Ytrend : Why Did You Rate Me Like That?

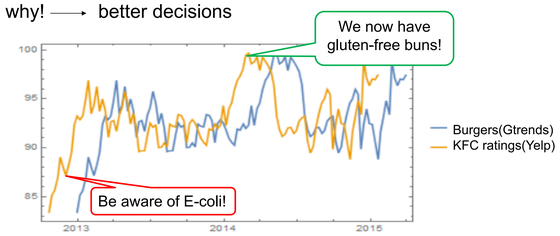
Jasmine Kim
UC Riverside
jkim509@ucr.edu

Shirin Haji Amin Shirazi
UC Riverside
shaji007@ucr.edu

Bailey Herms
UC Riverside
bherm001@ucr.edu

Introduction

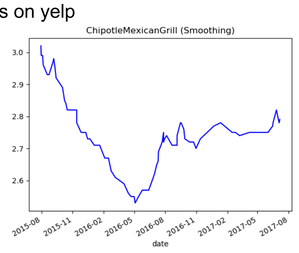
➢Yelp is a trusted source for reviews
➢No explanation on the reason behind the ratings
➢Lets explain why! ⟶ better decisions



It's always better to have more knowledge resources.
1-Yelp stars 2-Yelp text reviews 3-Google Trends

Example : Chipotle Mexican Grill reviews on yelp

Can we find out what happened during the dip?



Problem Definition

"Given the Yelp review data and Google Trends topic searches, we utilize time series similarity search and text clustering to explain certain businesses' Yelp review trends"

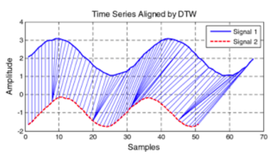
✓ **Where did the data come from?**

Yelp Dataset Challenge provides us with different businesses name, category, location, rating and peoples reviews and stars given.

- business.json reviews.json user.json categories.json

✓ **What tools did we use?**

- DTW (Dynamic Time Warping)



- Z **N**ormalization
- Normalization of Time using Quartiles
- Solve the problem of having time series with different lengths which is required when using DTW
- Rolling Window Mean
  - Average Yelp reviews over a window of multiple data points because the Yelp star review range is too small and discrete
- WordCloud
- Lexical dispersion plot

✓ **What are the main components of our work?**

Our work includes three different aspects of data mining:

1. Pattern recognition in time series
2. Measuring similarity and correlation between different time series
3. Text mining

Proposed Method

Steps:

- Crawl and clean the Yelp Dataset Challenge
- Extract the target business' stars and reviews
- Plot time series of Yelp rating for businesses
  - Check visually if business display's any interesting patterns or dips and peaks to determine if it is interesting
- Once determined interesting, extract the stars and reviews of businesses or categories to compare against target business
- Perform normalization on time series, such as time normalization using quartiles, Z normalization, and rolling window mean, before comparing
- Extract patterns and measure similarity using tools DTW
  - Euclidean distance of corresponding points on two time series
- Mark different time periods corresponding to a sudden fall or rise in the ratings
- Text mine the reviews over the given time periods to find most used words
- Use the words to explore Google Trends and extract time series of the given subjects
  - Use WordCloud as a representation
- Utilize Google Trends time series and the text mining results to explain the Yelp star reviews

Challenges:

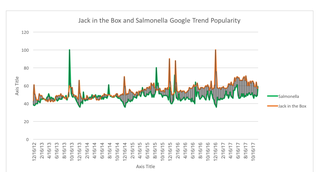
- Yelp star review data is sparse and discrete
- Data per business is sparse
- The range of the Google Trends' time series and Yelp star review time series are not the same
  - Example: Yelp reviews may go down but Google Trends may go up if a particular restaurant gets a lot of bad publicity
- Text reviews are full of sentiment related words and many general phrases
  - Not as professional and precise as predicted.

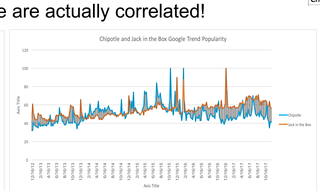
Results

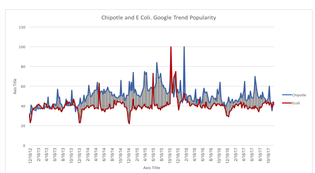
➢ After choosing some of the categories defined in the Yelp data, we used DTW to find similarities between different businesses.
  ➢ What is similar to Chipotle?
  ➢ Euclidean distance from DTW

Chipotle Mexican Grill	chipotle mexican grill	0
Chipotle Mexican Grill	dunkin' donuts	0.12659298
Chipotle Mexican Grill	jack in the box	0.39608637
Chipotle Mexican Grill	dairy queen	0.58918129
Chipotle Mexican Grill	kfc	0.64923899
Chipotle Mexican Grill	domino's pizza	0.9316575
Chipotle Mexican Grill	tim hortons	1.41996658
Chipotle Mexican Grill	jimmy john's	1.47542615
Chipotle Mexican Grill	panda express	1.95603364
Chipotle Mexican Grill	hot dog on a stick	3.12493355
Chipotle Mexican Grill	salsarita's fresh mexican	3.12493355
Chipotle Mexican Grill	mr sub	3.12493355
Chipotle Mexican Grill	fit for life	3.12493355

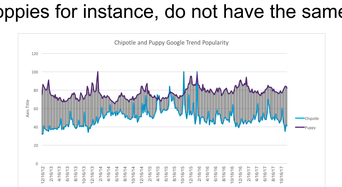
➢ Lets see which of these are actually correlated!





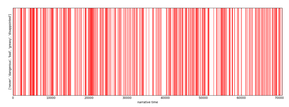


We can see that food poisoning has had a great impact of many businesses like Chipotle Mexican Grill and the correlation is reasonable. In contrast, we can see that other categories like poppies for instance, do not have the same trends!



Now lets utilize text mining!...

- WordCloud was used to find out what words were the most frequent in the reviews during the dips.
  - Mostly negative words ! These can be used to search for specific Google Trends. (example: employee strike and complaints about the service.
- The lexical dispersion plot also shows that many negative words have been used around the same time of the dip.

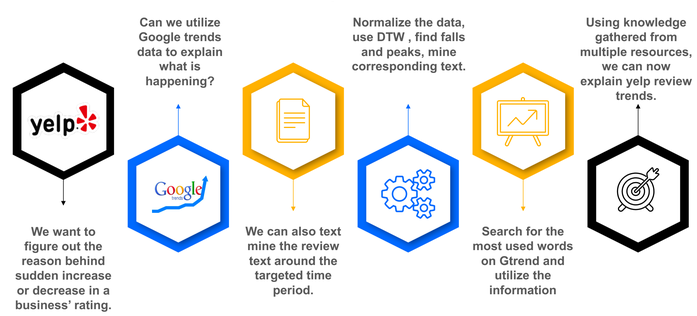


Related Work

- **Fast Similarity Search**(Agrawal et. al)
  - Two subsequences of a time series, are similar if one subsequence fits in the envelope of the other subsequence
  - Performs normalization and then subsequence ordering
- **Clustering Time Series Streams**(Rakthanmanon et al.)
  - Reducing cardinality of the data
  - Bit saving approach
  - Able to cluster without utilizing distance
- **Streaming Pattern Discovery in multiple series**(Berndt et al.)
  - Defines hidden values to project each time series
  - PCA approach
- **Dynamic Time Warping**(Papadimitriou et al.)
  - Defines a window and moves it along the series to find a match
  - Discovering pattern in one series or similarity in more than one
  - Template based recognition
- **Time Series Representation Methods**(Bettaiah et al.)
  - Ten requirements for a visualization method to be useful
  - Piecewise linear approximation
  - Must support classification and clustering
- **Comparing Averages in Time Series Data**(Correll et al.)
  - A method to visualize the data to see the big picture
  - Using color coded averages
  - No hidden variable

Conclusions

"Google Trends, Yelp Stars time series and Yelp review text can be utilized to understand and explain why a certain business has been rated in a specific manner."



UNIVERSITY OF CALIFORNIA
UC RIVERSIDE