Restaurant Menu Inference

Jason Jennings Nicholas Early

Objective

Develop a system to automatically extract menu items from Yelp restaurant reviews.

Yelp provides businesses the option of linking their menu, and customers of the restaurants often take photographs of the menu. This makes it hard to see, at a glance, what items a restaurant offers. Additionally, many restaurants have neither a link to their menu nor an uploaded menu image.

Our objective is to infer a restaurant's menu from the text of the restaurant's reviews. This will also allow a user to see the items on the restaurant's menu without visiting an external website, or viewing an image. Additionally, it may be used as a means to provide statistics, such as rating or popularity, for each dish on a restaurant's menu. This can help potential customers make informed decisions about restaurants they choose to attend, and what items they might order.

Data Mining Tasks

To implement this system, several data mining concepts will be used.

Classification

This project will likely involve multiple levels of classifiers. One classifier may determine if a sentence is talking about food or not, while the second classifier may determine which terms in the sentence are the food item.

Data Wrangling

The system will involve processing text data (natural language). The data must be cleaned and converted to a consistent format to ensure the system performs well.

Natural Language Processing

The system will be aided by the use of natural language processing techniques, such as part of speech tagging, to determine important features of sentences relevant to our goal (finding noun phrases, for example, will be very helpful).

Deliverables

Website

The primary deliverable will be a website which will allow a user to browse restaurants in the dataset and view the inferred menu items.

Presentation

The presentation will primarily be a demo of the website. Additionally, we may provide examples of data as it goes through the entire process of being classified.

Challenges

Data Labeling

The yelp dataset does not contain any ground truth data for restaurant menu items. This means we will need to hand label data in order to implement this system.

Resolution: We will develop a small application that allows us to quickly label data. We may attempt to use unsupervised learning techniques such as clustering to help us with labeling.

Data Cleaning

People misspell words, speak in incomplete sentences, use inconsistent capitalization and punctuation.

Resolution: We will rely on existing libraries such as NLTK to help us clean data so that it is as consistent as possible.

Natural Language

Natural language is tricky. It may be hard to recognize menu items in complex sentences. We may limit the scope of our project to identifying food items in sentences with certain structure.

Extraneous Data

It is not uncommon for restaurant reviews to contain mentions of dishes from other restaurants. Our system may not be able to eliminate these as possible menu items.

Resolution: We may limit our scope to only identifying potential food items.

Implementation

Our project will primarily be written in python, making use of popular data wrangling, machine learning, and natural language processing libraries such as pandas, scikit-learn, and nltk.

Preliminary Design

An early design (subject to change as we do more research) is that we will implement two stages of classifiers.

The first classifier will identify which sentences in a review are specifically mentioning a food item.

The second classifier will determine which part of the sentence is the food item.

Additionally, some information retrieval techniques may be used to match a possible food item to already identified food items, so the menu does not contain duplicate values.

Evaluation

Quantitative

Our classifiers will be evaluated quantitatively, based on common statistics such as error rate.

We may also compare the results to ground truth results for a small set of restaurants where the ground-truth menu items may be easily available.

Qualitative

Qualitative performance will also be taken into account. Do the menu items inferred from review text seem reasonable?

Project Roles

Nicholas Early

Will focus on classifier determining if a sentence is mentioning food.
Expected to do 50% of the data labeling for the entire project.
Will do front-end design for website.

Jason Jennings

Will focus on classifier determining which terms in a sentence are the possible menu item.
Expected to do 50% of the data labeling for the entire project.
Will do back-end design for website.
Will focus on natural language processing techniques.