Project Outline

Data Mining Team Project

presented by
Sophia Mai
Ivan Karachunskiy
Timo Strauch
Natalie Buchner submitted to the

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

April 2016

Contents

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.

Retreive additional aspects that describe the restaurant, bar, cafe in more depth and place them into the facilitie's yelp prole. This could help yelp users to make a

faster decision based on further attributes important to them.current market situation in general

1.2 The Yelp Dataset

The data we will use for this project is porvided by yelp in the course of their 7th "yelp dataset challenge". It includes information about businesses in 10 cities over 4 coutries. Besides data such as reviews, tips, business attributes, social network of users, aggregated check-ins over time for each business the set also 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

1.2.1 Background

The data is provided as an auxiliary file for download on the corresponding page.

1.2.2 Integration of JSON

1.3 Our approach

1.3.1 Preprocessing

1.3.2 Algorithms

1.4 Measure of Success

1.5 Possible Results