Hierarchical Topic Modelling

Ziv Epstein
ziv.epstein@pomona.edu

September 21, 2016

Abstract

Here we explore an application of non-negative matrix factorization (NNMF) in topic modelling. In particular, we consider a Hierarchical topic models, where topics are nested in a tree-structure. In this document, we formalize this notion, propose an algorithm, and showcase an visualization engine for this doman.

1 Implementation

For a given document matrix V, we use the python library `scikitlearn` to decompose V into document/topic matrix W and topic/word matrix H such that

$$V \approx WH.$$

The `scikitlearn` implementation uses alternating gradient descent with the following objective function to generate optimal guesses for W and H.

$$c(H, W) = \tfrac{1}{2}||X - WH||^2_{fro} + \alpha\lambda||W||_1 + \alpha\lambda||H||_1 + \tfrac{1}{2}\alpha(1-\lambda)||W||^2_{fro} + \tfrac{1}{2}\alpha(1-\lambda)||H||^2_{fro}$$

where $||\cdot||_{fro}$ is the Frobenius norm, $||\cdot||_1$ is the L1 norm, λ is the L1 ratio and α is a free parameter.

From the N topics t_n for $n \in \{1 \cdots N\}$[1], we populate an adjacency matrix A where

$$A_{i,j} = \frac{T_i \cdot T_j}{||T_i||\ ||T_j||}$$

is the cosine similarity between topics i and j. We then define a *threshold vector* σ by sorting all the elements of A.

$$\sigma = \{\sigma_1, \sigma_2, \cdots \sigma_{N^2} \mid 0 \leq \sigma_i \leq \sigma_j \leq 1 \forall i \leq j \text{ and } \sigma_k \in A\}$$

We then create an array of graphs $A^{(k)}$ thresholded using the values of σ, such that

$$A^{(k)}_{i,j} = \begin{cases} 1 & \text{if } A_{i,j} > \sigma_k \\ 0 & \text{otherwise.} \end{cases}$$

Observe that $A^{(1)}$ is the fully connected graph and $A^{(N^2)}$ is the completely disconnnected graph.

[1] observe that t_n is simply the nth row of H

Figure 1: Screen shot of hierarhical topic model application for sample data set

2 Related Work

References

[1] S. Pinker. *The Sense of Style: The Thinking Person's Guide to Writing in the 21st Century.* Viking, 2014. ISBN 0670025852

[2] J. M. Williams and J. Bizup. Style: Lessons in Clarity and Grace. Pearson, 2013. ISBN 0321898680