Classification for the Yelp Data Set Challenge

Nicolas Drizard & Virgile Audi

CS281 Final Project, Harvard University

Objectives

We are taking part in the Yelp Dataset challenge round 6. We chosed to solve the question of infering categories based on the reviews. The goal is to build a finer way to categorize the Yelp businesses.
Our approach consists in using Latent Dirichlet Allocation to build latent features from the text reviews which would depict the categories of each business and turn it into a vector representation which carries local geometry information. We then resorted to building a network of restaurant based on this underlying geometry and cluster this network using a graph theory algorithm.

The Data

- 1.6 million reviews
- 61 000 businesses
- 481 000 attributes (ie hours, parking availabity, take-out, ambience)
- aggregated check-in measurements over time

We cleaned and prepared restaurants data in the city of Las Vegas, NV to grasp in a finer manner the local relationships of the restaurants. This also makes more sense on from the user's perspective.

Latent Dirichlet Allocation

LDA [1] is a three-level hierarchical Bayesian model. Documents are represented as random mixtures over latent topics and each topic is characterized by a distribution over words. The generative process for a document $\mathbf{w}$ in a corpus D is the following:

1. Choose the topics representation $\phi \sim \text{Dir}(\beta)$
2. Choose the number of words: $N \sim \text{Poisson}(\xi)$
3. Choose the distribution of topics $\theta_w \sim \text{Dir}(\alpha_w)$
4. For each of the N words:
   1. Choose a topic assignement $z_{n,w} \sim \text{Multinomial}(\theta_w)$
   2. Choose a word $w_n \sim \text{Multinomial}(\phi_{z_{n,w}})$

We chose an online variational inference algorithm [2] which showed significant speed improvement when compared to the LDA python package using Gibbs sampling. To fit an LDA model with 50 topics on a 4000 document corpus and 10000 words took:

$$\text{Online} = 1m13s \quad \text{Gibbs} = 50m$$

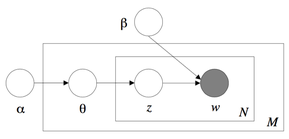
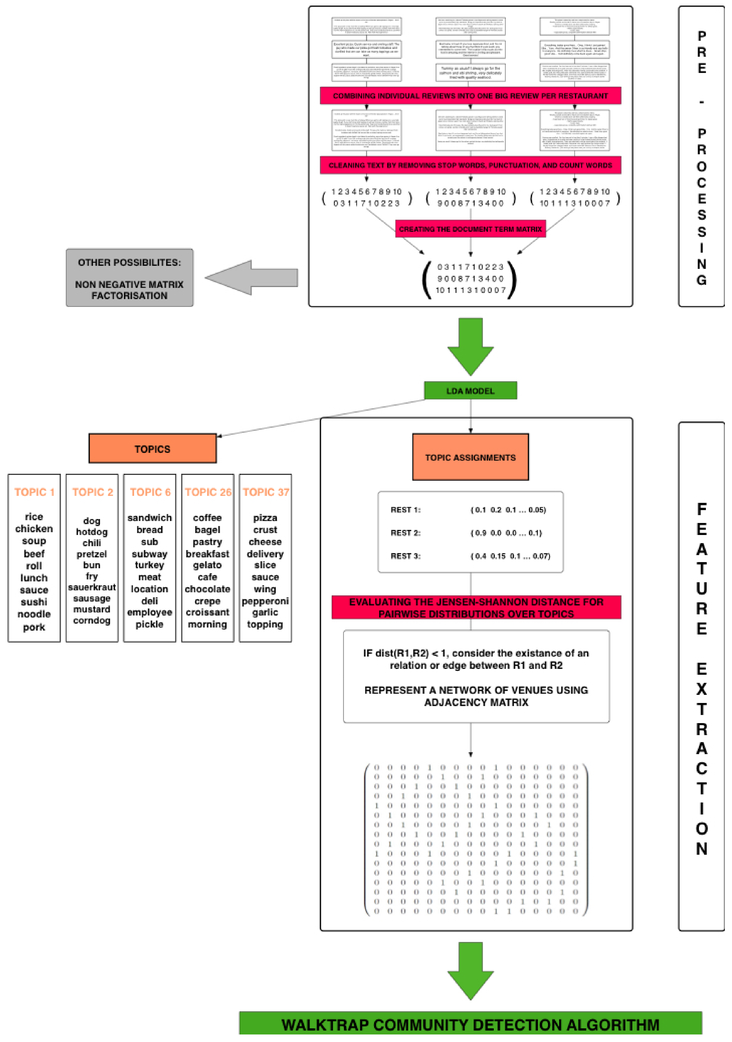


Figure 1: Graphical Model Representation of LDA



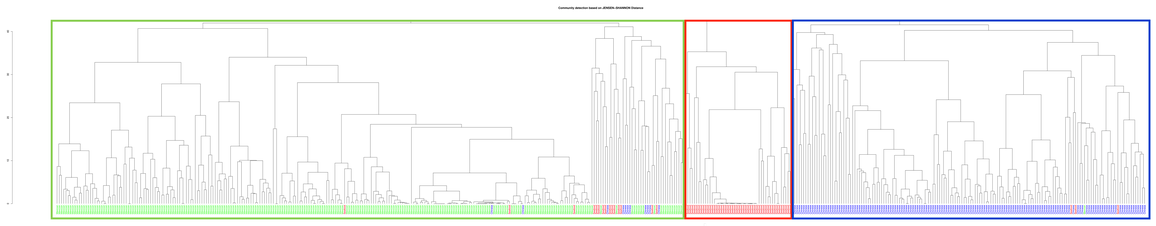


Figure 2: Overall Process

Evaluation

Perplexity is defined as the geometric mean of the log likelihood of the words in the held-out set of docuements given the trained model. In our case, for each document we held out 20% of the words which constitue the test set.

The formula is given by:

$$\text{perp}(D_{test}) = \frac{\sum_{d \in D_{test}} \sum_{w \in d} \log\left(\sum_{t \in topics} p(w|t)p(t|d)\right)}{\sum_{d \in D_{test}} |d|}$$

Graph Theory

The intuition behind the Walktrap algorithm is that random walks on a graph tend to get "trapped into densely connected parts corresponding to communities".[3] It is a clustering algorithm based on the definition of a new metric to evaluate the distance between two vertices in a graph:

$$r_{ij} = \sqrt{\sum_{k=1}^{n} \frac{(P_{ik}^t - P_{kj}^t)^2}{d(k)}}$$

where P_{ik}^t is the transition probability at time t and $d(k)$ is the degree of node k (number of edges incident to the vertex)

Results

- The validation indicate that **50** was the optimum number of topics, some of these topics are plotted in the diagram
- Based on the Shannon distance (symmetric KL divergence) between two documents, we set a threshold of 1 and created the adjacency matrix
- In order to inspect the graph and due to its extremely large size (210^6 edges), we sampled 500 restaurants having among their tags either "Chinese", "Seafood" or "Mexican"
- We ran the walktrap clustering algorithm in R. It initialy found 7 cliques, but a closer insepction of the dendogram, we actually noticed that it subcategories 3 bigger cliques boxed in green, red, and blue, which corresponded almost exactly to the 3 classes of restaurants !

Conclusion

By using a combination of unsupervised learning and graph theory, we managed to retrieve most of the original classification that is up to now still done at hand ! This work could result in an au tomation tool for Yelp to label properly its data based.

Next steps

Futur work could focus on the sub-cliques and try to get a more refine classification by combining them with other features like check-ins, parking availability, etc. As a variant, we could also try a supervised LDA, where we use the existing categories as a response variable associated with each document and infer the joint model of the documents and the responses.

References

[1] M. Jordan D. Blei, A. Ng.
Latent Dirichlet Allocation.
Journal of Machine Learning Research, 3, 2003.

[2] D. Bach M. Hoffman, D. Blei.
Online Learning for Latent Dirichlet allocation.
2010.

[3] M. Latapy P. Pons.
Computing Communities in Large Networks Using Random Walks.
JGAA, 2006.