MINI PROJECT CANVAS

Title (preliminary): BB

Group members: Suvi Hakala, Henna Roinisto, Reetta Sainio

Workshop # : 1

MOTIVATION

The target group and end-user of our mini-project is practically everyone interested.

Their objective is either to find a business that answers their needs in a specific area or establish a new business in a specific place.

They will benefit by gaining the best business information for their needs, or alternatively, gain information on what would be the best business in the specified location.

DATA COLLECTION

We are planning to use the Yelp dataset to avoid legal issues.

The data is from https://www.yelp.com/dataset/documentation/main.
Our mainly used tables (json files in our case) will be business, review and tip.

Yelp is already managing the data we use and taking care of preserving user anonymity.

PREPROCESSING

The goals of the preprocessing pipeline are to add (and in some cases delete) missing values, aggregate information and find good ways to make use of the data.

Preprocessing steps include e.g. cleaning review and tip strings (stemming etc.), filtering the reviews based on positivity/negativity, combining reviews to tips and locations. We will also categorize the businesses by their industry.

Some data cleaning/wrangling methods we're planning to use are stemming, removing stopwords, dealing with missing values and taking subsets of the data groups.

Our data is already in json so data transformations per say are not needed. We will, however, turn this data into python dictionaries etc.

We might do some feature engineering based on telling our algorithm which words in tips and reviews are truly relevant.

EXPLORATORY DATA ANALYSIS (EDA)

Look at the data!

data has been looked at

We will take some statistics on which industries are more represented than others, scan the reviews and tips from this perspective and possibly focus on some identifiable areas.

Meaningful summarization/visualization properties are areas where services are clustered, which services are clustered and where, as well as positive/negative review clusters to see how we want to handle this information.

VISUALIZATIONS

A map of best businesses (an app/web page) for the end-user is an important visualization as well as a list of these businesses in order. There could also be some slides aimed at to-be business owners.

The map is an interactive visualization.

As for interactivity types, the ability to move and zoom in on the map and scroll the list would be useful to the end user.

LEARNING TASK

(focus on problem definition)

Problem setting definition:

This is a supervised machine learning problem..

This is a classification problem.

We are planning to learn which are/would be the best businesses.

As input we use business' industry, other specified business features, location, reviews, and tips.

LEARNING APPROACH

(focus on solution implementation)

Naive Bayes model seems to be a relevant method for the problem setting, because it's good in text classification and with problems with multiple classes. Also sentiment analysis seems good, because it identifies emotional tone in text (for processing reviews and tips)..

Whether the results match results from other similar areas could be a relevant evaluation metric.

COMMUNICATION OF RESULTS

An interactive web page will benefit the end-user most.

The user can define themselves which are the best and most relevant results for them and they will be shown ordered top 10 etc.

The webpage will include a text field/dropdown to specify the filter, a map and a list.

DATA PRIVACY AND ETHICAL CONSIDERATIONS

(if applicable)

One fairness constraint should be falsification of data due to natural language processing of slang words.

No need to ask for consent to collect data (taken care of by Yelp).

No need for data anonymization (taken care of by Yelp).

Another privacy consideration could be users possibly leaving identifying information on their reviews/tips.

Special treatment is relevant regarding how we choose to split the data and how we cross-validate, as specified in other fields.

ADDED VALUE

There is a possibility for added value from the data we're planning to use.

The data could be used to develop areas and to plan what should change to make the area more prosperous. It will also show which areas lack which types of quality services.

The predictions are turned into added value for the end-user by being an extra feature.

LEGEND

WEEK 1: Data collection/preprocessing

WEEK 2: EDA & visualizations

WEEKS 3-4: Machine/deep learning

WEEK 5: Fairness & data privacy