CSC 591 – Algorithms for Data Guided Business Intelligence

CAPSTONE PROJECT

Team Members	Unity ID
Anuraag Motiwale	asmotiwa
Parag Nakhwa	psnakhwa
Abhishek Singh	aksingh5

Introduction

Trend Analysis using Time Series:

A time series is a series of data points indexed in time order. Time series analysis comprises A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus, it is a sequence of discrete-time data. Examples of time series are heights of ocean tides, counts of sunspots, and the daily closing value of the stocks. Time series are used in statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, intelligent transport and trajectory forecasting, earthquake prediction, electroencephalography, control engineering, astronomy, communications engineering, and largely in any domain of applied science and engineering which involves temporal measurements. Time series analysis comprises methods for analyzing time series data to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values.

It is important to recognize the presence of seasonal components in the data and to be able to remove them so as not to confuse them with long term trends and the better the stochastic model, the better the prediction. Time series models are used to separate (or filter) noise from signals.

Recommender System:

The recommendations systems are one of the most inevitable part of business intelligence in today’s world. These systems typically follow the generation of suggestions based on the explicit feedbacks from the user and implicit feedbacks associated with the user data.
Recommender systems are used to personalize your experience on the web / mobile etc. telling you what to buy, where to eat or even who you should be friends with.

People tend to like things that are like other things they like, and they tend to have similar taste as other people they are close with. Recommender systems try to capture these patterns to help predict what else you might like.
Social media, video and online news platforms have been actively deploying their own recommender systems to help their customers to choose products more efficiently, which serves a win strategy.

Two most omnipresent types of recommender systems are Content-Based Filtering and Collaborative Filtering. Collaborative filtering is a Unsupervised Learning algorithm which produces recommendations based on the knowledge of users’ attitude to items, that is it uses the “wisdom of the crowd” to recommend items. In contrast, content-based recommender systems focus on the attributes of the items and give you recommendations based on the similarity between them.

Business Intelligence Use Case:

The problem statement that we worked on is as follows:

1. There are multiple types of businesses found in the yelp data set. Our question of interest was to find the most trending places belonging to different business categories in a state using the time series analysis using stl() decomposition. Stl() is an acronym for “Seasonal and Trend decomposition using Losses”.
   Stl() is a very powerful method as it allows the users to control the smoothness of the trend component of time series.
   Stl() also allows to decompose the time series data into the seasonal, trend and noisy components. Our focus on interpreting the trend component of the time series obtained.

   The process pipeline is as follows:

   i. Data pre-processing:
      We started with converting the json data to csv files.
      The yelp data set contains large number of businesses registered of different cities for all the states. We started with first identifying the 34 broad categories of businesses. We categorized all the businesses present in these 34 categories using the ensemble method based on the categories provided by yelp. This left us with all the businesses categorized into 34 broad categories.
      We then required the check-in data to help find out the trend for the businesses. We used regular expressions to get the appropriate check-in format required for weekly period.

   ii. Time Series analysis:

For analyzing the check-in data, we used R. The function used to decompose check-in time series data, we have used “stl” method and extracted the trend component of check-ins. Based on the trend component we find out the top 3 trending businesses belonging to each category who have maximum slope to the trend component associated with it. The results for the trending places is stored in a csv file which contains the business categories and the respective trending businesses.

iii. Post-processing:
Once, we identify the trending places belonging to a business category, we must find out why are they trending. For this, we study the corpus of the reviews and filter the reviews removing the stop words from “sklearn” library for the trending businesses. We then find the most frequent words from the filtered reviews and associate it with the business. For the purpose of visual aid, we have used a word cloud to represent the most frequent words.

<u>Business value of this problem:</u>

- <u>For the business owners:</u>

  If we are able to identify the top trending places for a business category, then this will act as a further incentive for the businesses that are doing well by making them popular among the rest of the users of yelp and since, we are associating a word cloud with every business place, the owners can easily identify the most common words associated with their places and ensure that these things are taken care of.

  For the businesses not doing well, can look up to the words associated with the trending places and can work in that direction to identify the needs of the people and becoming trending.

- For the users:

  The users can easily view the top trending places. This will help the users to identify the most happening places in a new city which they are not familiar with. With the help of the word cloud, the users can also identify the most important traits associated to a particular business.

2. As mentioned earlier, there are various businesses as well as users on Yelp. Daily, millions of users review hundreds of businesses. So our aim is to connect like minded people by recommending reviewers to follow or be friends with. For the purpose of recommendation, we have used **Content based filtering** approach with the help of implicit.als package in Python.

Content Based Filtering:
Content-based filtering methods are based on a description of the item and a profile of the user's preference. In a content-based recommender system, keywords are used to describe the items and a user profile is built to indicate the type of item this user likes. In other words, these algorithms try to recommend items that are similar to those that a user liked in the past (or is examining in the present). In particular, various candidate items are compared with items previously rated by the user and the best-matching items are recommended.

Some applications of Content Based filtering:

- Pandora Radio is a popular example of a content-based recommender system that plays music with similar characteristics to that of a song provided by the user as an initial seed.
- There are also a large number of content-based recommender systems aimed at providing movie recommendations, a few such examples include Rotten Tomatoes, Internet Movie Database, Jinni, Rovi Corporation, and Jaman.
- Public health professionals have been studying recommender systems to personalize health education and preventative strategies.

The process pipeline is as follows:

i. Data Preprocessing:

We extracted all the reviews for North Carolina State containing the attributes “User”, “Business”, “Rating”. Then we made a Sparse Coordinate Matrix to provide it as an input to our model. Due to the vastness of the dataset, we worked only on the NC review data. This approach is easily scalable for other bigger states.

ii. Recommendation using Content Based Filtering:

Firstly, we are acquired the list of businesses (for eg. B1, B10, B21) reviewed by a particular user (for eg. U1). Then we extracted the list of all users who have reviewed or are likely to review those business in the future (for eg. B1- U200, U300, U400; B10- U200, U250, U300; B21- U200, U100, U300, U550). Since User-200 and User-300 have reviewed the same businesses which User-1 has reviewed, the User-1 may want to follow these reviewers and their reviews for other businesses, due to similarity of interests. So we will recommend these users to User 1.

Business value of Recommending Friends:

- For the business owners:

  As our recommender system will connect like minded users, a business which maintains great standards will gain positive reviews from its elite users. So the people following these elite users will get to know about this new business or product which they might want to try, thus helping the business to gain more customers. This will also urge the businesses to continuously improve and maintain their products as there reviews will reach new people exponentially faster due to the recommendation of friends.

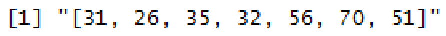
- For the users:

  The users will be able to follow new reviewers and make new friends who have similar interests. This will help them to discover new products which they are not aware of.

RESULTS AND PLOTS:

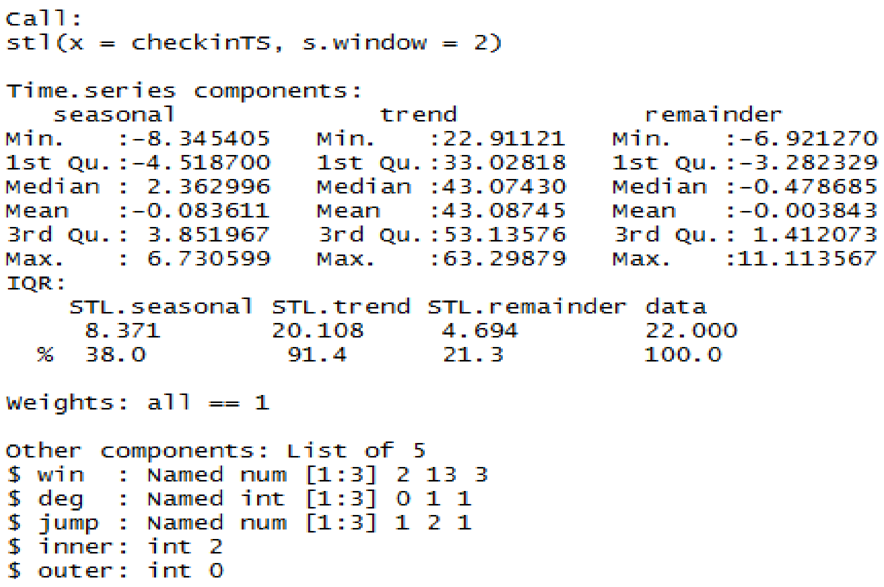
1. TREND ANALYSIS USING TIME SERIES:

Checking counts mon through sun



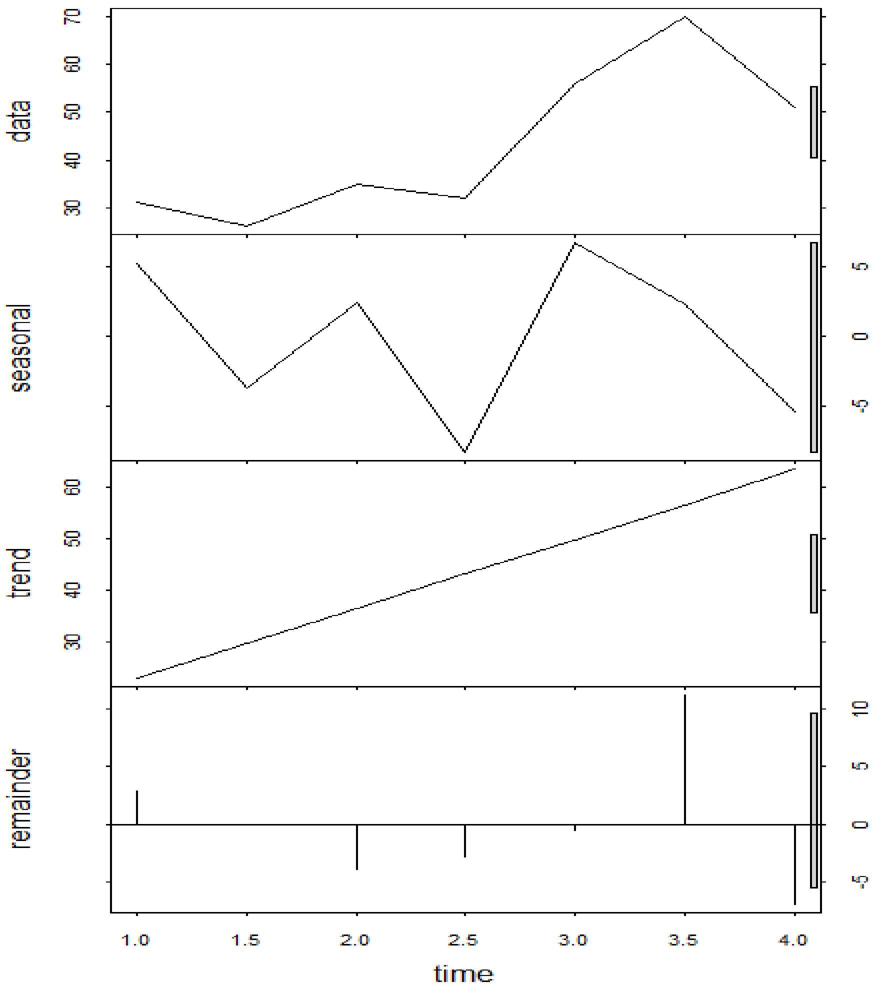
```
[1] "[31, 26, 35, 32, 56, 70, 51]"
```

Time series analysis:

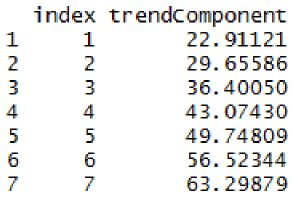


```
Call:
stl(x = checkinTS, s.window = 2)

Time.series components:
   seasonal             trend             remainder
Min.    :-8.345405   Min.    :22.91121   Min.    :-6.921270
1st Qu.:-4.518700   1st Qu.:33.02818   1st Qu.:-3.282329
Median : 2.362996   Median :43.07430   Median :-0.478685
Mean    :-0.083611   Mean    :43.08745   Mean    :-0.003843
3rd Qu.: 3.851967   3rd Qu.:53.13576   3rd Qu.: 1.412073
Max.    : 6.730599   Max.    :63.29879   Max.    :11.113567
IQR:
    STL.seasonal STL.trend STL.remainder data
     8.371       20.108     4.694         22.000
  %  38.0         91.4      21.3          100.0

Weights: all == 1

Other components: List of 5
$ win  : Named num [1:3] 2 13 3
$ deg  : Named int [1:3] 0 1 1
$ jump : Named num [1:3] 1 2 1
$ inner: int 2
$ outer: int 0
```

STL decomposition:



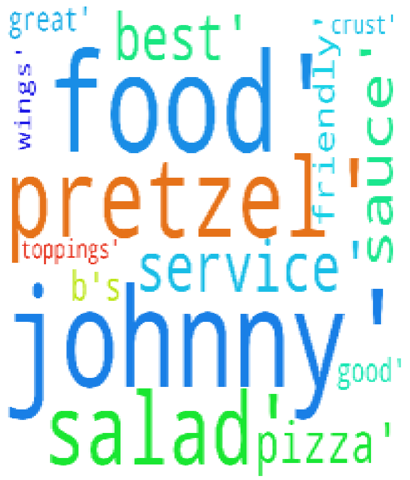
Extracted trend Component:

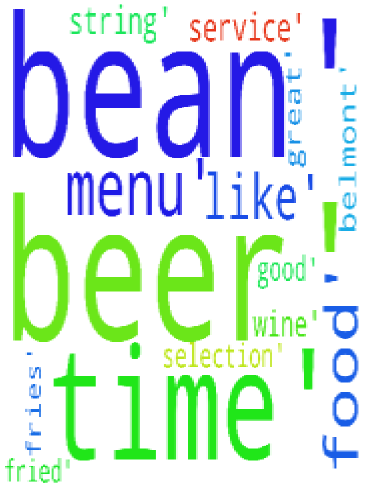


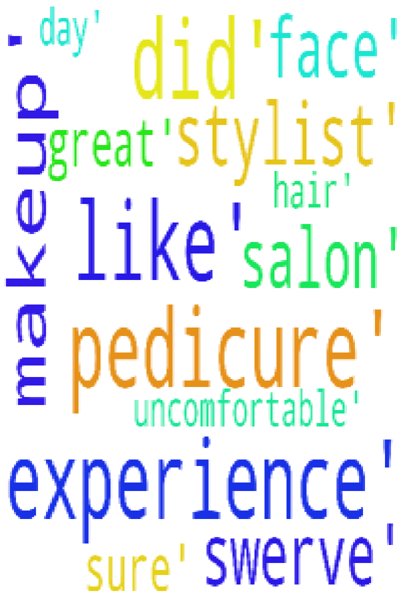
```
  index trendComponent
1     1       22.91121
2     2       29.65586
3     3       36.40050
4     4       43.07430
5     5       49.74809
6     6       56.52344
7     7       63.29879
```

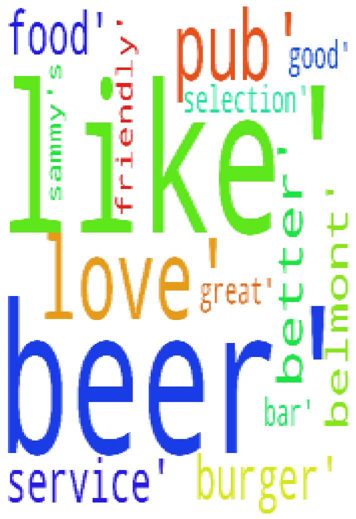
Fit linear model to extract the slope

WORD CLOUDS FOR TRENDING BUSINESSES:

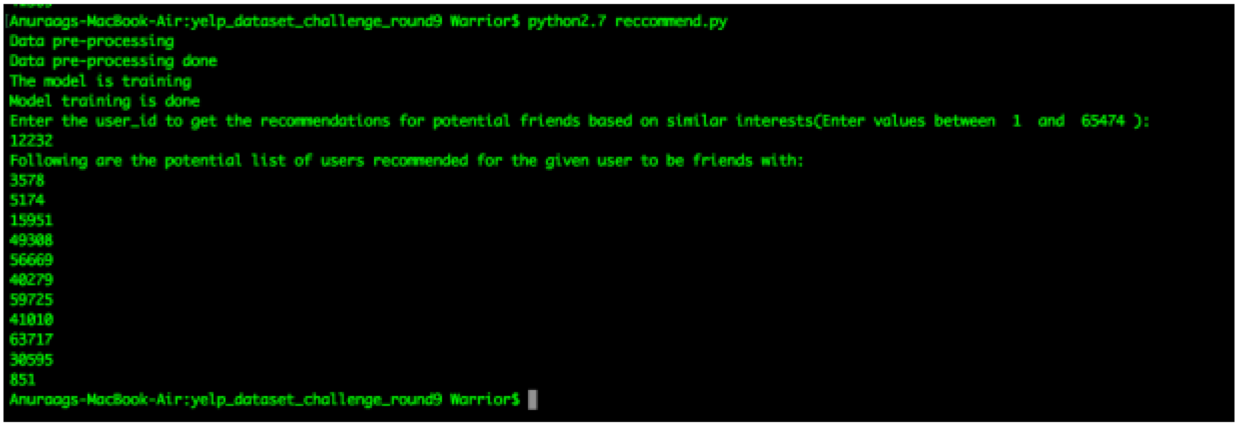








2. RECOMMENDATION SYSTEM USING CONTENT BASED FILTERING:



```
Anuraags-MacBook-Air:yelp_dataset_challenge_round9 Warrior$ python2.7 reccommend.py
Data pre-processing
Data pre-processing done
The model is training
Model training is done
Enter the user_id to get the recommendations for potential friends based on similar interests(Enter values between  1  and  65474 ):
12232
Following are the potential list of users recommended for the given user to be friends with:
3578
5174
15951
49308
56669
40279
59725
41010
63717
30595
851
Anuraags-MacBook-Air:yelp_dataset_challenge_round9 Warrior$
```

Papers for Reference:

1. Time Series and Trend Analysis
   - STL: A Seasonal Trend Decomposition Procedure Based on Loess
2. Recommender Systems using Content based filtering.
   - Collaborative Filtering for Implicit Feedback Datasets
   - Using Content-Based Filtering for Recommendation