CS598 Data Mining Capstone [Spring 2020]: task 7

denisa2@illinois.edu

Table of Contents

Introduction and general information

The goal of this task is to leverage data mining technics and to build a small-scale application system that would allow the envisioned end users (i.e., people who will benefit from the results that are generated by data mining algorithms) to upload a new data set and apply at least one algorithm that you developed or experimented with to mine the uploaded data set using a Web interface.

Address and access

The application is available by the following link: http://54.201.50.127:5000/
via any desktop browser with no authentication and available for any regular user.
The application hosted on AWS service.

Application description

During this task I developed a "Yelp wizard" application which allow users to evaluate a dataset of restaurant's reviews in order to find and choose a new restaurant for visit based on topics and cuisine which could be interested for a user.

Dataset description

For the task we will use Yelp's reviews data set "yelp_academic_dataset_review.json" and "yelp_academic_dataset_business.json" with ***703508*** reviews for ***14035*** business. Each business linked to a set of categories like type of business, cuisines and so on. Before topic mining we will pre-process this file and choose review only for venues, which are in "Restaurant" category and extracts the set of cuisines which get us a set of ***239*** cuisines in whole.

By default, the application operates by the whole this dataset but also, a user could upload any subset of this dataset for evaluation.

CS598 Data Mining Capstone [Spring 2020]: task 7

denisa2@illinois.edu

Functions and goals

The key goal of the application is to allow user to find a restaurant which could be interested for him based on data mining algorithms instead of the standard filters which available on Yelp service. So, in the application there are two key functions:

- ***Topic mining*** - which allow user to apply LDA algorithm with different parameters for all reviews in the dataset and choose which topic is interested. Based on this choose the application will show a list of cuisines and a list of restaurants for which chosen topic is a topic with the highs weight. It allows user to choose restaurant based on topics and key words of each topic, instead of normal search by keywords.
- ***Text similarities*** - which allow user to choose cuisine based on measure of similarities between cuisines based on review's texts. It allows user to evaluate a cuisines data set and find an interesting cuisine based on similarities between texts.

Base user's scenario

The main user's scenario is the following:

1) Evaluate and choose topic on the "Topic" page by trying different model's parameters.
2) Evaluate and choose cuisine on the page "Cuisines" by clicking to any cuisines name and evaluating a similarity between cuisines.
3) Evaluate a list of restaurants

Also, a user could upload their own subset of reviews (for example, upload only reviews only with 5 starts) and make an evaluation based on this dataset.

The dataset should be a subset of "yelp_academic_dataset_review.json" - pls use the same format, structure and business IDs as in the original dataset. Or just play with uploaded ones :)

Interface description

The application has a web-based interface with 4 pages and simple navigation. The first page is a "Topic" page with is an access point to the application and a start of customer journey. The general interface organized as 3 areas:

1) Navigation between key pages: Topics, Cuisines, Restaurants - by clicking the buttons a user could navigate between pages and also in the button hi could see how much items are able for evaluation.
2) Parameters panel: here there are parameters which a user could change for each page and visualization.
3) Evaluation area: here a user could evaluate data

CS598 Data Mining Capstone [Spring 2020]: task 7

denisa2@illinois.edu

Topic page

The "Topic" page looks like on the screen:

Here a user could evaluate a result of reviews topic mining, try different parameters for topic mining. Based on this parameter all reviews for all restaurants will be processed.

Parameters:

- Number of topics - 10 / 15/ 20
- Numbers of words for each topic - 10 / 15 / 20

By clicking of any word in the evaluation area a user could make a choose of topic for further analysis. In the navigation area a user could observe how much cuisines and restaurants correspond to this topic and go to other pages.

Cuisines page

The "Cuisines" page looks like on the screen:

Here a user could evaluate a result of cuisines similarity measure and choose a cuisine for further restaurant's list evaluation.

CS598 Data Mining Capstone [Spring 2020]: task 7

denisa2@illinois.edu

On this page a user could see a cloud of tags snapped from the cuisine names, which have a dominant topic selected on the "Topics" page. The font size in the cloud corresponds to how similar the cuisine are to each other. In other words, the larger the name of the kitchen is shown, the more similar it is to the cuisine you have chosen.

- If no topic is selected, you will see all cuisine that have reviews.
- By clicking on any cuisine, a user could select the appropriate cuisine and the list of restaurants will be filtered according to that choice.
- User can also cancel the cuisine selection by clicking on the "Cuisines" button in a header or go to the topics or restaurant selection pages.

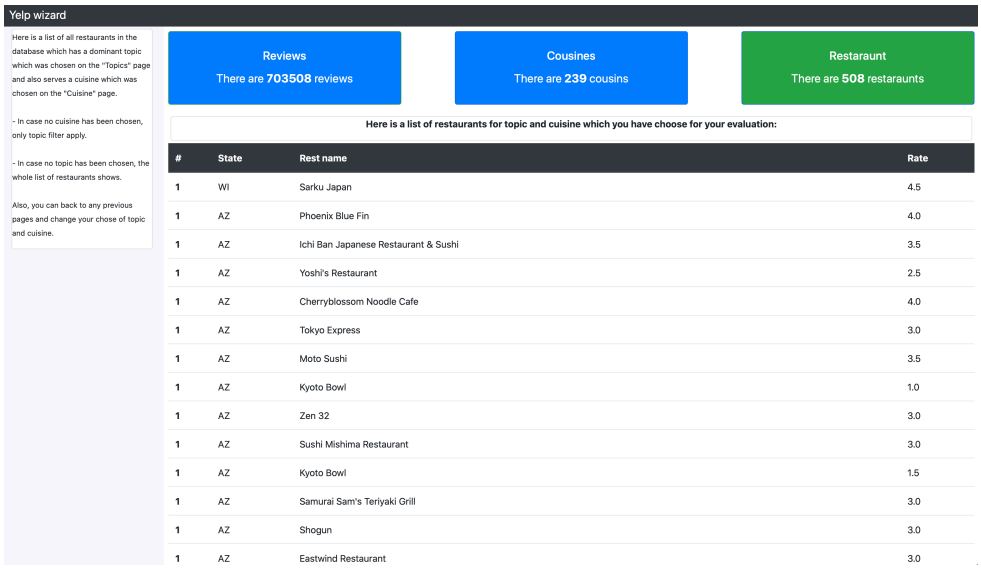
Parameters:

- Similarity function - Euclidean distance / Cosine distance

By clicking of any cuisine in the evaluation area a user could make a choose of cuisine for further analysis. In the navigation area a user could observe how much restaurants correspond to this topic and cuisine and go to other pages.

Restaurants page

The "Restaurants" page looks like on the screen:



Here is a list of all restaurants in the database which has a dominant topic which was chosen on the "Topics" page and also serves a cuisine which was chosen on the "Cuisine" page.

- In case no cuisine has been chosen, only topic filter applies.
- In case no topic has been chosen, the whole list of restaurants shows.

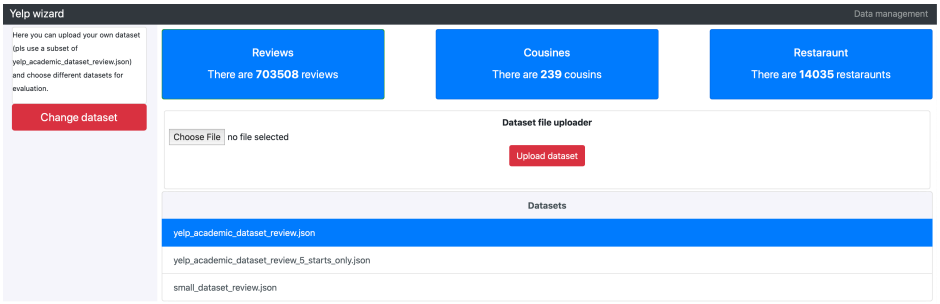
User could back to any previous pages and change your chose of topic and cuisine.

Data manager

The "Data maanger" page looks like on the screen:

CS598 Data Mining Capstone [Spring 2020]: task 7

denisa2@illinois.edu



User could find this page by "Data management" link at the top right corner of any pages.
On this page there are:

- ***Dataset file uploader form*** - a user could upload .json file with new subset of "yelp_academic_dataset_review.json" file
- ***A list of uploaded datasets*** - by clicking on the name of dataset a user could choose a dataset and set it to process by clicking on a "Change dataset" button.

Disclaimers and limitations

In the development process I based on the following assumptions:

- The application is optimized for desktop view, not mobile or any other small display;
- All users operate the same datasets;
- There is no format control for uploaded datasets, there is an assumption it's a subset of "yelp_academic_dataset_review.json";
- There is no function to delete dataset, it's out of scope the task.

Toolkit and libraries

To develop the application, I used:

- ***Python*** - as general language
- ***Flask*** - application server
- ***amCharts*** - for charts and visualisation
- ***Bootstrap*** - for user interface

For data mining I used the following tools and libraries for Python:

- ***Sklearn*** - for classification
- ***Gensim*** - for text processing
- ***Numpy*** - for some additional tool
- ***NLTK*** - for text processing

Conclusion

In this work I developed an application for wide numbers of users who could find an interesting place to eat by using data mining algorithms for text reviews analysis, topic mining and cuisine similarity evaluation. The application has a web-based interface and available on public server.

The application has two novel function (topic mining and cuisine similarity estimation) which not available in existed systems and services and the application could be very useful to many people.