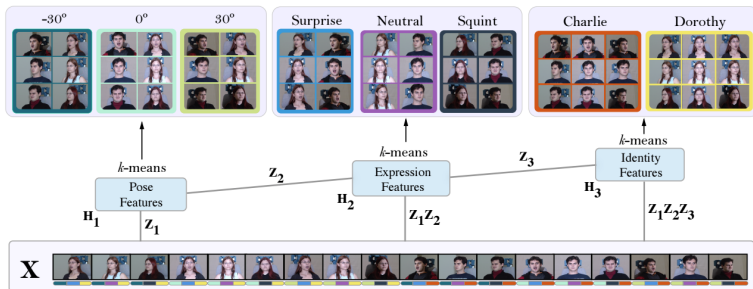
Instead of only factoring $X \approx S \times A$, we instead recursively factor A_i into S_{i+1} and A_{i+1} .

For example,

$$\begin{aligned} X &= S_1 A_1 \\ X &= S_1 S_2 A_2 \\ &\vdots \\ X &= S_1 S_2 \cdots S_n A_n \end{aligned}$$

Key insight : We could already do this, but neural network optimization scheme gives us a robust way to learn S_i

Example: Let X be a matrix of faces. Decompose X into $Z_1 \times Z_2 \times Z_3 \times H_3$. Learns hierarchy of features.



Now let's apply it to topic modeling!