Hierarchical Topic Modeling

Ziv Epstein
ziv.epstein@pomona.edu

Pomona College

October 28, 2016

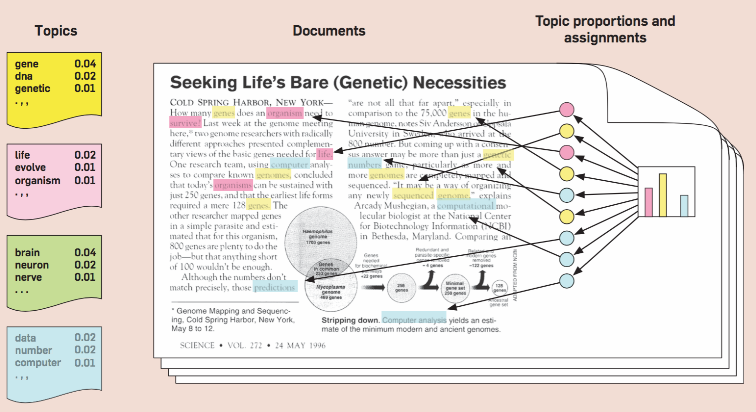
Topic Modeling

- 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

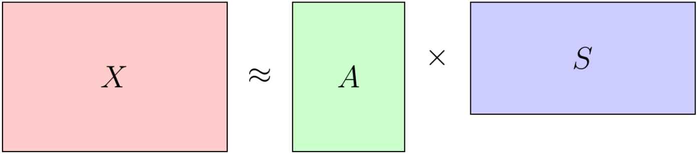
Given a corpus of documents $\mathcal{C} = \{d_1, d_2, \cdots, d_n\}$ with vocabulary set $\mathcal{V}$, a topic t_i is a frequency vector $\{f_1, f_2, \cdots, f_m\}$ where $|\mathcal{V}| = m$.

An example of Topic Modeling



Non-negative matrix factorization

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

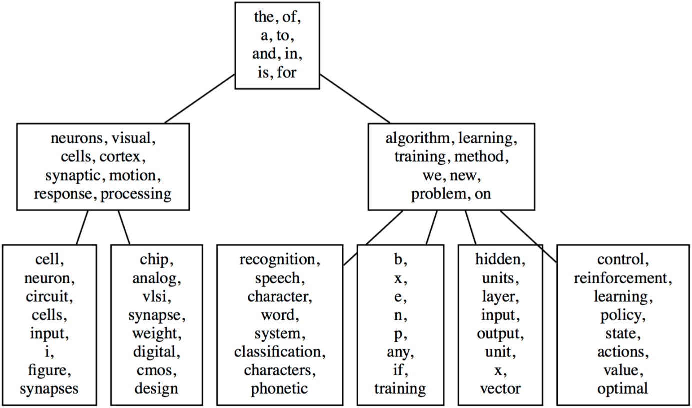


Let r be the inner dimension, the number of topics, such that $A \in \mathbb{R}^{n\times r}$ is the document/topic matrix and $S \in \mathbb{R}^{r\times m}$ is the topic/word matrix.

- Nice linear algebraic intuition
- Fast implementations

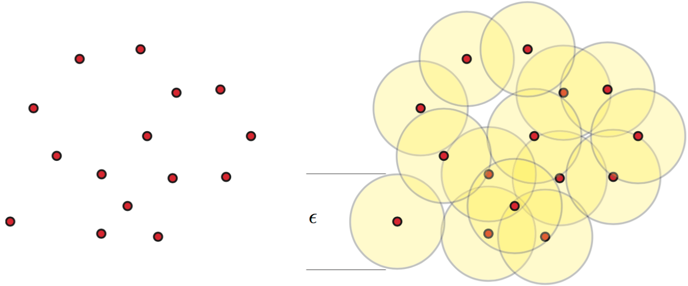
Hierarchical Topic Modeling

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 $\mathbb{R}^m$.

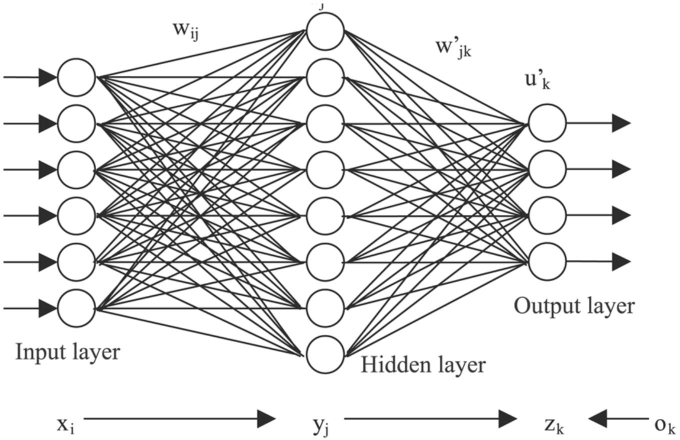


Grow ϵ-balls around each point, and merge leaves when ϵ-balls overlap.

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

Deep Semi Supervised NNMF

Quick Review of Neural Networks



Can be thought of as a matrix equation

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

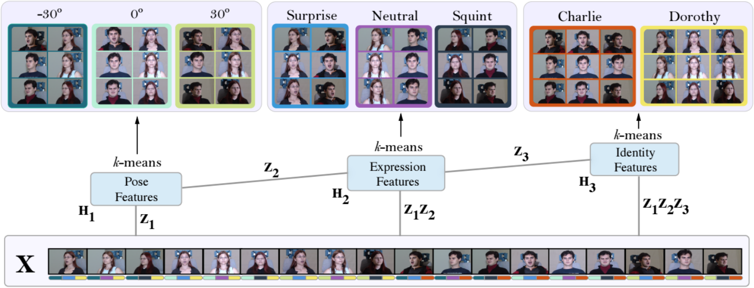
Deep Semi Supervised NNMF

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!