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 looking for a place to drink their coffee or have a nice meal. 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 business 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 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 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 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

As project outcome we would like to identify additional success factors and attractive business features important to customers. This knowledge could firstly be used by placing additional attributes important to customers into the facility's yelp profile. By that a more in depth representation of businesses (restaurants, bars, cafes etc.) is obtained, enabling yelp users to make a faster decision based on further determining attributes.

Secondly, 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 the current market situation in general. As a result existing and potential future market trends as well as market niches could be identified.