Project Outline

Data Mining Team Project

presented by
Natalie Buchner
Ivan Karachunskiy
Sophia Maier
Timo Strauch

submitted to the
Data and Web Science Group
Prof. Dr. Bizer
University of Mannheim

24th April 2016

Chapter 1

Project Proposal

1.1 Our problem

Why are certain gastronomy facilities more popular and economically successful than others?
Search and rating websites such as trip advisor, yelp etc. present business profiles including specific properties, such as location, price range, opening hours, cuisine, and user based rating. As a result these properties play a decisive role in the decision process of potential customers. When only taking those rather general attributes into account, many businesses seem similar while their economical success, however, often differs significantly. As a consequence we wonder which other determining factors make people choose one restaurant over another and in addition help businesses to retain those customers.
Some of these attributes can be found when taking a look at textual reviews where people state their opinion on extended characteristics of the business and give a concluding rating (e.g. through stars at a scale of 1 to 5). The goal of this project is to extract new success aspects of businesses from reviews and assign weights to them according to their importance within a customer's decision process.

1.2 The Yelp Dataset

The data we will use for this project is provided by yelp in the course of their 7^{th} "yelp dataset challenge" . It includes information about businesses in 10 cities over 4 countries. Besides data such as reviews, tips, business attributes, social network of users and aggregated check-ins over time for each business the set includes pictures from the respective businesses. For our purposes we will predominantly

use the following data:
- 566K business attributes, e.g., hours, parking availability, ambience
- 2.2M reviews and 591K tips by 552K users for 77K businesses
The data set is provided for download on the corresponding yelp data challenge page (`https://www.yelp.com/dataset_challenge`). It contains a file for each single object type, which itself contains one JSON-object per line.

1.3 Our approach

1.3.1 Preprocessing

Integration of JSON data: As the source format of the data is JSON, the dataset cannot simply be added as a RapidMiner repository but has to be transformed into a format RapidMiner can process.
We've considered two options: RapidMiner offers an operator "JSON to data", which takes a collection of JSON documents and transforms it into an example set with one example per document. With further transformation, including transposing columns into rows and transforming the array dimensions into separate columns using regular expressions, the JSON data can be transformed into a tabular structure. Unfortunately, the respective JSON-operator is expecting an array of JSON documents (or separate files for each object) in order to create the correct structure. As the dataset contains a large amount of objects in a one json-object per line style, this implies it would be necessary to further preprocess the provided files: Either splitting them into separate files for each JSON object (per type), or transforming the object type files into an array of all objects.
Thus, a second option would be to directly transform the existing JSON files into a format RapidMiner can work with, using external tools. In their github repository (`https://github.com/Yelp/dataset-examples`), contributers have already provided scripts that execute JSON to csv conversion. These scripts can be used as a starting point to transform the dataset into a usefule format for RapidMiner.

- As a first step we will filter the businesses to look at during the process-building-phase. The processing steps will take too long to apply to the entire dataset several times from the start. we will begin by sampling a rather small but still representative subset of the data to work on. This dataset will contain businesses meta data and textual reviews for around 50 businesses. We will also extract the reviews of 10 businesses as our validation set and another 10 as our test set. When sampling the data we will make sure to include businesses that have positive overall ratings and also businesses with negative overall ratings to create a balanced dataset. Both the validation and test set will need manual evaluation of

the sentiments for the different aspects mentioned in the reviews. Thus we will only include businesses in the validation and test set that have a manageable amount of reviews.
- Once we have the data sampled we will extract the reviews for each business and create a concatenated version of them in a single document. We do not need the information which sentences originated from which review since we want to look at the entire set of reviews to extract new attributes for the business.
- At this stage we will proceed with common text preprocessing methods. Those will include tokenization, stop-word removal and POS tagging. We will follow with our approach a method published by Bancken, Alfarone and Davis in "'Automatically Detecting and Rating Product Aspects from Textual Customer Reviews'". We will make use of the Stanford CoreNLP to apply the preprocessing steps and extract syntactical dependencies on a per sentence base.
- The next step will be to build an algorithm that is able to extract the relevant syntactical dependencies identified by the Stanford Parser. We will build on the relevant dependencies identified in the paper by Bancken, Alfarone and Davis. The result of this preprocessing step are tuples in the form ¡sentiment modifier, potential aspect¿.
- After that we will apply a clustering algorithm (k-means or k-medoids) to cluster tuples together that express a sentiment for the same aspect. To accomplish this we will use a Word-net based similarity metric calles Jcn. An implementation of this metric can be found in the WS4J library. The similarity metric will be calculated for each pair of potential aspects after they were stemmed to reduce the number of different words. The metric will then serve as input to a clustering algorithm which will output clusters that represent different aspects. As a result of this step we can reduce the number of potential aspects to increase our precision based on the cluster-size.
- Afterwards we will make us of a sentiment lexicon for the English language to determine sentiment values for the sentiment modifiers identified in a previous step. As a result we will be able to assign sentiment values to aspect-clusters which were identified in the previous step. We will then extract the most positive and the most negative aspects which form our result.

1.3.2 Algorithms

1.4 Measure of Success

1.5 Possible Results

As project outcome we would like to identify additional success factors and business features important to customers. This knowledge could firstly be used by placing additional decisive attributes into the facility's yelp profile. By that a more in depth representation of businesses is obtained, enabling yelp users to make a faster decision.
Secondly, the gathered insight into business success factors could be interesting for existing and new businesses as well as start-up consultants, as it could serve as a guideline for upgrading their facility in order to attract and retain more customers.
Based on our anticipated findings, further analysis could be performed in order to gain a sophisticated overview regarding the current market situation in general. As a result existing and potential future market trends as well as market niches could be identified.