Nicolas Drizard
nicolasdrizard@g.harvard.edu
CS281-F15

Project Proposal

Due: 23:59pm October 14, 2015

Virgile Audi
vaudi@g.harvard.edu

Introduction

Looking for an applied project, we decided to take part in Yelp Dataset Challenge round 6 ! The dataset is constituted of 1.6 million reviews for 61 thousands business in the UK, the US, Germany and Canada. It also contains 481 thousands attributes such as hours, parking availability or ambience as well as aggregated check-ins over time for the entire base of business. We chose to analyse this data set because it gave us the opportunity to come to conclusions that could eventually be implemented on Yelps platform. In more details, we focus in this project on the classification by categories. When looking for a restaurant, Yelp users enter keywords and get results corresponding to these tag words. But how are these tags associated to each venue? Until now, each owner manually associates the tags to their venue when creating the profile along many attributes that users can then update based on their experiences. Can we use Machine Learning techniques to create a more refine category system? Would it also be possible to find some new categories that are not as obvious as Pizzeria or Fast Food based on the features and characteristics of the venues? How much information can reviews give us on the type of restaurants ? These are some of the questions we are looking to answer in this project.

We now present a road map to the project with possible extensions if time permitted:

1 Data Preparation and Feature Engineering

The given dataset gathers many information under several formats. We first need to extract relevant features of our data and/or pre-process them to apply our algorithms.

1.1 Numerical Feature Extraction

We would like to map most of the data set attributes to features with numerical values to embed them in a given space. Some features are already in this format, e.g. stars, review_count, check-in. We will convert categorical features into binary features, with consideration of only a given top number if there are too many features, for instance for city, state, neighborhood, attributes in business. The reviews would be used through count of specific relevant words.

1.2 Text Processing

For text reviews, since we are not paying attention to who wrote the reviews, we will merge the different reviews that are for a particular venue in order to create a single text document per venue. Each document will then be stripped of usual stop words and punctuation and then transformed into a document-term matrix. We will also try to remove any type of opinion in the review as we will primarily look at the descriptive value of the review. We are indeed looking for any relevant information about the nature of the bar rather than its quality.

1.3 Dimensionality Reduction

This feature engineering part will lead to a large number of features which could be reduced with dimensionality reduction methods. We will project them in a lower dimensionality space with a Principal Component Analysis, and if time we may try also a Non-Negative Matrix Factorization as our data would

be constituted only by positive quantity. These methods may be adjusted based on first results of our classifiers of the next section.

2 Confirming Existing Categories

We will first evaluate the current labeling of the data to evaluate how it is relevant given the features.

2.1 Multinomial Logistic Regression

With the built features, we would like to first confirm that the categorizations used is coherent. We will implement a multinomial logistic regression on a set of filtered categories. For example, we can restrict our approach first to the restaurant with the categories related to the type of food provided which should give around 50 categories.

2.2 Latent Dirichlet Allocation

We will try to categorise the venues solely on their reviews. We will code up an LDA algorithm with online variational inference to extract the topics out the reviews and use them to classify venues.

3 Infering New Subcategories

Once we studied the current label, we would infer latent correlations among the existing categories to come up with new subcategories.

3.1 K-Nearest-Neighbors

The preliminary work on the features will embed the examples in a multi-dimensional space from which we can extract cluster based on the local geometry. We could apply an unsupervised method as the k-nearest-neighbors to infer new subcategories. For instance this clustering method could be used locally in each already known category to infer subcategories or more globally. Another approach if time could be to apply a decision tree, which has the advantage to work on categorical features. One drawback could be its exposure to overfitting.

3.2 Latent Dirichlet Allocation with network component

Using the results about the venues' clustering using the regular K-NN algorithm and taking them into account when performing LDA, we will look at a possible improvement of the performance. In other words, we will try to add an extra node in the graphical model representing the structure of the graph that has an influence on the topic node.

4 Evaluation methods

To assess the performance of the algorithms used in part B, we will divide the data set into a training and test sets in order to make prediction on the test set. We will evaluate the logistic regression through the classical classifiers metrics considering each class separately; we will compute the confusion matrix and extract the accuracy and the recall. Evaluating the new subcategories will be a challenge as it remains very subjective. We can always assess the performance of the LDA model using the perplexity metric but how good this will translate in terms of category classification still needs to be investigate.

5 Possible Extensions

- Online variational inference for LDA (https://www.cs.princeton.edu/ blei/papers/HoffmanBleiBach2010b.pdf) to speed up convergence
- Use Hierarchical Dirichlet Process to relax the assumptions of the number of categories

6 Work Division

Given our personal affinity, we subdivided each part.

- **Nicolas** : *Feature Extraction*, *Dimensionality Reduction*, *Multinomial Logistic Regression* and *K-nn*
- **Virgile**: *Text Processing*, *LDA* and *LDA with network component*