CSCI 5408 Final Report

Haoyu Sun B00786821

Yiwei Zhang B00779976

Business idea

In this project, the main idea is about to give visitors appropriate advice and help people start a business locally. Tourists who have came to Halifax on Twitter are generally saying true feeling, so through the project's analysis of Twitter's history, visitors can learn about Halifax's food, attractions, music and mall satisfaction. Based on that, users can optimize their travel itinerary. Companies can also get user preferences through our data analysis to develop more intimate services and more attractive attractions.

Problem Statement

This project aims to search and identify the most popular music, viewpoint, food, and the shopping mall within the Halifax as well as explore the difference in the summer and winter. Twitter, as an online social networking, provides users with a platform to post and interact with messages. The project based on the past five years twitter in Halifax to figure out the main entertainment and leisure activities that the people in Halifax like most in both the winter season and summer season.

Value Proposition

The value of the project can be divided into three main area

1. Business group:
    1) Hotel managers: the hotels can identify the peak season of the travelers.
    2) Real estates managers: the real estates' company can find the best location to build the new commercial buildings.
    3) Travel Agency: Travel Agencies can design better routes for their customers.
    4) Performance Company: Performance Companies can find the best location as well as the best time to hold shows and performances.
    5) Catering industry: The managers of the catering industry can decide which kind of food is included in the restaurant. At the same time, the restaurants can adjust the menus based on the changing taste of customers in different seasons.
    6) Commercial music band: The music bands can decide what type of music they are going to play. Meanwhile, the music bands can figure out the best season to hold a music show.
2. Government:
    1) Non-profit organization: The organizations can find the best time and best location to hold an event that would attract more people to come.
    2) Government: the government can figure which viewpoint might need more maintenance and service managers.

3) Government: the government can figure out when the city needs more maintenance since more people might like to go out in the special period of time during the year.

3. Residents and visitors:

1） Residents: The residents in Halifax can have a guideline to find and join local interest groups.

2) Visitor: The visitors can have a better understanding of Halifax both in winter and in summer.

Therefore, travellers can decide when to visit.

3) Visitor: The visitors can have a guideline for where is the most popular place in halifax. Therefore, travelers can have better plans of visiting.

Implementation and Data Sources

Data gathering:

First, a python program is used to gather the recent five years historical comments from the twitter. The program work principle is that when the Twitter page is entered, the scroll loader starts. If you scroll down, you start getting more and more tweets, all by calling the JSON provider. After imitating, we get the biggest advantage of Twitter search on the browser, it can search for the deepest and oldest tweet.

```
def main(argv):
    if len(argv) == 0:
        print('You must pass some parameters. Use \"-h\" to help.')
        return

    if len(argv) == 1 and argv[0] == '-h':
        f = open('exporter_help_text.txt', 'r')
        print f.read()
        f.close()

        return

    try:
        opts, args = getopt.getopt(argv, "", (
        "username=", "near=", "within=", "since=", "until=", "querysearch=", "toptweets", "maxtweets=", "output="))

        tweetCriteria = got.manager.TweetCriteria()
        outputFileName = "output_got.csv"
```

After we have gathered the data, using TweetManger to help to get tweets in Tweet's model.[4]

```
    def receiveBuffer(tweets):
        for t in tweets:
            outputFile.write(('\n%s;%s;%d;%d;"%s";%s;%s;%s;"%s";%s' % (
            t.username, t.date.strftime("%Y-%m-%d %H:%M"), t.retweets, t.favorites, t.text, t.geo, t.mentions,
            t.hashtags, t.id, t.permalink)))
        outputFile.flush()
        print('More %d saved on file...\n' % len(tweets))

    got.manager.TweetManager.getTweets(tweetCriteria, receiveBuffer)

except arg:
    print('Arguments parser error, try -h' + arg)
finally:
    outputFile.close()
    print('Done. Output file generated "%s".' % outputFileName)
```

We input commands to the terminal and gather data starts. For example, the command line is showing below,

```
Done. Output file generated "food/food for type/winter/fried food/2017 fried food winter.csv".
(venv1) T8DEF:streamtest haoyusun$ python Exporter.py --querysearch "bakery halifax" --since 2013-11-01 --until 2014-04-30 --maxtweets 2000 --output 'food/food for type/winter/bakery/2013 ba
kery winter.csv'
Searching...

More 100 saved on file...

More 89 saved on file...
```

Before we run this command, we need to set the keywords for query search, time of the tweets, the location of files and file name of the new output. We only input the keyword into the search field with Halifax and to get the limited result (maximum is 2000). We divide the season into summer (from May to Oct) and winter(from November to Spring) for comparing the performance of each stuff in different season. We put file address at the end of the command.

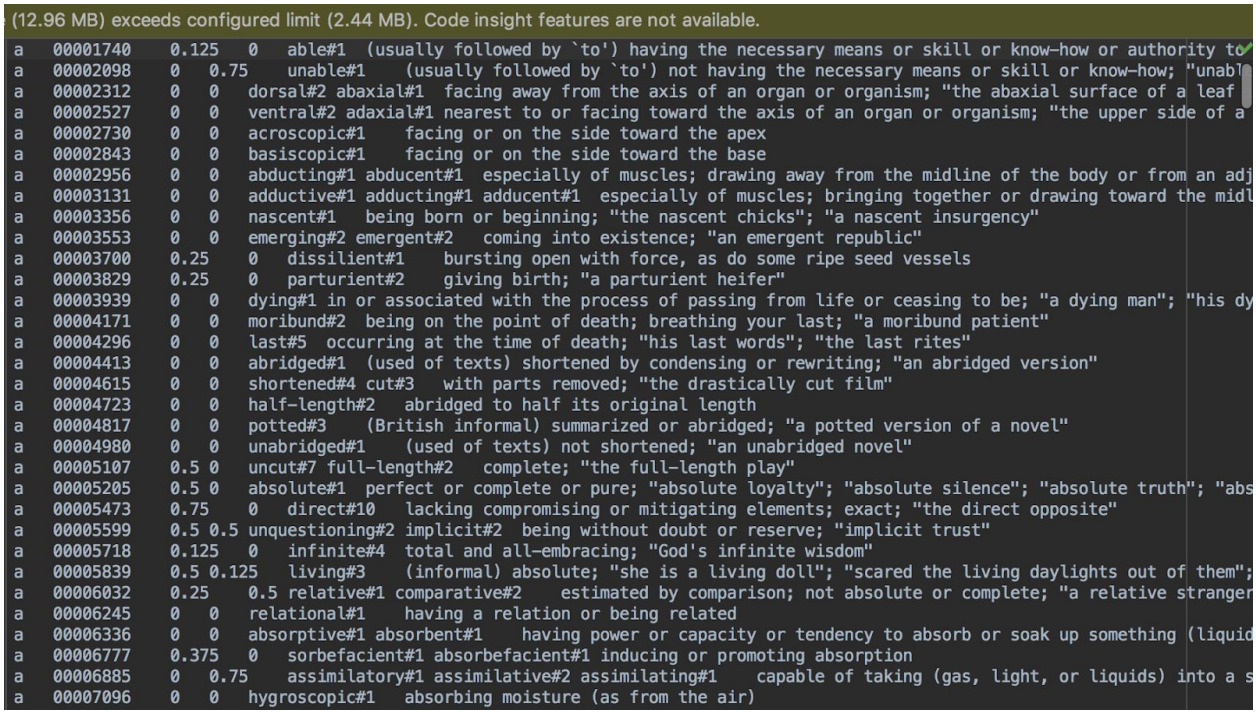
2. Data clean

   The data document must be loaded by data analysis tool, so we use pandas by python3 to do the data clean.

```
data = pd.read_csv('2017 boiling summer.csv', sep=';', error_bad_lines=False)

countNegative=0
countPositive=0
countNetural=0


df1 = data['text']

with open('2017 boiling summer output.csv', 'w') as outfile:
    writer = csv.writer(outfile)
    writer.writerow(["text","Sentiment","Score"])

    for row in df1.iteritems():
        s = SentimentAnalysis(filename='SentiWordNet.txt', weighting='geometric')
        textt = row[1]
        if s.score(textt)==0.0:
                senti="netural"
                countNetural=countNetural+1
        if s.score(textt)<0.0:
                senti = "negative"
                countNegative=countNegative+1
        if s.score(textt) > 0.0:
                senti = "positive"
                countPositive=countPositive+1
        writer.writerow(
                [textt.encode('unicode-escape'),senti, s.score(textt)])
    writer.writerow([ countNetural,countNegative,countPositive])
```

The Algorithm

The project used the sentiment analysis to get the attitude of the tweet text gathered from the previous steps. The basic dictionary is SentiWordNet 3.0



```
(12.96 MB) exceeds configured limit (2.44 MB). Code insight features are not available.
a   00001740    0.125   0   able#1  (usually followed by `to') having the necessary means or skill or know-how or authority to
a   00002098    0   0.75    unable#1    (usually followed by `to') not having the necessary means or skill or know-how; "unabl
a   00002312    0   0   dorsal#2 abaxial#1  facing away from the axis of an organ or organism; "the abaxial surface of a leaf
a   00002527    0   0   ventral#2 adaxial#1 nearest to or facing toward the axis of an organ or organism; "the upper side of a
a   00002730    0   0   acroscopic#1    facing or on the side toward the apex
a   00002843    0   0   basiscopic#1    facing or on the side toward the base
a   00002956    0   0   abducting#1 abducent#1  especially of muscles; drawing away from the midline of the body or from an adj
a   00003131    0   0   adductive#1 adducting#1 adducent#1  especially of muscles; bringing together or drawing toward the midl
a   00003356    0   0   nascent#1   being born or beginning; "the nascent chicks"; "a nascent insurgency"
a   00003553    0   0   emerging#2 emergent#2   coming into existence; "an emergent republic"
a   00003700    0.25    0   dissilient#1    bursting open with force, as do some ripe seed vessels
a   00003829    0.25    0   parturient#2    giving birth; "a parturient heifer"
a   00003939    0   0   dying#1 in or associated with the process of passing from life or ceasing to be; "a dying man"; "his dy
a   00004171    0   0   moribund#2  being on the point of death; breathing your last; "a moribund patient"
a   00004296    0   0   last#5  occurring at the time of death; "his last words"; "the last rites"
a   00004413    0   0   abridged#1  (used of texts) shortened by condensing or rewriting; "an abridged version"
a   00004615    0   0   shortened#4 cut#3   with parts removed; "the drastically cut film"
a   00004723    0   0   half-length#2   abridged to half its original length
a   00004817    0   0   potted#3    (British informal) summarized or abridged; "a potted version of a novel"
a   00004980    0   0   unabridged#1    (used of texts) not shortened; "an unabridged novel"
a   00005107    0.5 0   uncut#7 full-length#2   complete; "the full-length play"
a   00005205    0.5 0   absolute#1  perfect or complete or pure; "absolute loyalty"; "absolute silence"; "absolute truth"; "abs
a   00005473    0.75    0   direct#10   lacking compromising or mitigating elements; exact; "the direct opposite"
a   00005599    0.5 0.5 unquestioning#2 implicit#2  being without doubt or reserve; "implicit trust"
a   00005718    0.125   0   infinite#4  total and all-embracing; "God's infinite wisdom"
a   00005839    0.5 0.125   living#3    (informal) absolute; "she is a living doll"; "scared the living daylights out of them";
a   00006032    0.25    0.5 relative#1 comparative#2    estimated by comparison; not absolute or complete; "a relative stranger
a   00006245    0   0   relational#1    having a relation or being related
a   00006336    0   0   absorptive#1 absorbent#1    having power or capacity or tendency to absorb or soak up something (liquid
a   00006777    0.375   0   sorbefacient#1 absorbefacient#1 inducing or promoting absorption
a   00006885    0   0.75    assimilatory#1 assimilative#2 assimilating#1    capable of taking (gas, light, or liquids) into a s
a   00007096    0   0   hygroscopic#1   absorbing moisture (as from the air)
```

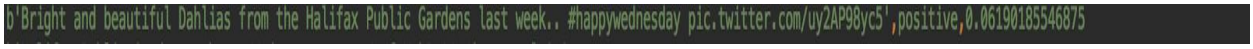
SENTIWORDNET 3.0 is a lexical resource explicitly devised for supporting sentiment classification and opinion mining applications[1]. The SENTIWORDNET 3.0 is not only publicly available but also aiming for research purpose. The widely used of the SENTIWORDNET 3.0 show the creditable of the dictionary.

The project used a python program to do the sentiment analysis. Sentiment scores are between -1 and 1, greater than 0 for positive and less than 0 for negative. This Dictionary-based sentiment analysis does not perform as well as a trained classifier, but

it is domain-independent, based on a priori knowledge of words' sentiment values.[2].

There are three example results from the sentiment analysis:

Positive score sentence:



Negative score sentence:

Neutral sentence:

The extra dependency is the nltk including tokenizers. And the mathematical way to calculate the score the geometric method.

```
def geometric_weighted(self, score_list):
    weighted_sum = 0
    num = 1
    for el in score_list:
        weighted_sum += (el * (1 / float(2 ** num)))
        num += 1
    return weighted_sum
```

At the same time, the program handles negations and multiword expressions.

```
if (self.is_multiword(word_minus_two)):
    if len(scores) > 1:
        scores.pop()
        scores.pop()
    if len(neighborhood) > 1:
        neighborhood.pop()
        neighborhood.pop()
    word = '_'.join(word_minus_two)
    pos = 'unknown'

elif (self.is_multiword(word_minus_one)):
    if len(scores) > 0:
        scores.pop()
    if len(neighborhood) > 0:
        neighborhood.pop()
    word = '_'.join(word_minus_one)
    pos = 'unknown'

# perform lookup
if (pos in impt) and (word not in stopwords):
    if pos in non_base:
        word = wnl.lemmatize(word, self.pos_short(pos))
    score = self.score_word(word, self.pos_short(pos))
    if len(negations.intersection(set(neighborhood))) > 0:
        score = -score
    scores.append(score)
```

Visualizations and data analytical tools

This project used the Tableau to do the visualizations as well as data analysis.

Tableau is a software that produces interactive data visualization products focused on business intelligence[3].

The criteria of the analysis are based on to main standards:

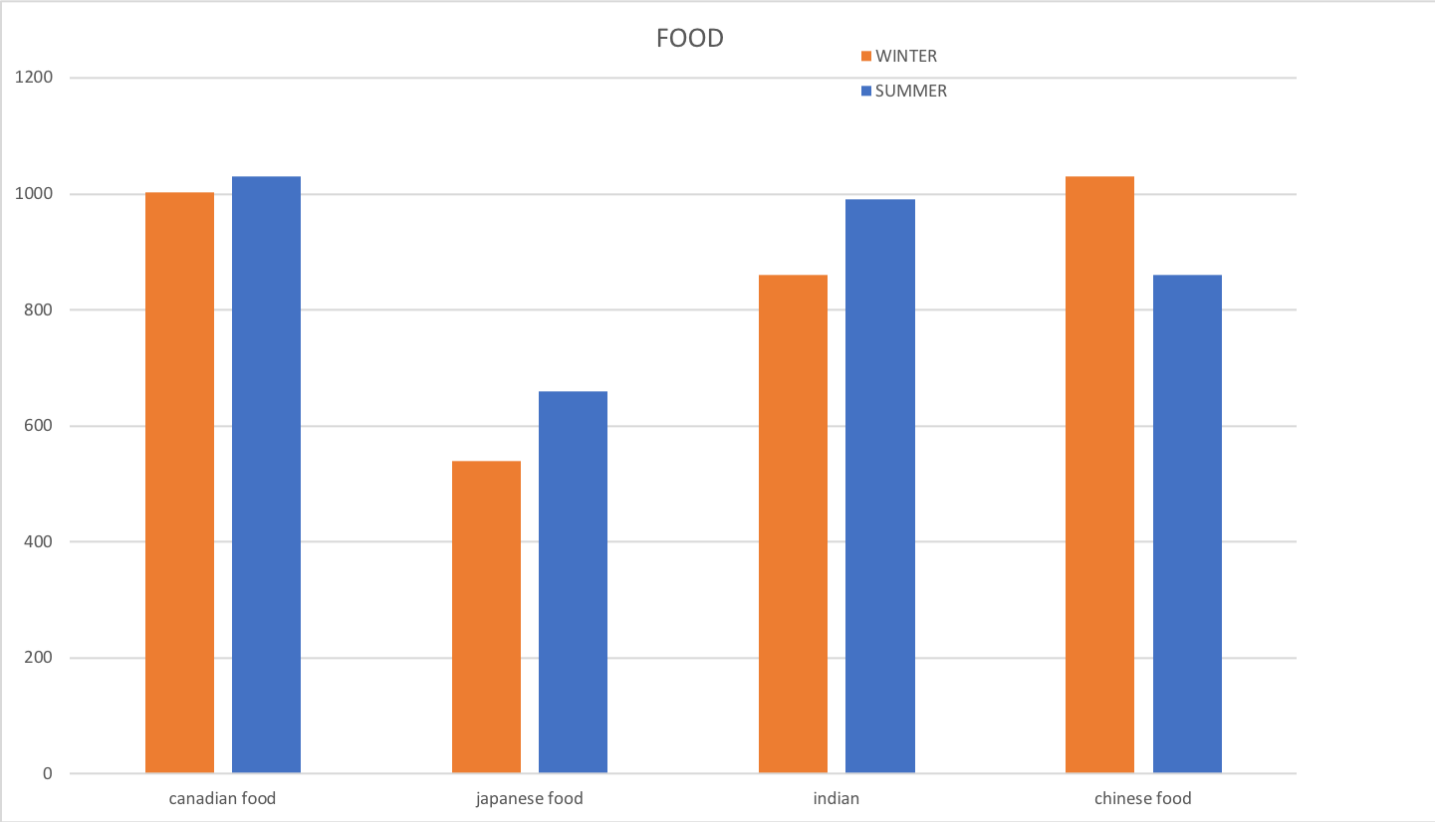
1. The number of tweets in the special time period.
2. The satisfaction score based on the dataset.

   The equation to get the satisfaction score is:

$$satisfaction\ score = \frac{number\ of\ \ positive\ tweets}{total\ number\ of\ tweets}$$

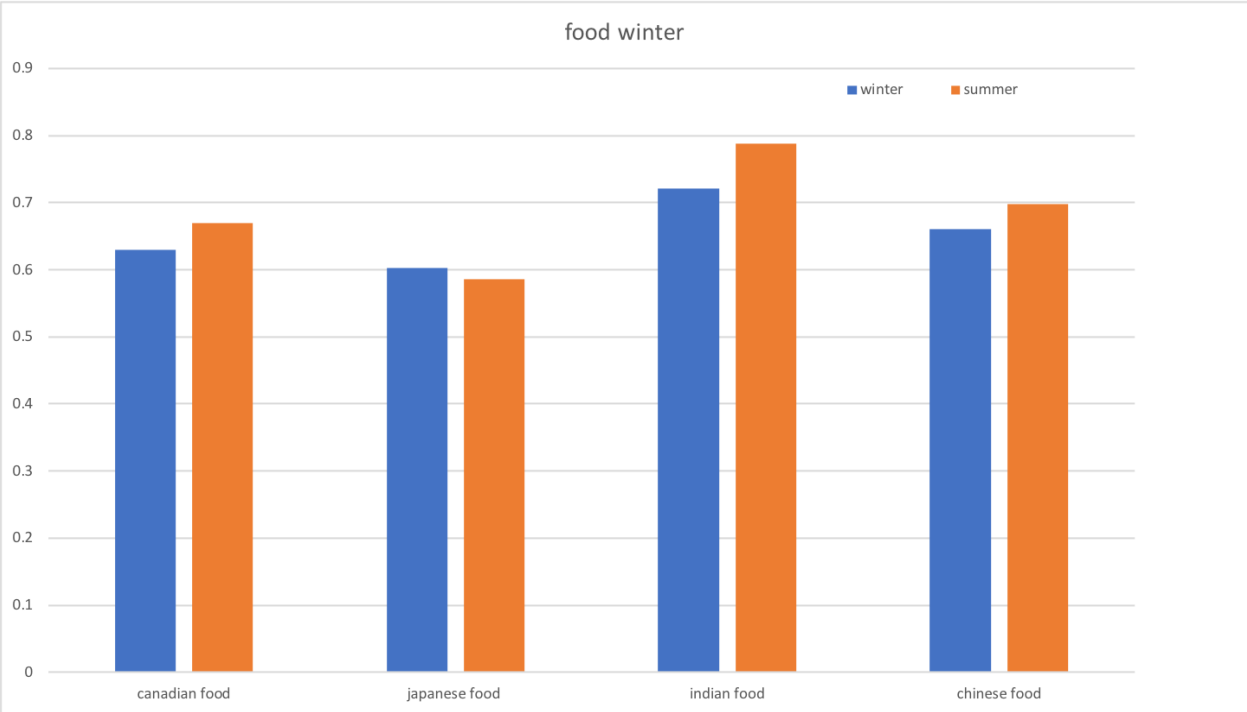
This two standard shows the choices of the people as well as the altitude of the people toward the specific study object. A large number of tweets indicates that more people would like to try the specific food or go to a viewpoint. The satisfaction score shows the whether people enjoy the food or not.

The result of visualizations are shown below:

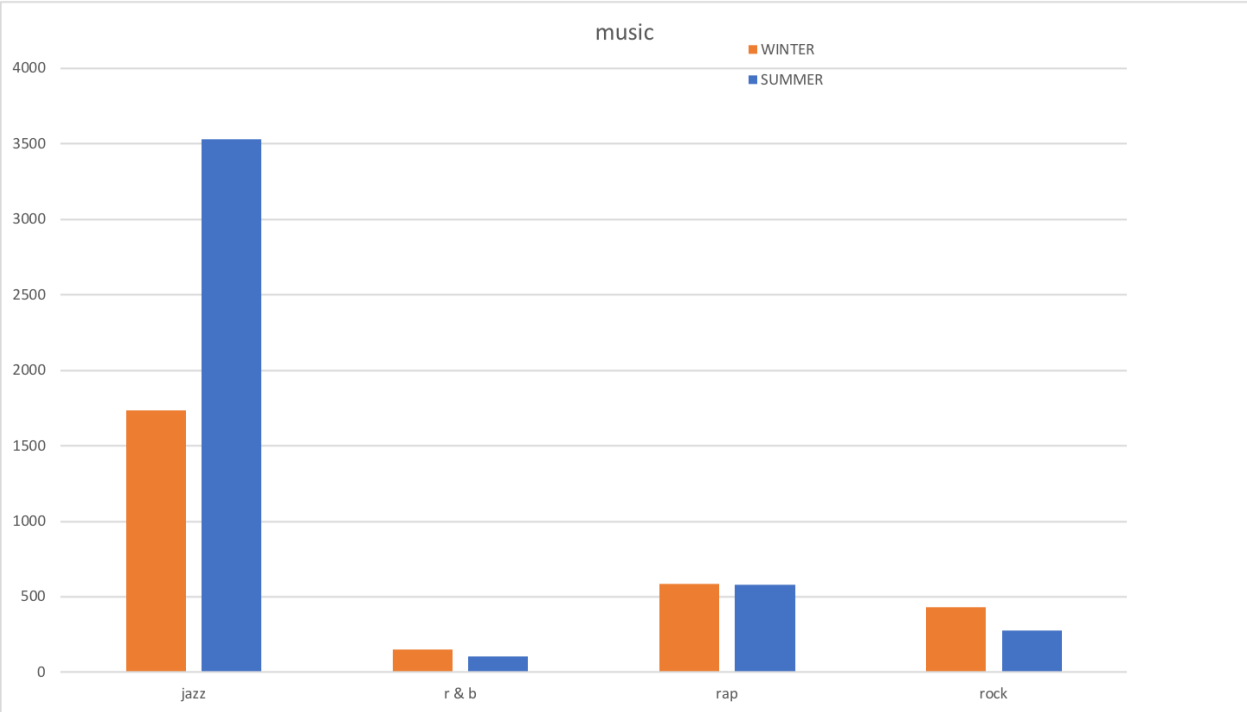


Result:

1. Most people in Halifax choose to have Canadian food in summer
2. Most people in Halifax choose to have Chinese food in winter
3. People choose to dine out more in the summer time

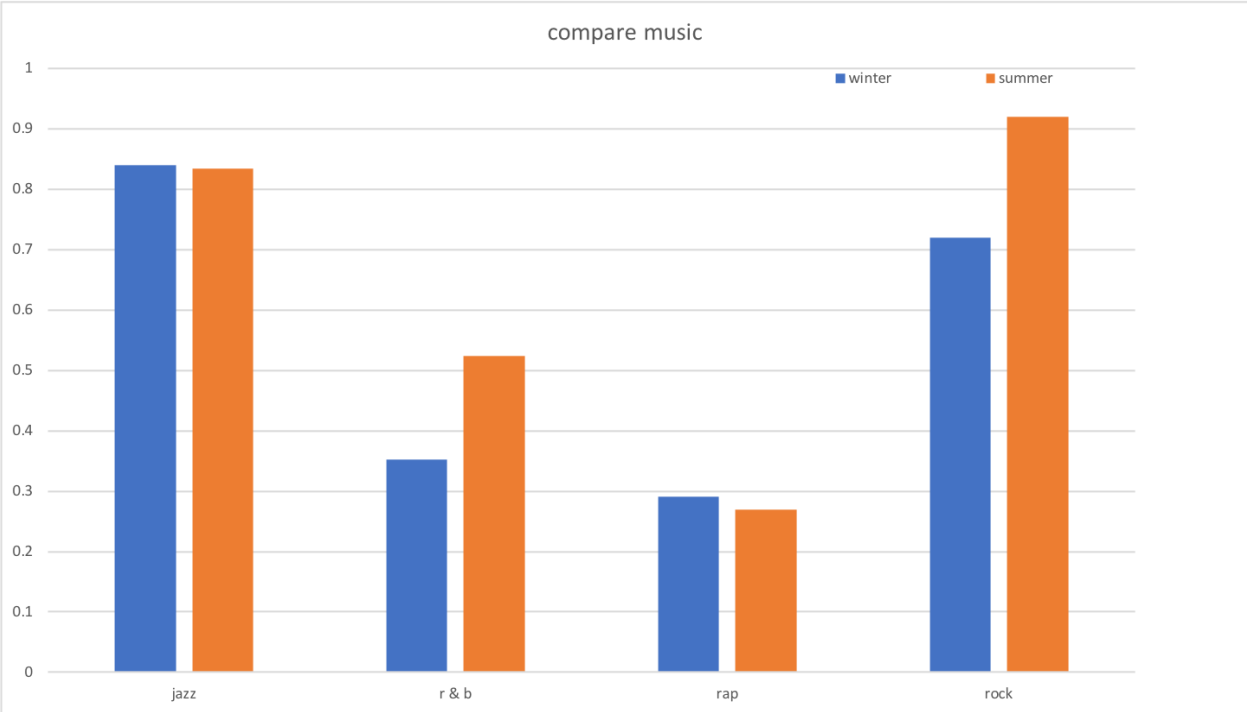


Result: The satisfaction score of Indian food is highest in both summer and winter in Halifax



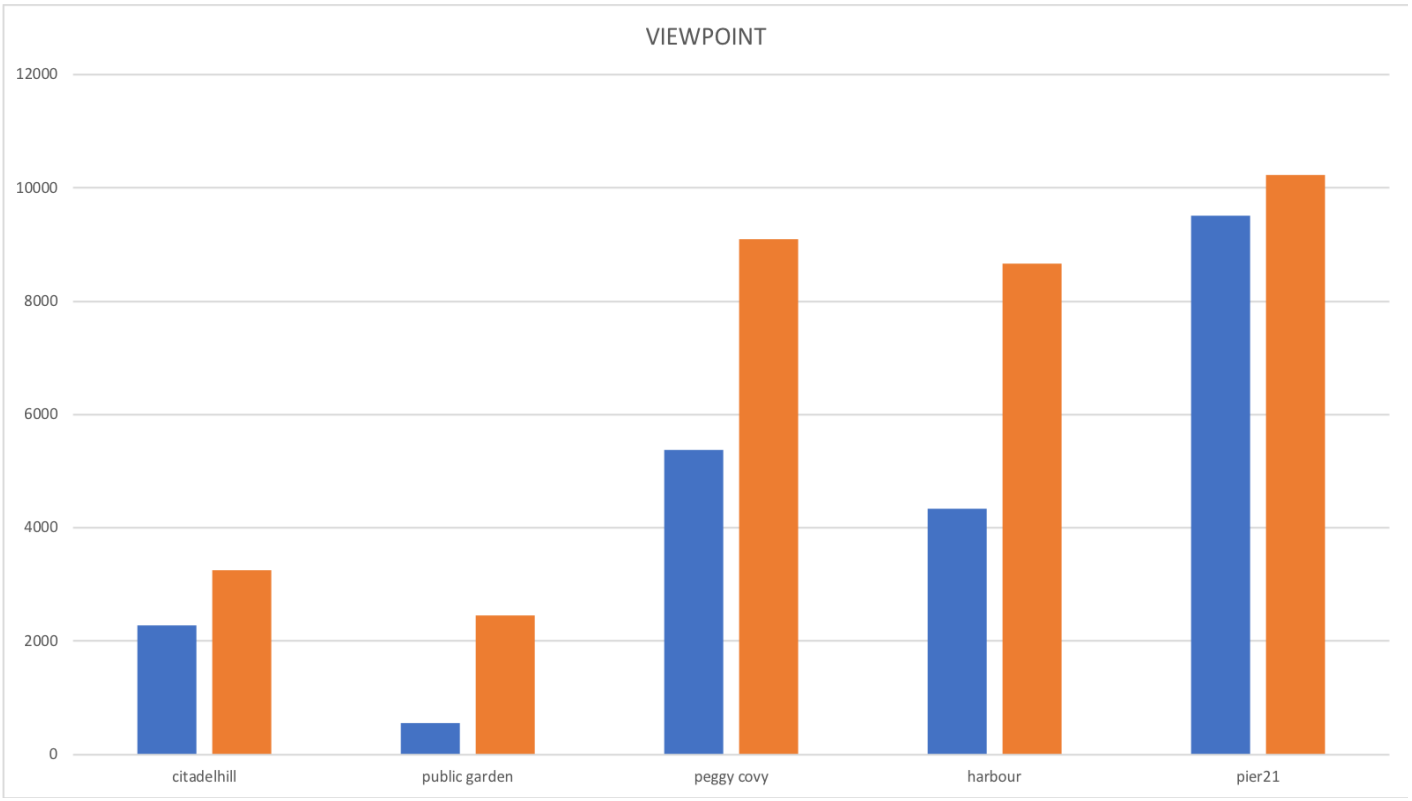
Result:

1. Most people in Halifax like to listen to Jazz music in both winter and summer
2. The reason for it might be the Jazz Festival held in Halifax annually.



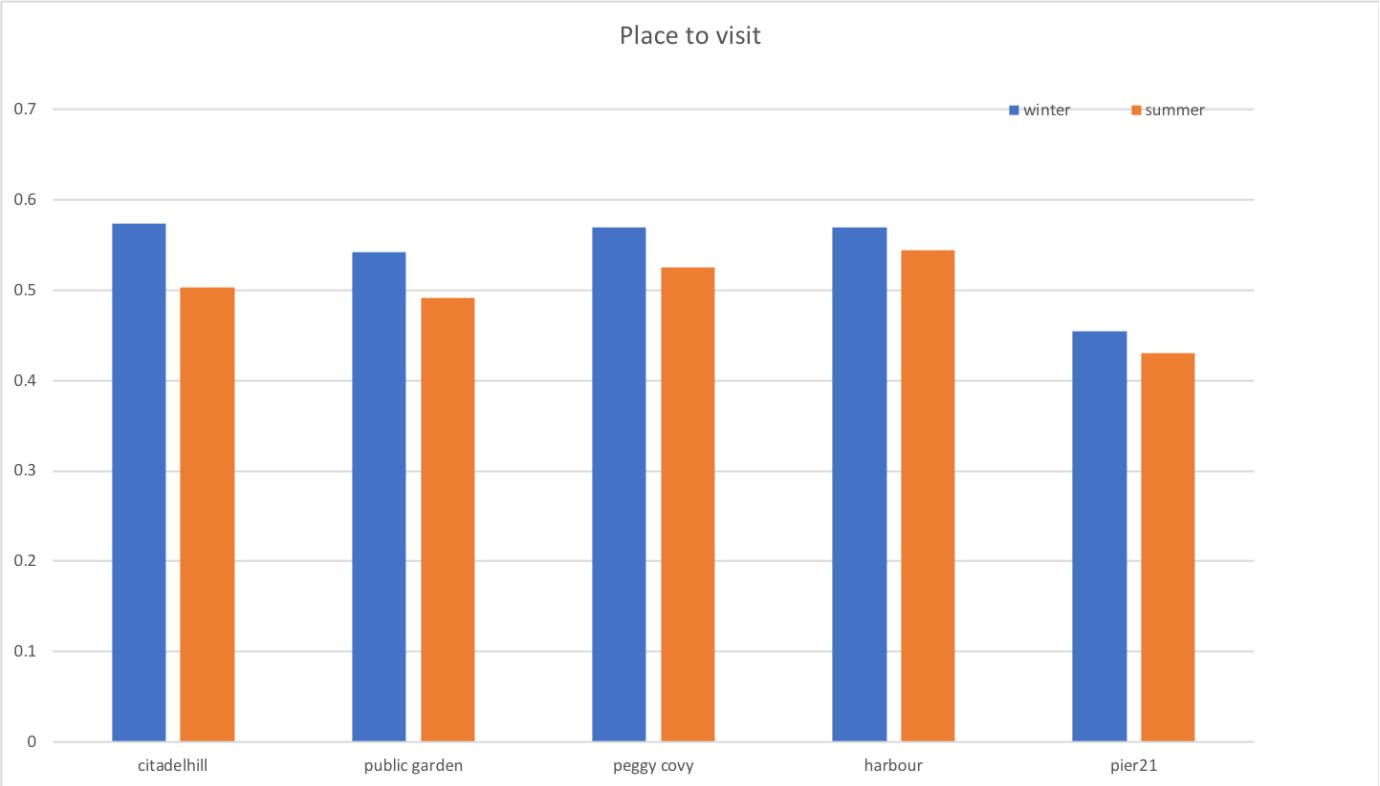
Result:

1. The satisfaction score of Jazz and Rock are both quite high.
2. Combining with the last graph, the Jazz music is the most popular music in Halifax.



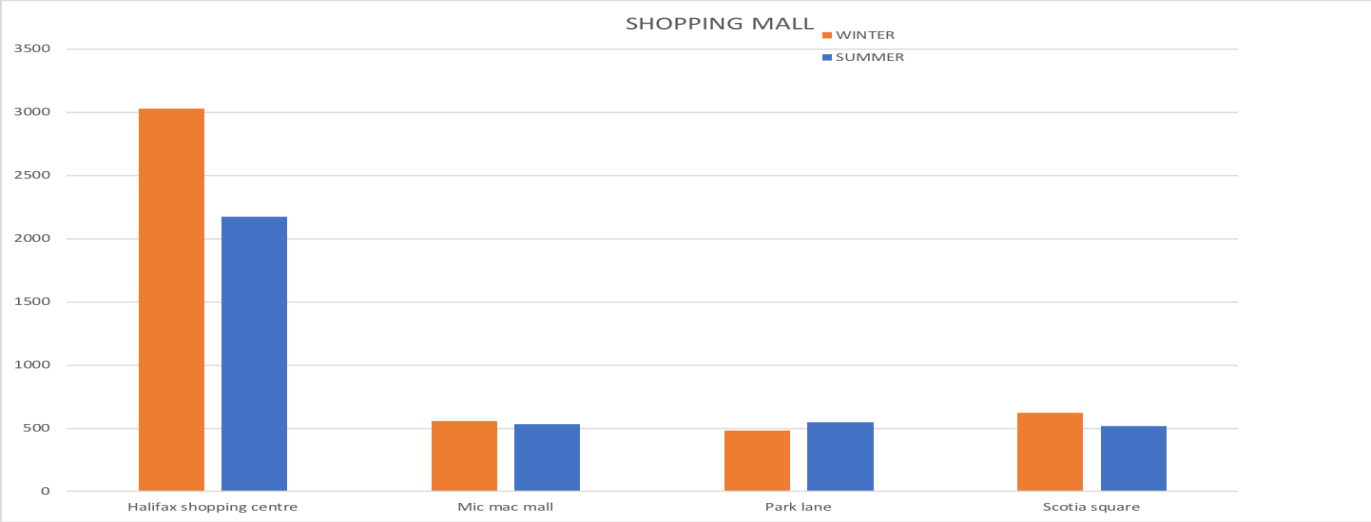
Result:

1. Pier 21 has most travelers around the year.
2. People like to go out more in the summertime.



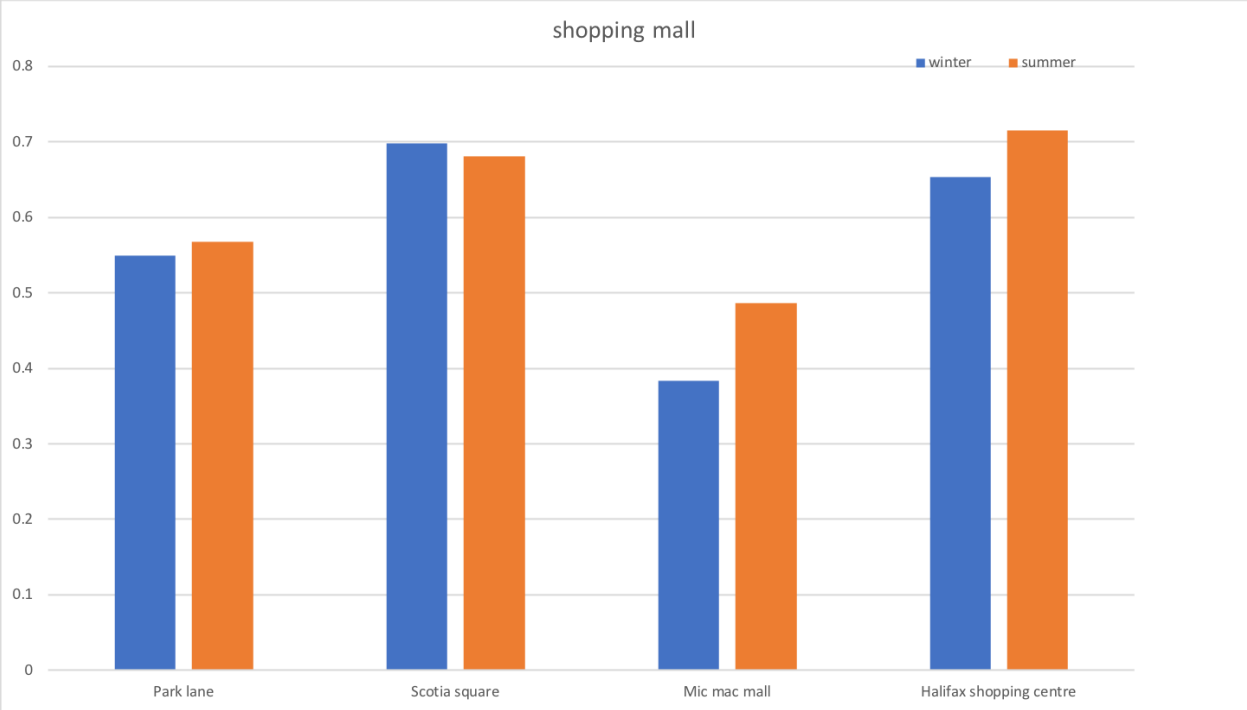
Result:

1. The satisfaction scores are higher in the winter for all the viewpoints. The reason for this might be Halifax is one of the warmest cities in Canada. Thus, people enjoy their stay in Halifax.
2. The Peggy's Cove and Harbour have the higher satisfaction scores. Yet the pier 21 has the most number of tweets.



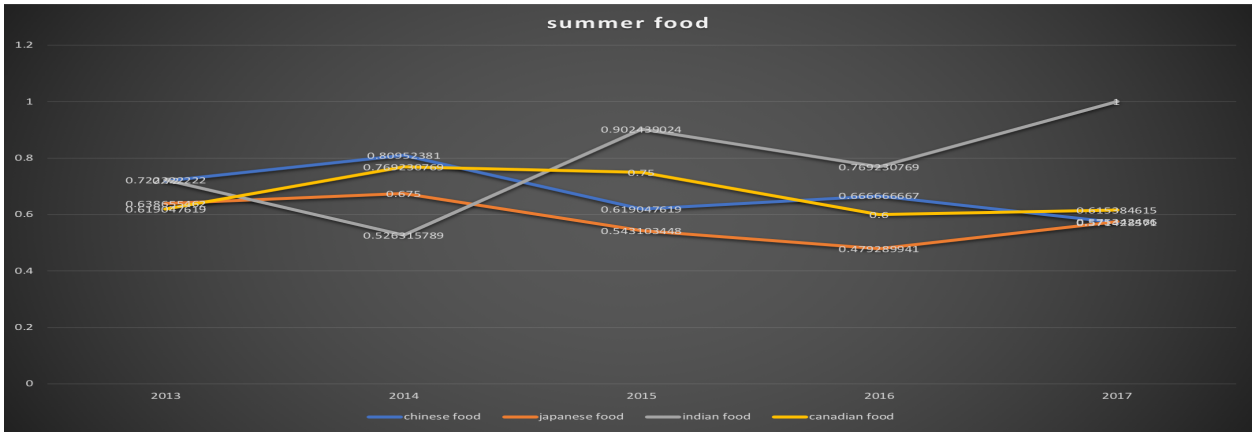
Result:

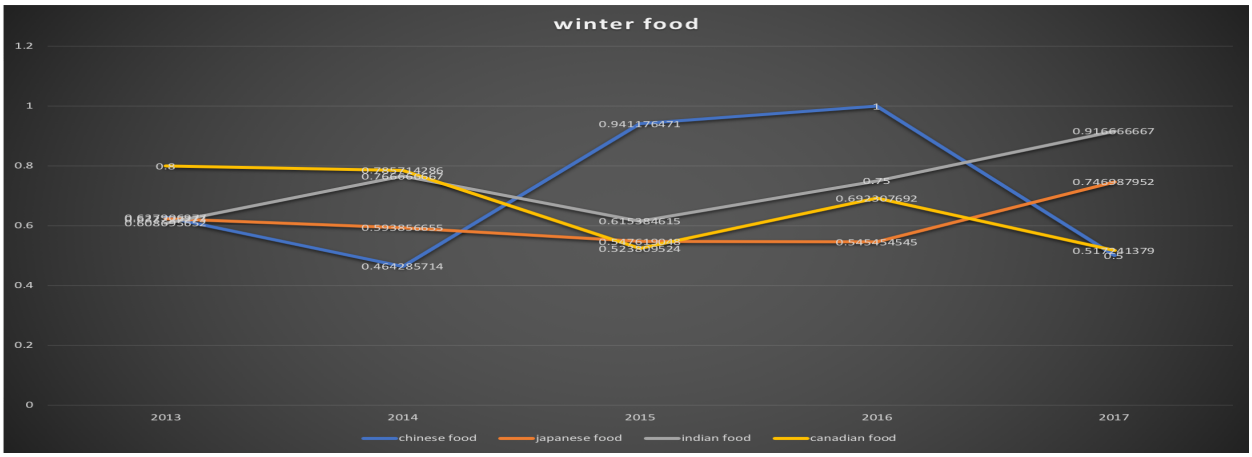
1. The Halifax shopping center is the most popular shopping mall in both winter and summer.
2. More people do shopping in the winter. The reason might be there are more vacation and holidays in the winter time, such as boxing day.



Result:

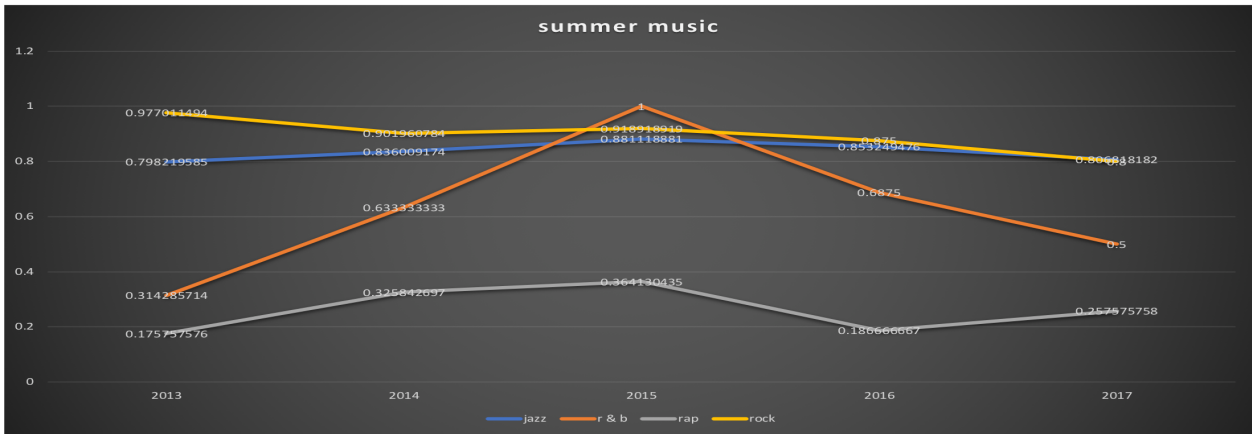
1. The satisfaction scores of Halifax shopping center and Scotia square are both quite well in winter and summer.
2. Combining with the last graph, the Halifax shopping center is the most popular shopping mall in Halifax .

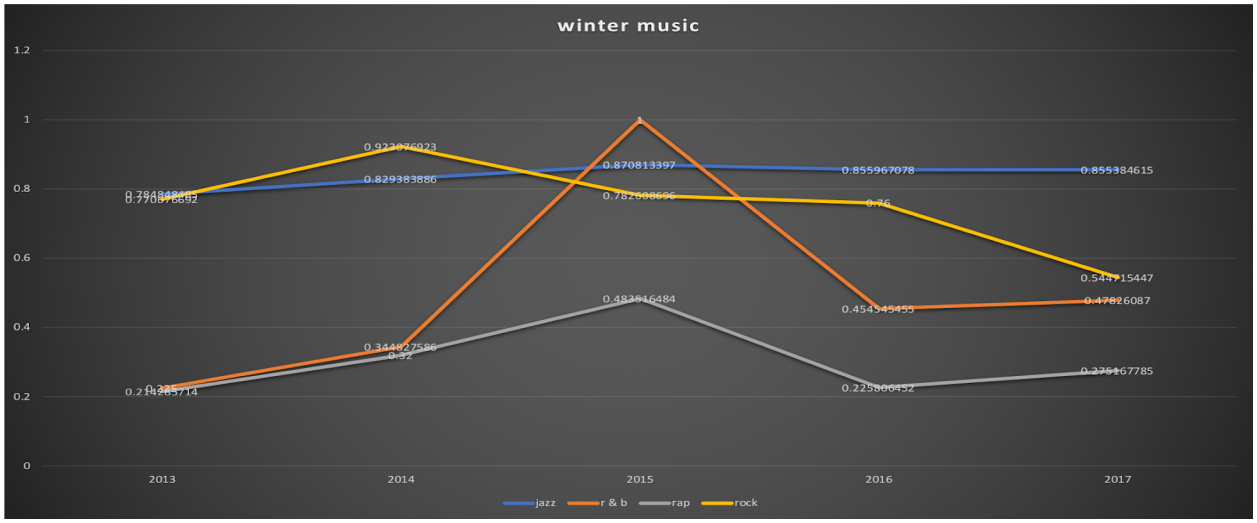




Result:

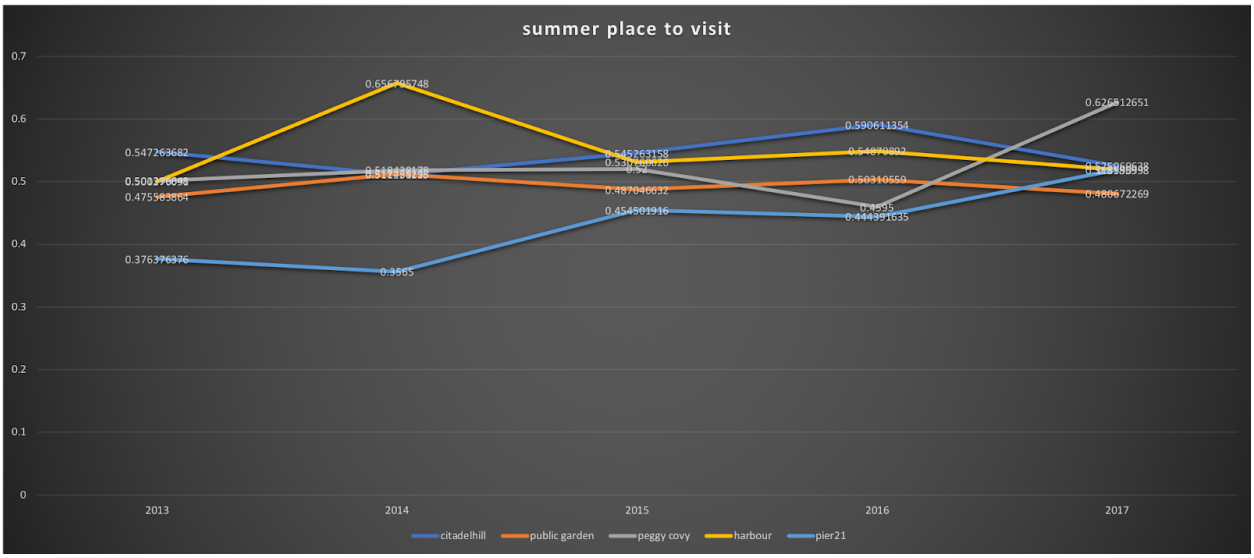
1. In winter and summer, the satisfaction score of Indian food is increased in past five years.
2. In winter and summer, the satisfaction score of Chinese food is decreased in past five years.
3. The reason might be there are more and more Indian students are in Halifax. And more and more local people enjoy the India food in past five years.

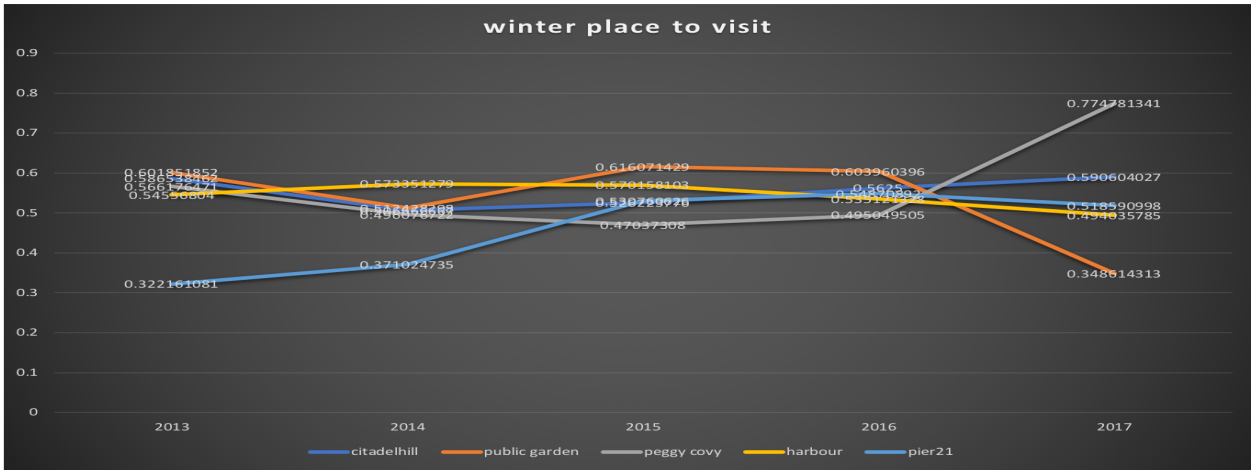




Result:

1. In winter and summer, the satisfaction scores of Jazz and Rock music are quite stable in past five years.
2. In 2015, the satisfaction score R&B surprisingly high. The reason might be there is a R&B show in Halifax in 2015





Result:

1. In winter and summer, the satisfaction score of Peggy's Cove is increased in past five years.
2. In winter and summer, the satisfaction score of pier 21 is decreased in past five years.
3. These results indicate that there might be more and more people would like to visit Peggy's Cove in the future.

Limitation of the work

During this project, we used sentiment analysis as the way to get the emotion of each tweet. The sentiment analysis sticks with the particular keywords to judge neutral, negative and position, which means it cannot judge the context through self-learning. Second, due to the single data platform, we are using, the trending of these comments about Halifax is easily biased. For example, when some Twitter users take the lead in Halifax to boycott some food or attractions, the negatives of the day's twitters may become very high, which is not conducive to our later analysis of the data.

Future work

For the future work, we are going to optimize our analysis approach. We need to find a way of using machine learn to instead of sentiment analysis. Once we use machine learning to process comments, the accuracy will improve. In the future, we also need to expand our data sources, such as retrieving data in Facebook.

According to the characteristic of machine learning, we are supposed to change the historical data to streaming data that the program can extract data at any time.

Work Breakdown

Week1

Data gathering: using the python program to gather the data from tweet

Week2

Data clean: using the python program to extra the tweet context

Week3

Data analysis: analysis of the result of the dataset

Week4

Visualization: presentation of the result data and find the final conclusion.

Critical Review

Overview of this project we did gathering data, data clean, sentiment analysis, data analysis, and virtualization. Based on the analysis of our data, the most popular food is Canadian food, the most popular music is Jazz, the most popular mall in Halifax is Halifax shopping center and the most popular viewpoint is Pier 21. But the satisfaction is not as same as the popular; the most satisfying food is Indian food, the most satisfying music is Rock, the most satisfied viewpoint are Peggy's Cove and Harbour, and the most satisfied mall is Halifax shopping center. We can see that only the Halifax shopping center have both. In other options, no one has both high satisfaction and popular. If you are a tourist, I would recommend that you pay attention to the popular places, because the popular places are generally the center of the entire area. If you are a start-up person, I suggest that you pay attention to satisfaction, which means the

options for high satisfaction in the development of the next few years have a strong development momentum, and there will be more opportunities than the already popular places.

Role Based Distribution of Work

Yiwei zhang:

1. Data gathering: Gather the data for the music and viewpoint parts
2. Data clean: Do the data clean on the music and viewpoint parts
3. Data analysis: Do the sentiment analysis based on the music and viewpoint datasets
4. Visualization: Visualized the music and viewpoint datasets.

Haoyu Sun:

1. Data gathering: Gather the data for the food and shopping mall parts
2. Data clean: Do the data clean on the food and shopping mall parts
3. Data analysis: Do the sentiment analysis based on the food and shopping mall datasets
4. Visualization: Visualized the food and shopping mall datasets.

References:

[1] A. Cernian, V. Sgarciu, and B. Martin, “Sentiment analysis from product reviews using SentiWordNet as lexical resource,” 2015 7th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), 2015.

[2] Anelachan, “anelachan/sentimentanalysis,” GitHub. [Online]. Available: https://github.com/anelachan/sentimentanalysis/blob/master/ReadMe.md. [Accessed: 10-Jul-2018].

[3] “Tableau Software,” Wikipedia, 23-Jul-2018. [Online]. Available: https://en.wikipedia.org/wiki/Tableau_Software. [Accessed: 29-Jul-2018].

[4] Jefferson-Henrique, “Jefferson-Henrique/GetOldTweets-python,” GitHub, 23-May-2018. [Online]. Available: https://github.com/Jefferson-Henrique/GetOldTweets-python. [Accessed: 04-Jul-2018].

URL: https://github.com/yw349762/CSCI5408A3

Project directory