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 containing 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 looking for a place to drink their coffee or have a nice meal. However, if you only take those rather general properties into account, many businesses seem similar while their economical success often differs significantly. As a consequence we wonder whether there could be other determining factors that make people choose one business over another and 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 certain extended characteristics of the business and often 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. These can be used to give a more sophisticated view on the restaurant, bar or caf.
The knowledge concluded as well as the gathered insight into business success factors could be interesting for existing and new businesses as well as start-up consultants or even venture capitalists as it could serve as a guideline when wanting to upgrade 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 current market situation in general. As a result existing and potential future market trends as well as market niches could be identified.

1.2 The Yelp Dataset

The data we will use for this project is provided by yelp in the course of their 7th "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.

1.3.2 Algorithms

1.4 Measure of Success

1.5 Possible Results

With the project outcome, we want to retreive additional aspects that describe the business ()restaurant, bar, cafe etc.) in more depth and place them into the facility's yelp profile. This could help yelp users to make a faster decision based on further attributes important to them.
(current market situation in general)