Contribution title

Hannes Quade, Johannes Weber, Joshua Braun, Julian Ruppel, and Tobias Rasbold

[1] SRH Fachhochschule, Heidelberg, Germany
https://www.hochschule-heidelberg.de
[2] @uni-heidelberg.de

Abstract. By that time many people have the ambition to be self-employed and want to start a restaurant business. But to run a restaurant can be risky due to lots of different requirements. A well-conceived idea for a restaurant is necessary to be successful and competitive. The purpose of this thesis is to make a recommendation for a new restaurant business in Germany based on various data sources, especially from the community platform of yelp. The collected yelp data includes lots of different significant facts such as the restaurant name, the rating, the text review and the location. For a significant analysis more information was necessary. Therefor we collected other data like the population, the buying power and the rent average for industrial buildings for all cities of Germany. The aim was analyse the collected data and fit a statistical model for the mentioned key figures to find the best cities for starting a restaurant business in Germany. Also a sentiment analysis for each review on yelp was done to decipher the text whether it is positive or negative sentiment. On top the menus of an example city was analyzed to get more insights about the existing restaurants. For the financial aspects it was used a benchmark of Germany's catering industry to get reference values for the investors restaurant.

Keywords: First keyword · Second keyword · Another keyword.

1 Introduction

Nowadays lots of people want to realize their dream to start an own business - often its associated with a ownn restaurant. But an own restaurant business is not for everybody because you have to deal with a lot of different requirements. First of all, a well-conceived concept is absolutely necessary. Therefor you have to ask yourself several types of questions: What kind of customers would you like to have in the restaurant? How should the restaurant look like? Where should it be located? What kind of food should you offer in your menu? All these factors should be well-matched to have a unique feature as a restaurant and to be competitive. The purpose of the thesis is to make a recommendation for a new restaurant business in Germany. Especially to find a good location for opening a restaurant based on different facts as well as reaching the mentioned financial specifications of the investment. A good indicator to discover a new restaurant business is to analyse restaurants that are already running. By that time the world wide web and it's community platforms are inevitable. Customers can leave reviews and rate the restaurant visits online. One of the most popular community platforms is yelp with it's thousands of restaurants and other businesses which can be accessed. Yelp offers an API for developers which was connected with python scripts to collect the restaurant data, e.g. name, rating, location and so on. There were some other important key figures that help to find an appropiate location, especially the dataset with all of Germany's cities and its population as well as the datasets including the buying power and the rent average for industrial buildings. All the collected data were stored in the Google Cloud Platform with the help of python scripts, more precisely in it's Datastore which is a NoSQL document database. From there it was possible to create a postgreSQL istance with the fully-managed relational database service, the so called Google Cloud SQL. For a significant analysis of the raw datasets it was necessary to clean it thoroughly. This was made with the help of python scripts and different datebase queries. The idea was to fit a statistical model for the mentioned key figures to find out the best cities for starting a restaurant business in Germany. In addition to that we implemented a sentiment analysis to categorize each of the users reviews on yelp into positiv or negative sentiment due to the words that were used in the text. Also the menus for restaurants in the case of Mannheim were analyzed to get more insights about the existing restaurants like the favorite items of different restarant categories. For the financial aspects it was used a benchmark of Germany's catering industry which included lots of different key figures like the number of guests per day or the average consumption per guest as well as the revenues and expenses in detail.

2 Literature Synthesis

Essentially there are lots of different papers in the world wide web which are running in that restaurant business direction but most of them do have entirely different points of view. The focus of other papers was only on a specific partition of it like siting of a restaurant or the impact of online ratings.

Last-named aspect belongs to the paper from Havard Unversity[6]. In this the author writes about the significant influence of yelp user ratings on the revenue of restaurants from data of Washington's State Department of Revenue and proves it with statistical models. This is definitely a good insight but focusses as mentioned on just one specific partition of the restaurant business. Further it was not possible to use restaurants revenues as a key figure for this thesis because there are no profits and losses published for independent restaurants in Germany. Only restaurant chains with a huge number of restaurant branches and a yearly turnover were available. Due to too vague estimate this was renounced for statistical purposes in this paper.

In the bachelor thesis [5] it's all about the market analysis of existing restaurants based on various research methods to discover new possible restaurant businesses. This approach relates to the evaluation of surveys for a part of Imatra's population with single questions and does not focus on online review platforms like yelp and its variety of attributes. In addition to that this paper it was not neccessary to find a suitable location because there aldready is one with Imatra, Finland.

In the paper [2] the authors proposed a method to identify the sentiment tendency of each review on yelp based on a statistical model and to classify it to a specific type of restaurant. The analysis of this authors is based on the yelp challenge dataset and does not belong to a specific use case. We also used this approach for our paper and started to train the model with the yelp challenge dataset. After that the collected data from the yelp API was used with certain constrictions because it is only possible to collect three reviews per restaurant each in german and english and not more than 200 characters per text review. As previously mentioned, this paper refers to text analysis only it is just a small part of our paper.

3 Material and Methods

3.1 Usecase Description

SCHULPS are investors who own various restaurants around the world. SCHULPS is headquartered in India and operates various restaurants and catering services. Thanks to a long-term expansion policy in recent years, the investor is also represented in the rest of the world with numerous gastronomic offers. SCHULPS is now also planning to open a restaurant in Germany. An investment of 650,000 euros is planned, whereby the monthly turnover must be at least 40,000 euros. The type of offer has not yet been precisely defined, but should be explained in an understandable way. Since the existing restaurants of SCHULPS differ completely due to the multitude of countries and cultures, the German market is still largely unknown. After some research, SCHULPS became aware of THD's offer to use data from the World Wide Web to search for a suitable location. With the help of search engines and advanced programming techniques, the World Wide Web offers an incredible range of possibilities. The comparison portals and enterprise platforms are more concerned to give their contents into the hand of Communities for different reasons [9]. This also creates more possibilities for access to the data and its evaluation and analysis. The evaluation refers thereby a large part to the differentiation between honestly meant evaluations and the so-called 'Fake News'. THD has given SCHULPS the ability to analyze locations and ratings and make recommendations for a suitable location.

3.2 Data Processing Pipeline

In order for our project, described in subsection 3.1 - *Usecase Description*, to succeed, it is of inestimable value to have a architecture that supports this in the best possible way.

The requirements for our architecture are as follows:

1. Multiple data sources must be supported.
2. All data collected must be stored in its original representation to be able to answer still unknown questions at a later point in time.
3. The areas of data acquisition, analysis and visualization must be separated.

When dealing with large amounts of data, the following two architectural approaches distinguish between each other:

Lambda Architectur: Named after the Greek λ, this approach is characterized by the fact that the streaming data is processed in two ways. On the one hand, the data stream is routed directly to a *Speed Layer*, which is often based on in-memory technology, processes it in real time and makes it available to the *Serving Layer*. Since streams are by definition infinite and the speed layer is expensive and physically finite, the stream is routed from *Ingestion Layer* into the *Batch Layer* in parallel. This layer stores the data persistently and starts a batch job after a defined interval to process the started data and transfer it to the serving layer. So it is the responsibility of the serving layer to mix the aggregated inventory data from the batch layer with the real-time data of the speed layer, which has not yet been processed by the batch layer.[8] [1]

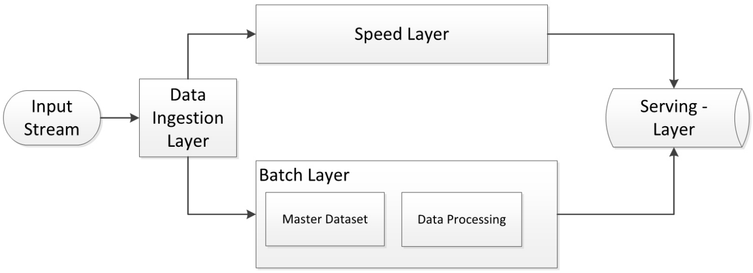

Fig. 1. Scheme of λ architecture[1]

Kappa Architectur: This approach, designed by Confluent co-founder and CEO Jay Kreps, dispenses with batch processing and therefore only requires *Ingestion-*, *Speed-* and *Serving-Layer*. This saves the development and operation of two separate layers and the time-consuming mixing of batch data with live data in the *Serving-Layer*. A prerequisite, however, is that the *Ingestion-Layer* does not only pass through data volatilely, but rather holds the raw data persistently in the *Master-Dataset* as a *Buffer* in order to make the raw data available again in case of a new, not yet precalculated request or change in the *Speed-Layer*. To ensure a correct replay of the messages, the buffer is based on a canonical log, in which only messages can be added unchanged, but already saved messages can no longer be changed or moved in their order. This approach was named after the Greek κ in order to illustrate both proximity and demarcation to the λ architecture. [11] [10]

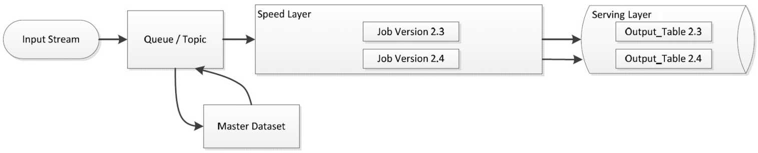

Fig. 2. Scheme of κ Architectur[1]

Both the Kappa and the Lamba architecture cover our requirements. Via the input stream, any number of data sources can be accessed, in the data ingestion layer, all collected data is stored, and then the data is processed and can be persisted in the serving layer. The analysis part can now be executed on the serving layer. The separation into different layers also fulfils our third requirement.

As explained in chapter section 3.2 - *Data sources* our biggest data source is the Fusion Application Programming Interface (API) of Yelp and the Search API of ImmobilienScout24. Due to the fact that Yelp data - with the exception of the Business ID - may not be stored for more than 24 hours,[12] the amount of data collected is limited.

Due to the manageable amount of data that should be available to us in the end and the limited time period to complete this project, we consciously forgo a batch layer - and therefore also the Lambda architecture - in order not to unnecessarily increase the complexity of our architecture but also to reduce the processing time of data at the same time. So we decided for a simplified version of the Kappa architecture.

In step 1, we collect the data from different data sources and write them unchanged into a schemalless database. Step 2 is the transport of the data into a relational database from where it could be visualized and/or analyzed. The transport included not only the simple moving of the data from the source to the target database but also the conversion of the data into a format that has been adapted to the relational database. If the data needs any enrichment with additional attributes to enrich the subsequent analysis and visualization in a more meaningful way, this is happening also in the Transport layer. The last step, the analysis of the data, was done depending on the scenario with the Bi- and Software Analysis Software Tableau or with a Python Script and the libraries `pandas` and `matplotlib`. The following picture shows our previously described variant of the Kappa architecture.

The following chapters are showing the used data sources and describe the single processes to get the data ready for the analysis. Our processes are Data Ingestion, Data Storage and Data Cleaning.

Data sources As mentioned previously, a lot of various key figures were collected for the following data analysis relating to the restaurant business. The main data source with helpful key figures should be at least from one restaurant online review website. In this case the online platform of yelp was chosen. yelp is a popular community site that allows restraurants to present themselves, give visitors a star scale rating of one to five, or give visitors the opportunity to leave comments. For example, the interface Yelp Fusion provides the name of the restaurant, the type of food or the coordinates of the restaurant with latitude and longitude.
In addition to this some other datasets relating to Germany's cities were used for the data analysis:

– CSV file with the population and the area

Purchasing power is an important factor in the catering industry. It makes little sense to open a three-star restaurant in a social focus area or vice versa. A fast food business will not last long between high-end haute cuisine restaurants. In 2016, public television broadcast a study on the quality of life in Germany. These studies included different population, city, and quality reports, with reference to external sources. One of the studies contained information on purchasing power in the cities of Germany. The report led to an external provider who had published this survey [7]. This evaluation was publicly available in the download portal of the website in portable document format and was very well suited to adapting this with a python converter in the postgres database. This attribute complements the model to determine the price level for a restaurant.
For our use case and in general, it is absolutely necessary to know how much rent you can pay in the month or year or whether you can afford a property. In search of a rent index in Germany, we first came across an open-source evaluation in .json format [3]. Unfortunately, this was only the average land prices. Thereupon the possibility opened by an interface to real estate platforms [4] to get more accurate and more up-to-date rental and land prices.

Data ingestion For collecting the yelp data we used the yelp API. Therefor we wrote a script with python. In the following you can some code:

Data storage As a storage for the datasets the Google Cloud Platform was used. Before that, it is necessary to configure it first. You have to create a GCP-Project and then put the collected data in the Google Cloud Datastore which is a NoSQL document database. From there, it was possible to create a postgreSQL istance with Cloud SQL, the fully-managed relational database service of Google Cloud. For setting up a connection a proxy or current ip address is required.

Data cleaning The collected datasets in the Datastore were not messy, but there was still a lot of cleaning activities to do. The yelp data itself had some inconsistencies. Especially the columns price_range and review_count had a lot of empty values. The price_range was filled with the average value of €€. For the column review_count ...

Moreover the data from yelp obviously did not match exactly with the other datasets. The buying power dataset did not include all the cities, that were listed in yelp. To solve this problem the empty values were filled with the average value of Germany. The same problem occured with the rent average dataset which was solved the same way.

4 Analysis Results

In this chapter, the results of data analysis are described through different experiments followed by an evaluation of each result and descriptive visualizations.

4.1 Inductive Analysis

Setup & Assumptions Starting with the technical setup, it should be mentioned that the following analyses were calculated with the R programming language. Each approach follows the same procedure: Preparation and loading of the to be analyzed data via SQL, application of the appropriate analysis methods, interpretation of their results and finally storage of them to be available for further reuse. A foundational assumption of the following approaches is the positive impact of increasing Yelp's star rating on the revenue of a restaurant, which is a essential finding of [6][3]. Based on this, predictive methods are applied to investigate which features lead to good restaurant ratings.

[3] Under certain circumstances like no affiliation of the restaurant to a chain.

1st Approach - Linear Regression

2nd Approach - Classification Tree Partitioning

3rd Approach - Analysis of Potential

4.2 Visualizations

5 Discussion

15-20 Zeilen

Vergleich zu bisherigen Algorithmen Gab es Ausreißer, die nicht untersucht wurden?

Offene Probleme?

Ausblick

API Application Programming Interface

References

1. Berle, L.: Streamingarchitekturen in der praxis: Lambda vs. kappa (2017), https://jaxenter.de/streaming-lambda-kappa-64573
2. Boya Yu, Jiaxu Zhou, Y.Z.Y.C.: Identifying restaurant features via sentiment analysis on yelp reviews (2017), https://arxiv.org/ftp/arxiv/papers/1709/1709.08698.pdf
3. F+B, I.d.d.W.K.: Wohnen in deutschland 2017 (2016), http://www.sparda-wohnen2017.de/was-kostet-eine-eigene-immobilie/uebersicht-deutschland/
4. GmbH, I.S.: Immobilienscout24 api (2018), https://api.immobilienscout24.de/
5. Hasan, S.F.: Market analysis to discover new restaurant business opportunities in imatra region (2015), https://www.theseus.fi/handle/10024/95430
6. Luca, M.: Reviews, reputation, and revenue: The case of yelp.com (2016), https://www.hbs.edu/faculty/Pages/item.aspx?num=41233
7. MB-Research: Mb-research kaufkraft für deutschland (2018), http://www.mb-research.de/
8. Michael Hausenblas, N.B.: Lambda architecture (2018), http://lambda-architecture.net/
9. Ogneva, M.: Social business is just good business (2012), https://www.cmswire.com/cms/social-business/social-business-is-just-good-business-016036.php
10. Pathirage, M.: Kappa architecture (2018), http://milinda.pathirage.org/kappa-architecture.com/
11. Soutier, M.: Die kappa-architektur und noetl (2017), http://www.soutier.de/blog/2017/01/29/kappa-architektur-und-no-etl/
12. yelp.com: Faq - general questions (2018), https://www.yelp.com/developers/faq