Final Reprot

Discovery of the recent music trends based on Sentiment Analysis in Twitter or Spotify

Author: Wenjie Chen

Supervisor: Dr. Daniela Tsaneva

Abstract

This report is about using Machine Learning to conduct analysis on data from the music industry. This project aimed to predict recent music trends based on Twitter by Sentiment Analysis. By applying computer science technology, the new insights can be provided and it can benefit from the data analysis.

In this project, some AI techniques, such as machine learning, databases and data visualization have been applied. There are three main steps to build this project. These are collecting data from Twitter and Spotify, analyzing data by Sentiment Analysis and showing the result by Data Visualization.

This project can analyze the sentiment of a sentence. The sentiment can be positive or negative. However, it could not predict the recent music trend accurately, this is because there are many elements that can affect music trends. By data visualization, the project shows some keys to affect music trends and provides more information for the music industry.

Although this project could not draw a conclusion to predict future music trends, perhaps only indirectly predict those trends, it can provide some insight in the field of music. What this research brings is to show the influence of multiple dimensions on the popularity of a singer or a song.

Acknowledgements

I would like to thank my supervisor Daniela Tsaneva for her help and encouragement throughout the project. We have much meeting and she gives me some suggestion to help me and improve my confident.

Contents

Table of Contents

Table of figure

Introduction and background

Music occupies an important position in people's daily life and there are many people who take some time to listen to music. Music could even affect all aspect of people's lives. This is because according to (Abraham et al. 2015)music listening is a means of stress reduction in daily life. For example, a lot of work puts a lot of pressure on office workers but Sandstrom and Russo (2010) found music high in valence and low in arousal positively affected recovery of heart rate and skin conductance levels after a stressor. Therefore, people will make some personal comments on the degree of music love on social media. The degree of love for music will also be shown on music chart, and music websites will rank these songs in a variety of ways. Through the above-mentioned comments on social media and music charts, a prediction of music trends could be made by Sentiment Analysis.

This project is to predict recent music trends by Sentiment Analysis in Twitter or Spotify. Sentiment Analysis (Feldman 2013) (or Opinion Mining) is defined as the task of finding the opinion of authors about specific entities. The decision-making process of people is affected by the opinions formed by thought leaders and ordinary people. Therefore, in the field of music, through the method of sentiment analysis, we could make some prediction about people's preference for music. For example, someone said this song is really nice on social media. This could get a positive result about this song. Someone said this song is bad. This could get a negative result about this song. Therefore, the purpose of sentiment analysis is to analyze a text to differentiate between a positive or negative comments. As a matter of fact, sentiment analysis could analyze multiple different types of problems, and they are document-level sentiment analysis, sentence-level sentiment analysis, aspect-level sentiment analysis, comparative sentiment analysis and sentiment lexicon acquisition. Moreover, there are many methods of sentiment analysis, but they are roughly divided into two methods. The first one is rule-based sentiment analysis and it uses a dictionary of words labelled by sentiment to determine the sentiment of a sentence. Sentiment scores typically need to be combined with additional rules to mitigate sentence containing negation, sarcasm, or dependent clauses. The second one is machine learning (ML) based sentiment analysis. ML model could be trained to recognize the sentiment based on the words. This approach depends on largely on the type of algorithm and the quality of the training data used.

I decided to use ML model to build the project. The following are the specific steps to use machine learning to do sentiment analysis.

1. Data can be collected by web crawler on social media. People prefer to rate songs and singers on Twitter or forums, and most comments about singers and songs could be collected by Twitter API.
2. It is necessary to preprocess the data. This is because there are many words in English that cannot affect the results of sentiment analysis, and these words should be deleted.
3. Text vectorization. As a matter of fact, the purpose of text vectorization is to turn it into numbers, because computers could not understand these characters string directly.
4. Use the Naïve Bayes algorithm to get the results of sentiment analysis. There are many kinds of sentiment classification algorithm, such as Supported Vector Machine (SVM), Naïve Bayes and logistic regression. Different algorithms have different accuracy of word segmentation results. Here I choose to use the Naïve Bayes algorithm.
5. Naïve Bayes is a good algorithm and it is easy to understand. It has a good accurate of classification.

   Naive Bayes is an algorithm of machine learning. Machine learning could be divided into supervised learning and unsupervised learning. Naive Bayes is a classification algorithm that divides the analysis results into positive and negative in sentiment analysis, so it belongs to supervised learning. Predicting the categories of the test data set is simple and fast, and it performs well in multi-category predictions.

   There are many applications of sentiment analysis in real life. For example, sentiment analysis is helpful in language learning. In the learning of a second language, emotion vocabulary has not received considerable attention. M. Chen, W. Chen and L. Ku (2018) developed a system, which is a context-aware emotion synonym suggestion system for educational purpose to help people to learn second language. The results indicate that the participants achieved substantial progress on emotion word use with the help of the proposed system. Similarly, this project, predicting music trends, is to help the music industry, musicians, etc. better understand user preferences, and help them invest more flexibly. Therefore, sentiment analysis can help analyze and get useful feedback from users in many industries, and facilitate industries to optimize their products.

   Machine learning is an evolving branch of computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment (El and Martin 2015). Sentiment Analysis (Feldman 2013) (or Opinion Mining) is defined as the task of finding the opinion of authors about specific entities. The decision-making process of people is affected by the opinions formed by thought leaders and ordinary people.

Project goals

1. Will positive comments keep the song in a high ranking?

Most people want to ask this question, and it is also a question that I personally want to answer the most. If there are some positive comments on Twitter, the popularity of a song or some songs of an artist will increase significantly. Between positive comments and the number of music played, there are many other variables that could affect the result, such as time, the type of the song. Therefore, between the two variables, there are some unpredictable elements. For example, a song is known to the public as the theme song of a movie, so many people may hear it on the music platform. Just as the world is unpredictable, my research could not establish a perfect model. Because of the small sample size, the results may be very unstable. Maybe there is a connection between these two variables, or there is no connection between these two variables at all. For instance, some singers may have very positive reviews, but he or she rarely releases digital albums or singles, and instead he or she releases physical albums or singles. Therefore, it is difficult to determine the relationship between these two variables. Specific to the research method, I will collect some positive comments about the singer from 2019 to 2020, and I will collect the number of followers of the singer from 2019 to 2021 and the top 200 songs on music platform, and then I will analyze the results by the number of followers and the number of play before and after.

2. Identify if negative comments will affect the number of times a song has been played

Compared with the first one, this question only seems to have changed one precondition, but in fact there is a big gap between the two variables. Although negative comments could have a negative impact on the song or the singer, it could also help in the reverse direction and increase the exposure rate. For example, there are many reviews who say that a song is terrible, and not just one person is saying it. Therefore, this will arouse some people's curiosity, and they want to hear how bad this song is. Therefore, negative comments may bring unexpected results to the song or the singer. For example, a lot of music that surpassed the cognitive level of the times attracted many people's incomprehension when it was just released, and they thought it was a very strange song. However, after the improvement of people's overall cognition, many people can understand what the author wants to express by listening to this song, and people's attitude towards this song has undergone a great change. I will collect the negative comments about the singer from 2019 to 2020, and will collect the number of followers of the singer from

2019 to 2021 and the top 200 songs played on music platform. I will make a before-and-after comparison to analyze the impact of negative comments on the singer.

3. Does the type of song affect the ranking of music?

Because of different generation and different cultural backgrounds, most people's music tastes are different. Therefore, the music style would have a certain degree of influence on the number of play and trend of music. For example, if some people like playing basketball, they may prefer listening to hip hop music. In addition, as a matter of fact, music is divided into positive and negative lyrics. There is music that expresses sadness and music that brings happiness to people. Therefore, the mood of the listener is also part of deciding what music to listen to. However, there are too many classifications of music and it is difficult to define, so I will take the classification method on the music platform Spotify to classify. I will calculate the total playback volume of each music genre from 2018 to 2021 and make a table for intuitive comparison and analyze whether the music type really affects the number of playing, that is, a certain music genre is actually more popular in the eyes of the public. Analyze music from this perspective will show a general understanding of music trends, which could help the music industry make capital investment.

4. Identify if the number of followers of a singer affect the number of playing of the singer's song

The number of followers of singer is an important element that determines the number of playing of singer's song. When the song is released, the singer's followers may support the singer he or she likes and listen to the music they release. However, while paying attention to singers, it will actually bring some negative effects. For example, for follower who follow singer, the negative news about the singer will be noticed by them for the first time. This may affect the fans' following the singer, which may lead to a decrease in the number of songs played. However, there are some singers who seldom use social media, or rarely appear in the public eye, and have less interaction with their followers, which may lead to the phenomenon that these people have low followers but high played volume. Therefore, I will collect the number of followers and songs played by some singers to show how they change over time.

Implementation

In this project, there are five steps, namely collecting data from Spotify, collecting data from Twitter, training a machine learning model, using the machine learning model and visualizing the results of the project with data.

1. COLLECT DATA FROM SPOTIFY

The premise of making a prediction is to obtain enough data to analyze the past data and predict the future. Therefore, the first step was to obtain music data from Spotify. Spotify is a Swedish audio streaming and media services providers founded on 23 April 2006 by Daniel Ek and Martin Lorentzon. It is one of the world's largest music streaming service providers, with over 365 million monthly active users. Spotify offers digital copyright restricted recorded music and podcasts, including more than 70 million songs, from record labels and media companies. Therefore, it is a great platform for collecting data. However, the official Spotify API does not provide information about artists for individual tracks, so after searching for a period of time, I found another website Spotify Chart that could provide data. Spotify Chart is a snapshot of what music followers are loving around the world and it is easy to get chart data from Spotify. It provides the top 200 daily singles played worldwide, from January 1, 2017 to present.

Spotify Chart has many advantages, simple arrangement, and easy access to data. Finding the URL where you want to get the data, it can be obtained easily through a web crawler. A web crawler is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion. Web crawling is an important method for collecting data on, and keeping up with, the rapidly expanding Internet. (Dhenakaran and Sambanthan [no date]). In this project, I decided to use Python web crawler. The unique advantage of Python is the key to writing web crawler. Firstly, it is cross-platform and has good support for Linux, Windows and macOS. Secondly, using scientific calculations and numerical fitting are convenient because of numpy and scipy which are libraries in Python. Thirdly, it has very convenient visualization libraries which are Matplotlib and Mayavi. Fourthly, in the presence of complex networks, Networkx and scrapy crawlers could still provide good help. Therefore, compared with other languages, python crawlers have very good advantages in these situations.

There are three main methods to crawl data with Python and the methods are Regular Expressions, Beautiful Soup and Lxml. The first is Regular Expressions and there are five main steps to collect data. Firstly, an html web page is deployed on the Tomcat server. Secondly, the URL is used to establish a connection with the web page. Thirdly, the input stream will be got,

and the input stream is used to read the content of the web page. Fourthly, Regular Expressions will be wrote and it is based on the data that you want. Finally, the data will be stored in the database. Beautiful soup could also get the content of the web page. It uses the structure and attributes of the webpage to parse the webpage. To use it, it is not necessary to write very complex regular expressions, and only a few simple sentences could complete the extraction of elements in the web page. The last method is Lxml. It is a Python parsing library and it supports HTML and XML parsing, and it is very efficient. There are three main steps. Firstly, the XML file would be found and then parsed. Secondly, after parsing, a certain label will be found or located. Finally, the attributes and text content will be found. Crawlers mainly have the above three methods. I decided to use the second one which is Beautiful Soup. This is because it is relatively simple to use, and the amount of data I collect is not large and there is no need to pursue efficiency.

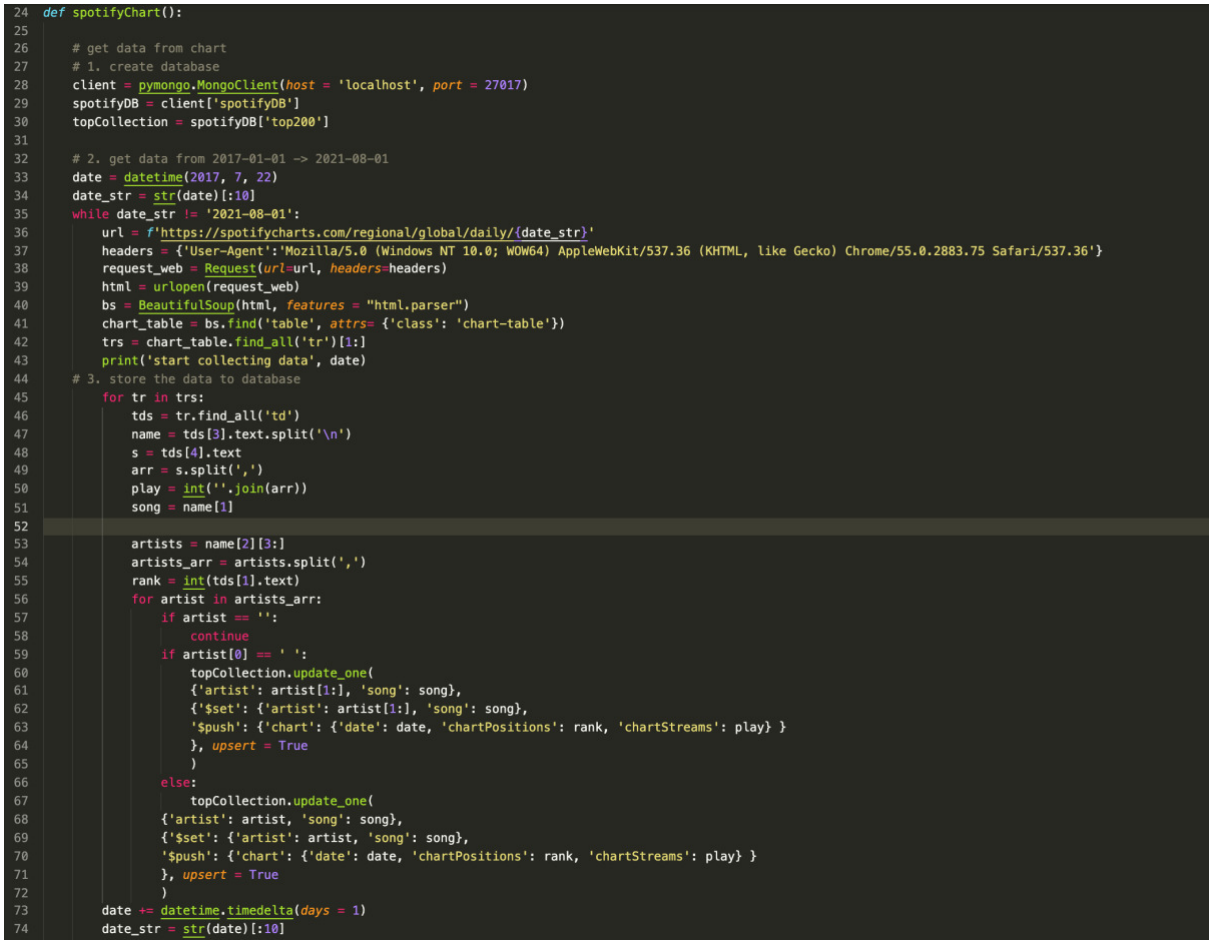
After collecting the data, how to store the data is a very troublesome problem. There are two types of databases and they are Structured Query Language (SQL) and NoSQL, and they have their own advantages.

SQL was initially developed at IBM by Donald D. Chamberlin and Raymond F. Boyce after learning about the relational model from Edgar F. Codd in the early 1970s (Chamberlin and Boyce 1976). In the following decades, the concept of relational has been fully developed and has gradually become the mainstream model of database structure. A relational database is a two-dimensional table model, and a relational database is a data organization composed of two-dimensional tables and the connections between them. There are three main advantages of relational databases. Firstly, the relational database is easy to understand, and the two-dimensional table is very close to the real world, so it is easy to understand by the learner and beginner. Secondly, the relational database is convenient to use and the SQL sentences make it very convenient to operate. Thirdly, the relational database is easy to maintain, and the attributes of SQL reduce the probability of database redundancy and data inconsistency. However, SQL has own shortcomings. First of all, the efficiency of reading and writing massive amounts of data will be very poor. For example, some websites have high concurrency, with many requests per second. For relational databases, it is difficult for the input and output of the hard disk to handle these requests. Finally, it is difficult for the database to expand horizontally. For example, in a web application, the number of users and visits of the system has increased gradually over time, and the database cannot expand the performance and load capacity by adding more components and server nodes like a web server.

The biggest advantage of relational databases is the consistency of transactions. This feature allows relational databases to be used in many systems with relatively high consistency, such as the banking system. In the application of web pages, such consistency requirements are not strict, allowing a certain time to respond, so this feature is not important for the systems. On the contrary, it takes too many resources to maintain consistency to read and write data. However, some applications have high requirements for concurrent reading and writing, such as Facebook, so the relational database cannot have a good solution, so a new data structure would be used to store data to replace the relational database. Therefore, no-relational databases were created to solve this problem.

The term NoSQL was first used in 1998 for a relational database that omitted the use of SQL. The term was picked up again in 2009 and used for conferences of advocates of non-relational databases such as Last. Fm developer Jon Oskarsson, who organized the NoSQL meetup in San Francisco (Pagán et al. 2015). NoSQL uses a different concept to store key and values, but the structure is unstable. Each table may have different fields, so this is not limited to a fixed structure. Non-relational databases have three advantages. Firstly, non-relational databases do not need to be parsed, and the reading and writing performance are relatively high. Secondly, non-relational databases are based on key and values, so it is easy to expand. Thirdly, non-relational databases could store different types of data, such as documents, pictures, videos, etc. However, non-relational databases also have some disadvantages. Firstly, there is no relationship between the data, which is more difficult for beginner. Secondly, non-relational databases could not guarantee the security and integrity of data.

In this project, I think non-relational database is more suitable for storing the data that I want, so I used MongoDB to store the data. MongoDB is a NoSQL database. Through the above techniques I used, I obtained two databases table, one is SpotifyAPI and the other one is top200. The first table, SpotifyAPI, stores the artist's name, music genre, and the data that is the highest number of followers from January 1, 2017 to August 1, 2021 and the songs released during this period. First of all, the purpose of storing music types is to compare the differences in the number of playing songs between different types of songs. This difference may affect the next music trend, because most people may prefer a certain type of music. Similarly, storing the number of followers of a singer is also for this reason. The trend of music could be predicted by comparing the number of followers of different types of music.



```
def spotifyChart():

    # get data from chart
    # 1. create database
    client = pymongo.MongoClient(host = 'localhost', port = 27017)
    spotifyDB = client['spotifyDB']
    topCollection = spotifyDB['top200']

    # 2. get data from 2017-01-01 -> 2021-08-01
    date = datetime(2017, 7, 22)
    date_str = str(date)[:10]
    while date_str != '2021-08-01':
        url = f'https://spotifycharts.com/regional/global/daily/{date_str}'
        headers = {'User-Agent':'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/55.0.2883.75 Safari/537.36'}
        request_web = Request(url=url, headers=headers)
        html = urlopen(request_web)
        bs = BeautifulSoup(html, features = "html.parser")
        chart_table = bs.find('table', attrs= {'class': 'chart-table'})
        trs = chart_table.find_all('tr')[1:]
        print('start collecting data', date)
    # 3. store the data to database
        for tr in trs:
            tds = tr.find_all('td')
            name = tds[3].text.split('\n')
            s = tds[4].text
            arr = s.split(',')
            play = int(''.join(arr))
            song = name[1]

            artists = name[2][3:]
            artists_arr = artists.split(',')
            rank = int(tds[1].text)
            for artist in artists_arr:
                if artist == '':
                    continue
                if artist[0] == ' ':
                    topCollection.update_one(
                    {'artist': artist[1:], 'song': song},
                    {'$set': {'artist': artist[1:], 'song': song},
                    '$push': {'chart': {'date': date, 'chartPositions': rank, 'chartStreams': play} }
                    }, upsert = True
                    )
                else:
                    topCollection.update_one(
                {'artist': artist, 'song': song},
                {'$set': {'artist': artist, 'song': song},
                '$push': {'chart': {'date': date, 'chartPositions': rank, 'chartStreams': play} }
                }, upsert = True
                )
        date += datetime.timedelta(days = 1)
        date_str = str(date)[:10]
```

Figure 1: SpotifyAPI code

The second table is top200. In this table, I collected the name of the singer, the name of song that the it appeared on the Spotify Chart website, and the time it appeared each time, the number of the songs played on that day, and the ranking of the song on that day. These are key information for analyzing changes in music trends.

```python
# get detail of artist
def artistsInformation():
    #create a new collection
    collection = getMongoDB('spotifyAPI')

    #get all artist
    allArtist = getAllArtists()

    spotify_access_header = header()

    for index, artist in enumerate(allArtist):
        if '&' in artist:
            artistName = artist.replace('&', '')
        else:
            artistName = artist

        artist_obj = requests.get(f'https://api.spotify.com/v1/search?q={artistName}&type=artist', headers = spotify_access_header).json()
        bestIdx = 0

        for indx, result in enumerate(artist_obj['artists']['items']):
            if artist_obj['artists']['items'] != []:
                if result["followers"]["total"] > artist_obj["artists"]["items"][bestIdx]["followers"]["total"] and result["name"] == artist:
                    bestIdx = indx
                    print("Idx changed: ", indx)
            else:
                continue
        if artist_obj["artists"]["items"] != []:
            print(bestIdx)
            bestObject = artist_obj["artists"]["items"][bestIdx]

            genres = bestObject['genres']
            spotifyId = bestObject['id']
            followers = bestObject['followers']['total']
            artistId = bestObject['id']
            document = {"name": artist, "genres": genres, "spotifyId": artistId, "followers": followers}
            print(index, ": Going into database: ", document)
            update_result = collection.insert_one(document)
        else:
            continue
```

Figure 2: top200 code

2. COLLECTING DATA FROM TWITTER

Nowadays, the Internet is particularly developed, and people could send messages on social media anytime and anywhere, such as Facebook and Twitter. Twitter is an American microblogging and social networking service on which users post and interact with messages known as “tweets”. Users could interact with other users all over the world by sending tweets, as well as like, comment and other functions. By 2012, more than 100 million users posted 340 million tweets every day. Therefore, Twitter is a good platform to get people’s views on music. People may comment on the song, or comment on the singer and all of these factors may have an impact on music trends. The Twitter company provides an API, which could get users’ tweets by keywords. However, Twitter’s official API has restriction for individual users. After the collected data reaches a certain amount, Twitter will no longer provide API services for free, so I plan to change the way to collect tweets. After searching for a while, I found twint. Twint is an advanced Twitter scraping tool written in Python that allows for scraping tweets from Twitter profiles without using Twitter’s API. Therefore, I collected the tweets with the singer’s name as the keyword and the time when the tweet was sent, and stored these data in the database. The name of the table I saved is twintDB.

```python
def twintGet():
    collection = getMongoDB('twintDB')
    all_artist_100 = getAllArtists()
    all_artist_100.remove('Trío Los Josefinos')

    c = twint.Config()
    c.Lang = 'en'
    c.Store_csv = True
    c.Count = True
    c.Hide_output = True
    c.Email = False
    c.Phone = False
    c.Limit = 10
    # c.Verify_ssl=False
    c.Output = 'twint.csv'
    startDate = '2019-1-1'
    endDate =  '2020-1-1'

    nltk.download('stopwords')
    stopwords_list = set(stopwords.words("english"))
    all_artist = all_artist_100[103:105]

    for artist in all_artist:
        print('start searching tweets ', artist)
        cur_start_date = '2019-1-1'
        cur_end_date = '2019-2-14'
        c.Search = artist
        c.Output = f'twint_{artist}.csv'
        while datetime.datetime.strptime(cur_start_date, '%Y-%m-%d') < datetime.datetime.strptime(endDate, '%Y-%m-%d'):
            c.Since = cur_start_date
            c.Until = cur_end_date
            print('Searching for artist: ', artist, ' between the dates: ', cur_start_date, ' and ', cur_end_date)
            twint.run.Search(c)
            cur_start_date, cur_end_date = incrementDates(cur_start_date, cur_end_date)

        with open(f'twint_{artist}.csv', 'r', encoding = 'utf8') as file:
            csv_reader = list(csv.reader(file, delimiter = ','))

        for line in csv_reader[1:]:
            tweet = line[10].lower()
            subject = artist
            language = 'en'
            time = line[3]
            clean_tweet = re.sub('[^a-z\s]', '', tweet)
            tweet_no_stopwords = " ".join([i for i in clean_tweet.split() if i not in stopwords_list])
            tweet_no_links = " ".join([i for i in tweet_no_stopwords.split() if "http" not in i])
            if line[11] == language:
                collection.insert_one(
                    {
                        'subject': subject,
                        'text': tweet_no_links,
                        'Date': datetime.datetime.strptime(time, '%Y-%m-%d')
                    }
                )
        os.remove(f'twint_{artist}.csv')
```

Figure 3: twint code

3. Building a machine learning model

In my opinion, machine learning is a scientific approach to make machines think and learn like humans, so as to learn how to complete the corresponding work, handle a large amount of work in a short time, and improve work efficiency. For example, that can be applied in places where identity information needs to be verified, such as airport. The face recognition function under artificial intelligence could help airport managers identify and verify the identity information of passengers, thereby replacing unnecessary waste of manpower and allowing people to do valuable things. Therefore, machine learning is particularly useful in life, and it can assist people to complete their work. Similarly, artificial intelligence can also help people by providing some new information and perspectives, and make better plans and predictions for

work or the future. There are many fields of machine learning, such as Natural language processing, computer vision, data mining, search engines, etc.

Machine learning is divided into supervised learning and unsupervised learning. Supervised learning is called classification or inductive learning. This type of learning is analogous to human learning from past experiences to gain new knowledge in order to improve our ability to perform real-world tasks (Liu B. 2011). Unsupervised learning allows the system to identity patterns within data sets on its own. Therefore, different types deal with different problems. Under each type, there are different models to help deal with different problems. In fact, there is another type called reinforcement learning. Reinforcement learning studies how machine take a series of behaviors in the environment to obtain the greatest return.

In the content of machine learning introduced above, machine learning is mainly divided into two types, supervised and unsupervised. Under these two types, there are respective models to accomplish different types of tasks. In my project, machine learning model is a Python file and it could recognize certain types of pattern after being trained. The model is trained with a set of data and provide it with an algorithm, and it could make inferences based on previously unseen data and make a prediction. Therefore, the model is used to solve a variety of different types of problems. For under supervised learning, the model could be divided into many categories, one of which is a classification model. For example, the classification model could be used for spam detection, sentiment analysis, etc. The other type is regression model. For example, regression models could be used to predict housing prices, stock prices, or height and weight. In this project, sentiment analysis is the core part of the entire project. The classification model could complete the task of sentiment analysis and help me get the results I want.

After obtaining the database from twintDB, the sentiment analysis method could be used to analyze people's comments on the artist, so as to obtain the result of sentiment analysis. There are two results, and they are positive and negative. I assigned a positive result as 4 and a negative result as 0. For example, I love John Lennon, which is a positive sentiment, has sentiment analysis result of 4. I hate John Lennon, which is a negative result, and its sentiment analysis value of 0. There are two main ways of sentiment analysis, one is to establish an emotional dictionary, and the other is to obtain results through machine learning. Although the accuracy of sentiment analysis through the establishment of sentiment dictionary is relatively high, it takes a long time for manual work. While machine learning could reduce the time consumed by manual work under the same accuracy rate, but it takes longer time to train the model. Therefore, in this project, I chose to use machine learning methods for sentiment analysis.

The core of sentiment analysis is a classification problem, and as I mentioned above, the positive result is assigned a value of 4, and the negative result is assigned a value of 0. Therefore, the model used for sentiment analysis is a classification model, which classifies the results as positive or negative. In sentiment analysis projects, commonly used models are K-NN, Naïve Bayes and SVM.

K-NN is the k-nearest neighbors algorithm. It is a relatively simple classification algorithm. It predicts the classification of new sample points by identifying data points that are divided into several classes. The advantage of K-NN is that the algorithm is easy to understand and it is very friendly to beginners. However, its disadvantage is that the accuracy rate is not high, and it cannot be used in real life.

Naive Bayes classifiers are a family classifier algorithm. Naive Bayes has a very important premise, that is, in this model, even if the features in the model are interdependent or affecting each other, the Naive Bayes algorithm considers these features to be independent. The advantages of Naive Bayes are simple algorithm, fast running speed and good scalability. However, the disadvantage is that the accuracy rate is not high, and it is not practical.

SVM is support-vector machine. Support vector machines can be used for both classification models and regression models. The advantage of support vector machines is that the accuracy is very high, probably above 85%. But its disadvantage is that training is very time consuming.

After learning about several classification algorithms, I decided to use the Naive Bayes classifier as the core algorithm of my project. The naive Bayes classifier model is based on the Bayes formula. Therefore, to understand the principle of the Naive Bayes algorithm, I would understand the Bayes formula firstly. The Bayesian formula is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Figure 4: formula 1

P(A) is the probability of occurrence of event A, and P(B|A) is the probability of occurrence of event B under the premise of occurrence of event A. Similarly, P(A|B) is the probability of event A occurring under the premise of event B. Therefore, the Bayesian formula is to study the probability of two independent events affecting each other. For example, to judge whether a person is a basketball player or not, it may be judged by other characteristics such as height, weight, etc. Therefore, judging the result of an event may have particularly many

characteristics. Hence, the naive Bayes formula is the problem of multi-feature classification under the Bayes formula, so its formula is derived as

$$P(y|x_1,\ldots,x_n) = \frac{P(y)P(x_1,\ldots,x_n|y)}{P(x_1,\ldots,x_n)}$$

Figure 5: formula 2

There are three naive Bayes algorithms under Naive Bayes, and they are Gauss Naive Bayes algorithm, multinomial Naive Bayes algorithm, and Bernoulli naive Bayes algorithm. This is because independent events may be Gaussian distribution, multinomial distribution and wave effort distribution. Therefore, under different distributions, there are three different naive Bayes algorithms.

After choosing which algorithm to use, you can start to build a machine learning model. Machine learning is roughly divided into seven steps, which are collecting data, preparing data, selecting models, training models, evaluating models, adjusting parameters, and practical applications. Firstly, collecting data is to obtain some classification criteria through a large number of labeled data sets, and make the machine record the criteria. For example, the sentence I love you is marked as 4 because it is positive. However, the sentence I hate you is marked as 0 because it is negative. Secondly, preparing the data is to use 80% of the data as the training set and make the machine learn this classification rule. Using 20% of the data as the test set to see how the machine learning is performed after machine learning, and how accurate it is. Thirdly, a suitable model could be selected. Fourthly, through programming, the model could be trained. Fifthly, under the programming library, there are special functions to help output the training results, so the specific accuracy of this model could be shown. Sixthly, by adjusting the parameters, the model may have a higher accuracy rate. Finally, once the model is well trained, the model could be applied in my project.

After understanding the specific machine learning process, I started to build machine learning models through programming. In this project, I used scikit-learn in Python. Scikit-learn is a free machine learning library for Python programming. Firstly, I downloaded the open source and free sentiment analysis data. There are 160,000 tagged sentiment analysis data in this data. Secondly, through the function of scikit-learn, 80% of the data is used as the training set and 20% of the data is used as the test set to obtain the accuracy of the sentiment analysis model, ensuring accurate results and applying them in actual projects. Thirdly, in this project, I chose

the Naive Bayes algorithm. There are three algorithms under the Naive Bayes algorithm. After training the model, I found that the Bernoulli Bayes algorithm has the highest accuracy, so I chose Bernoulli Bayes as the core algorithm for the model. Fourthly, run the program and train the model. By testing some simple sentences, I need to verify whether this model could recognize simple sentiment analysis tests. Fifthly, I checked the accuracy of the model through the predict function. After checking again, the accuracy of my model is 80%. Finally, I used the machine learning model to get all the comments about the singer in the twintDB table collected before to get the sentiment analysis result.

4. Data Visualization

After getting the data results, how to display the data clearly and easy to understand is also very important. Data visualization is to visualize complex data through graphical analysis and charts. Data visualization can be achieved through some charts, such as pie charts, bar charts and line charts. The advantage of the pie chart is that it could clearly show the proportion of each factor. The advantage of the bar chart is that it can easily compare various factors, so that people can easily observe the differences between these factors. The advantage of the line chart is that it can reflect the trend of the factors and let people understand the dynamics and direction of the development of things. Therefore, data visualization means that complex data becomes clear through simple charts, which are easy to understand and visually appealing. Therefore, data visualization has several advantages.

Firstly, in the same area, data visualization can display more information. The expression of text messages is subject to the limitations of text size. However, when we establish a coordinate, a point in the coordinate system could represent a piece of data, and a point is much smaller than the text. Secondly, the use of data visualization charts could quickly discover information characteristics. By connecting one data point after another into a line, the data could be formed as a whole. Observing the whole in this way, we could find the law and trend of data changes. Finally, data visualization makes the difference in data size more obvious. As data sets become larger and larger, people need to perform calculations and comparisons to discover the differences between the data. However, through data visualization, with the help of the difference of the position or color of the point, the difference of the data can be perceived through a clear image, so that the data could be analyzed faster and conclusions could be drawn. There are many techniques that could be used to visualize data. In this project, I chose dash board to make data visualization charts. Dash is a Python framework created by plotly for

creating interactive web applications. With dash, you don’t have to learn HTML, CSS and Javascript in order to create interactive dashborads, you only need Python.

I made three charts to show the direction of data changes. The first is that the chart is a pie chart, which shows the percentage of fans of each music genre in all music genres. Through this pie chart, we can easily observe that the type of music that most people prefer is popular music. The second graph is the comparison of the number of fans of some singers in each music genre. It is a bar graph. It can be observed that even for singers of the same musical style, there is still a big difference in the number of fans. Finally, there are two line charts. The first line chart is to show the trend of the singer's sentiment analysis value on Twitter in a period of time. The second graph is to show the singer's play volume change of a song in the same time. Through the comparison of these two charts, it can be analyzed whether the sentiment analysis value has a significant impact on the playback volume of the singer's song.

Result and Analysis

1. People's preference for music genre

In different countries and regions around the world, people may have different preferences for music types. Therefore, I collected the number of followers of the top 200 artists ranked by Spotify users worldwide from January 1, 2017 to August 1, 2021. Every singer releases many songs, but the types of these songs are basically the same. Few singers release many types of songs. Even so, most of their songs still belong to the same genre. Therefore, I collected the number of followers of the singer and the type of music style of the singer in the table SpotifyAPI. However, this line of music style is too specific. For example, the band ARIZONA, their music types is pop, electropop, pop edm, pop rock, etc. Therefore, it is impossible to treat each music type as an individual genre. If I classify music genre based on what I collected, there are too many music genres in the table. Therefore, I took the classification of the 7 main music types on Spotify's official website, and then classified the singers according to the key words of each singer's music type. In the end, I obtained the sentiments of various countries and regions around the world for different music types, and it made the following figure 6. From that chart and table, it is evident that popular music is the most popular music genre, and it occupies 51.7%. In addition to other particularly niche types, funk music ranked last, accounting for 1.71%. Maybe this is not in line with the perception of many people, and they will think it is unreasonable, because there will be many people around the world who like funk music. However, this is global music data, and music may be regional, so this result is relatively reasonable.

From the results, pop music is the most popular music, which actually meets my expectations. This is because the name of popular music is pop. On the whole, I think the entire chart is relatively accurate. However, due to the large amount of data and too many types of music, some of the music classification is incorrect, and some of the data is also inaccurate, so I could only say that this graph reflects a certain trend, but it is not 100% accurate. For example, singer Amanda Shires, has song genres including alternative country, folk, new Americana. From her song genre alone, it is difficult to classify her into a specific genre, because these are three different genres and cannot know which type of music is her main style, so she could be classified under the existing folk genre. Therefore, the data in this chart is roughly accurate, but there is no way to give 100% data support to illustrate people's specific preferences for

music types. However, a specific type of music, such as pop music, is still relatively more popular in most cases.

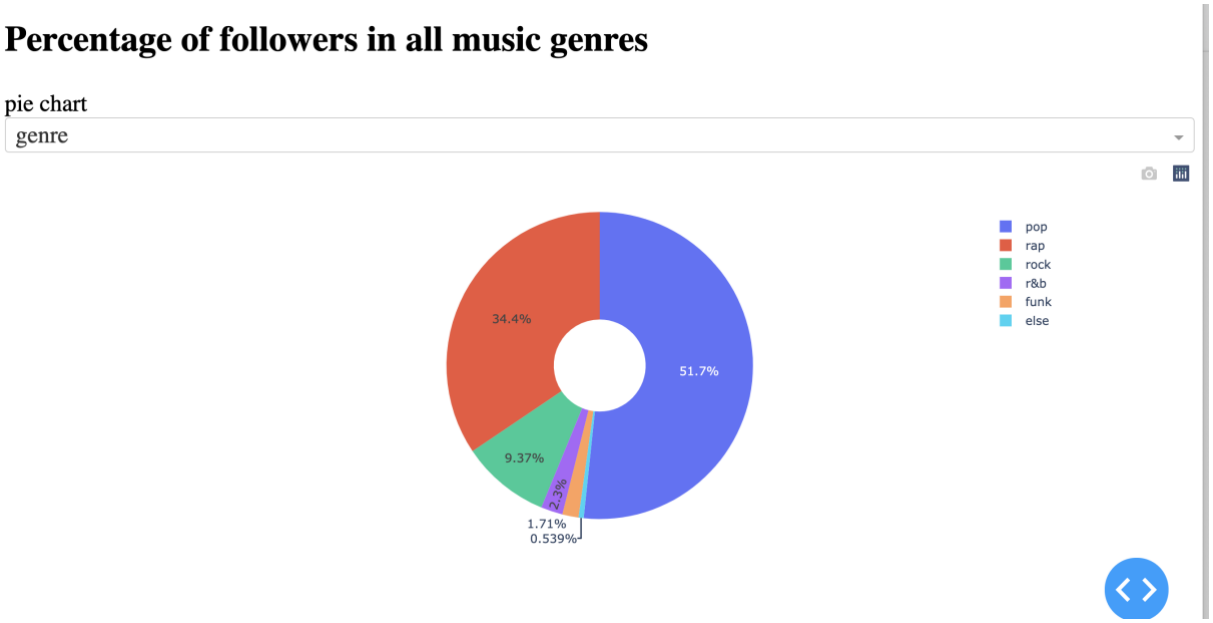


Figure 6: pie chart

2. Comparison between singers of the same genre

In the first result, pop music is the most popular type of music. However, as long as you make popular music, can you definitely be welcomed and loved by people? Under the same music genre, are there still big differences between singers and to what different degrees they are loved? Therefore, according to the classification of the first chart, on figure 7 a comparison of the number of fans among the 7 different types and the same type of singers to analyze whether there is a big difference between them. The bar chart I made is as follows, showing the number of fans of different singers of the same music genre. For example, for the same pop singer, the number of followers of the singer 5 Seconds of Summer is 7,787,543, but the number of followers of A Touch of Class is 18,779. Even though they are both pop singers, there is still a huge gap between them. As mentioned earlier, funk music ranks low, but there are still particularly popular singers. For example, the singer Anitta has a number of followers and the number is 10,767,933. Therefore, it is difficult to conclude that singers of a certain music genre will be welcomed by more people. This is because, there are still singers who are particularly popular even in a niche music style.

Therefore, through the analysis of this bar chart, it is impossible to support the previous conclusion. However, it is not inconsistent with the conclusion of the first discussion, because the previous analysis was a whole, and now this bar chart is analyzing specific individuals.

Therefore, the individual differences are very large, but the group differences will be much smaller, so they are not contradictory.

Figure 7: bar chart

3. If the sentiment analysis value affect the amount of music play by singers

On Twitter, people will tweet about singers or singer-related topics, which can actually reflect people's true views about singers to a certain extent. Therefore, I used the name of the singer as a key word, collected people's tweets about the singer's comments from Twitter, and saved it them the database. Through the sentiment analysis model, the sentiment analysis results are obtained. During the time when the Twitter user tweeted, I also saved the change in the number of music played of the singer's relatively popular songs in the database. Therefore, I intend to analyze whether there is a relationship between the sentiment analysis value and the amount of

music played by singers, that is, whether the results of positive or negative sentiment analysis will have a positive or negative impact on the amount of music played by singers. Two sets of line graphs were made through dash to compare the results of sentiment analysis and the amount of music played. The first line chart is the change trend of sentiment analysis value in a period of time. The second line graph is the changing trend of the singer's music playing volume in the same time period. This is because the amount of data is too large and there is too much data, I will show two examples to analyze the relationship between these two factors.

For example, the first singer is $NOT. From March to May, his sentiment analysis results are very unstable. The highest sentiment value is 3.11 and the lowest sentiment value is 2. However, his sentiment value tends to be stable from May, and it is around 2.4. The second line chart shows the trend of the amount of music played of his popular song. It reached the peak of its played before May 30, making it broadcast 1052774 times a day. After that time, the amount of his songs being played showed a downward trend, until he disappeared from the top 200 charts. Therefore, according to the trend change, it could be observed that the sentiment analysis value hardly affects the amount of music played by the singer. Let's look at another example, the singer is (G) I-DLE. His sentiment analysis results continued to rise from February 18, and began to stabilize after March 18, stabilizing at around 3.01. However, the amount of his songs being played in November is very unstable. On November 9th, it reached the highest number, which was 110224, but it began to continue to decline. Through these two examples, it is not difficult to see that, in fact, the quality of the sentiment analysis results has little impact on the amount of played of singer songs. It could be said that the impact of positive or negative comments on music is very low, almost negligible.

Although the results of sentiment analysis show that the results of sentiment analysis have little effect on music trends, results are only a possibility. This is because the analysis of data does not reflect the real situation. First of all, the amount of data about singers' comment is relatively small. Because of the twint tool and twitter API, I only collected less than 100 comments for each singer, and the time distribution is not very uniform, so the amount of data may have a great impact on the analysis of the results. Secondly, there are many reasons that affect the amount of music played, and sentiment analysis may only be a factor that has a small impact. Therefore, even if some influential data visualization results are obtained, it could not be concluded that it will cause an impact. After careful analysis, I think sentiment analysis may have a lower impact on music trends.

Changes in results of sentiment

Changes in the amount of music played by artist

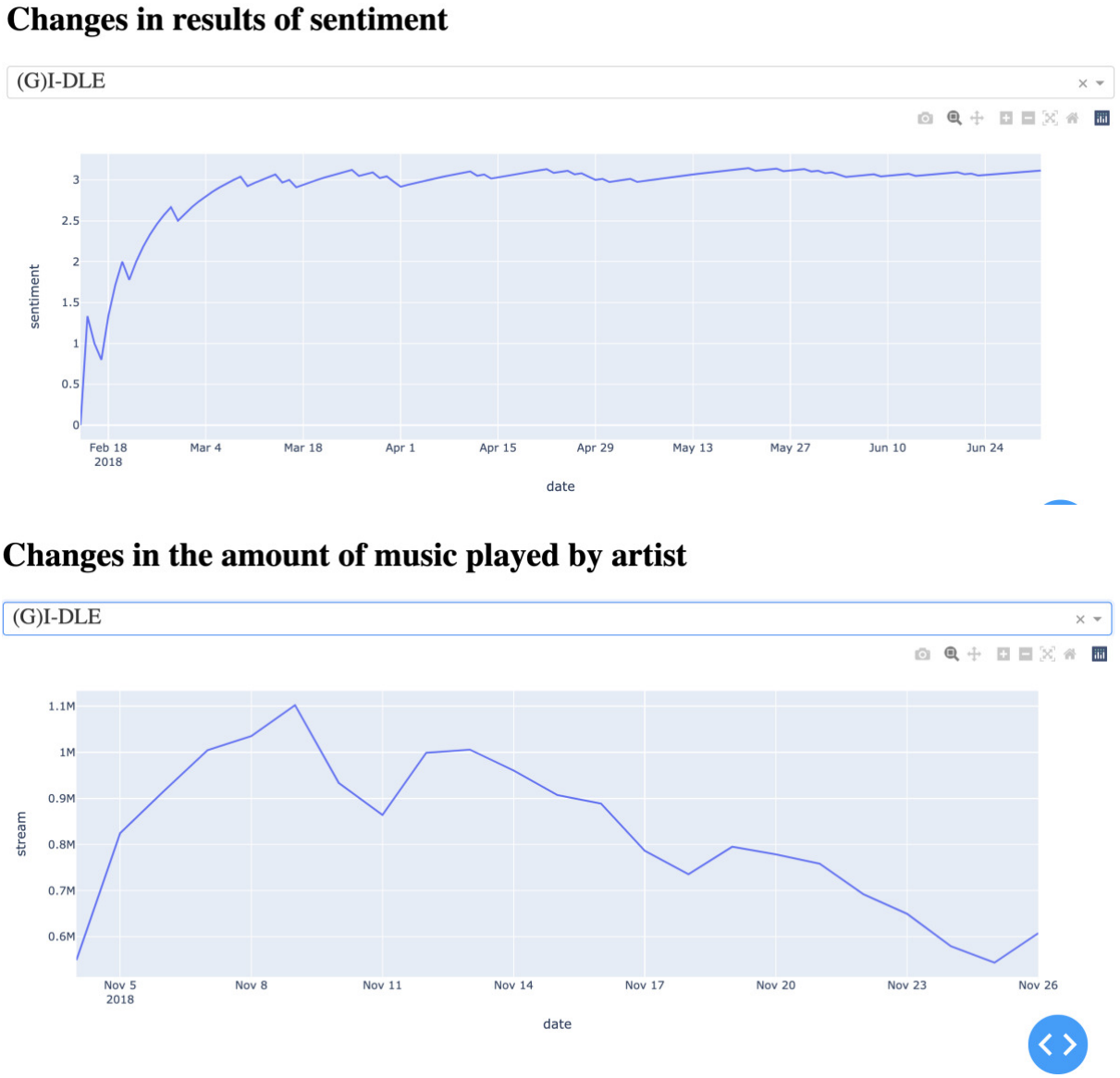


Figure 8: line chart

Conclusion

In the previous chapter, I proposed 4 goals. After the previous analysis, we could now get some conclusions to answer the previous questions.

The question of goal one is whether a higher sentiment analysis value will keep the singer's song in a higher ranking on the music chart. In the previous chapter, I analyzed the changes in the trends of the line graphs, and also analyzed the relationship between the sentiment analysis value and the singer's music playback. Therefore, it could be concluded that more positive comments about the singer on Twitter will not keep the singer's music ranking in a higher position. On the contrary, the music rankings of singers may fluctuate significantly.

The second question is whether negative comments will affect song rankings and the amount of music played. This question is actually very similar to Goal One, and it's all about the impact of comments on the amount of music played. Therefore, we could also conclude from the first result that, in fact, negative comments have a very low impact on song ranking.

The question of goal three is whether the genre of the song will affect the ranking of the music. According to the pie chart in the previous chapter, we know the comparison of the number of followers of each music genre. Therefore, as a whole, the type of music will affect the ranking of music and other related data. This is because certain types are more popular with people as a whole.

The question of goal four is whether the number of followers of a singer will affect the amount of music played. Same as goal three, the high number of followers of the singer represents the popularity of the singer. Therefore, the overall amount of music played by a singer is related to the number of followers of the singer, and the greater the number of followers, the higher the overall amount of music played.

Future work

This is the first time I have done a machine learning project, so I spent a lot of time learning basic theoretical knowledge, and I spent more time on how to use this knowledge in my project. Therefore, there are still many shortcomings in my project that need to be overcome, but I also gained experience in the process and learned how to better complete machine learning projects. Therefore, after submitting the dissertation in the future, I will personally continue to do some research to make this project more perfect and try my best to make it applicable in real life, whether it is to analyze music trends or other forecasting tasks. In my future work, I hope to improve in three areas. They are data collection, core machine algorithm models, and analysis methods.

1. COLLECT DATA

The first is to improve how to collect data faster. In forecasting projects, data is the most basic and most important, and collecting data will consume a lot of time. For example, when collecting user comments on Twitter, it is impossible to let a program run for a long time, so that data collection could not be completed at one time. On the contrary, due to the long-term establishment of a network connection with the server, the server itself will block such a long-term connection. Therefore, many problems arise when collecting data, and it takes a lot of time to deal with these problems. In this project, I collected a total of 1,600 singers' information. About every 10 singers' comment information, the server will automatically disconnect the link, so I could only solve this problem manually, which consumes a lot of time. Similarly, due to some problems with the server itself, I cannot collect tweets other than 12 o'clock that day, so I cannot guarantee whether the data is comprehensive. Therefore, it is better to collect data completely automatically. After knowing these possible problems, I could also write a program that is almost automated to collect these data at once to avoid wasting time on this, and I could concentrate more time to deal with follow-up work.

2. STORE DATA

After the data is collected, the structure of the data stored in the database also needs to be re-adjusted. This is because before organizing the data and visualizing the data, I did not know how the data should be stored, so I stored the data according to the source of the data. However, storing data in this way is very inconvenient when writing some machine learning code and data visualization code. Instead, I need to write more code to organize the data. Therefore, how

the data is stored is also a factor that greatly affects efficiency. Before the next time the data is collected and stored, it is necessary to plan how the data will be stored. The storage of data is designed visually according to the subsequent steps, so I need to make sufficient preparations to plan the purpose and direction of this project from the beginning, and it will make the code more redundant in designing the structure of the stored data and avoid wasting time and space.

3. SORT DATA

In fact, due to the large amount of data, I have great doubts in my mind about whether certain data is really what I need. For example, when collecting Twitter comments, I use the artist's name as a key word to collect comments. However, the names of some of the singers may be repeated with others, may be a very common word and appear in sentences frequently. Therefore, it is impossible to be 100% sure, this Twitter comment is a comment on this singer. If this is the case, a lot of the data is wrong, and this will even affect my final conclusion, and maybe the conclusion I got is also wrong. Therefore, the processing of data is a very important part of this project. I plan to write a program or find a mature third-party library to determine whether the comments I collect are related to the singer. If it is, then store the data. If not, the program will delete this piece of data and continue to judge whether the next piece of data meets the requirements. On the other hand, I also plan to deal with emojis. This is because people not only express opinions through words, but also express comments through emojis. However, in this project, I filtered out all emojis, so in future plans, by adding an analysis of emojis, it may be possible to more accurately judge whether the information sent by people is positive or negative.

4. ALGORITHMS

In this project, I used the Bernoulli Bayes algorithm as a Naive Bayes algorithm. In fact, this algorithm is not the best algorithm among the classification algorithms for sentiment analysis. There are many algorithms that are more accurate than Bernoulli Bayes algorithm in sentiment analysis classification. Therefore, I plan to use three methods to improve the accuracy of the classifier. The first is to still use traditional machine learning, but I will use other classifiers with higher accuracy, such as support vector machine algorithms. Although I will still spend time learning new algorithms and learning to apply algorithms in projects, I already have experience in machine learning, so the learning time will be greatly shortened, thereby improving the accuracy of the sentiment analysis classifier. The second method is to use the

latest deep learning model to improve the accuracy of the classifier. Compared with deep learning, machine learning models are relatively simple and elementary. Deep learning models are better at processing classification tasks and have higher accuracy. Much sentiment analysis is now done using deep learning, not machine learning. Therefore, using deep learning is a more advanced and accurate method to complete sentiment analysis classification. The third method is to use machine learning and sentiment dictionary at the same time to complete sentiment analysis. As I mentioned before, sentiment analysis could use the method of building a sentiment dictionary. Therefore, here I intend to use the method of combining sentiment dictionary and machine learning to experiment in the project. Perhaps the accuracy of the combination of the two is higher than using them alone.

5. DATA ANALYSIS

As a separate subject, data science has many analytical methods. Data analysis is to rationally analyze the data collected and obtain relatively objective results. By processing, sorting and analyzing a large amount of data, we could extract useful information and form conclusions to reflect a relatively objective and true result or phenomenon. Data analysis has many advantages. Firstly, data analysis could help companies make better decisions. For example, the pressure of competition among enterprises is great, so enterprises want to acquire more users. Data analysis could help companies predict user needs, which could help companies make better decisions and improve user experience. Secondly, data analysis could help companies increase productivity. Through data analysis, companies could fully understand their strengths and weaknesses, so this could help them to maintain their strengths and make up for their weaknesses. Therefore, this could help companies increase productivity. Thirdly, data analysis could help companies increase flexibility. Through the results of data analysis, companies could better adjust their business, which could help companies change their strategies and directions.

In this project, the analysis method I used is just to analyze and predict the entire research through some simple data and data changes. From the perspective of data analysis, such an analysis is not enough to get the following conclusions. Therefore, I need to supplement my knowledge in data analysis to analyze my project through a complete data analysis method.

Reflection

Driven by my curiosity about the future world, I chose the project of predicting music trends. For me personally, most of the computer technology in this is something I have never used before, so this is still a big challenge for me who has just started to learn computer science for a year. Before that, I was actually a person who was afraid of challenges. I had never lived and studied alone. However, in the year I came to the UK, I think I have grown a lot, maybe my improvement in this year is comparable to the previous 5 years or even 10 years of progress, because I am walking on the right path It's my best effort. In terms of learning English, I was unable to communicate with native speakers at the beginning, and now I can communicate with everyone in English normally. On the computer side, I only wrote some simple codes from the beginning, and now I can complete a project with the help of my instructor. In terms of interpersonal communication, from when I was afraid to communicate with strangers, I have learned a lot of social skills and made some friends from all over the world. I think these are things that I have never done before, so I want to thank my mentor and my friends. The most important thing is that I have clarified the direction of my life, set a life goal for myself, and am moving towards this goal. I think these are the changes after I came to the UK and started learning computers. Of course, in the process of studying and living, I also realized that many things were not done very well.

In the following, I will reflect on the things I did well and badly in the whole project from three aspects.

1. Programming

The first is reflection on computer technology. In this project, the first new technology I learned was crawler technology. Crawler is actually a relatively simple and very friendly technique for beginners. After learning the crawler framework, I could get some web page data with some simple codes. In this project, my landing page is also a relatively simple page. Therefore, I spent about 3 days to learn the crawler technology, so as to successfully obtain the data I want to collect. I think that in terms of using crawler technology, I completed the task very well, and learned the crawler framework Beautiful Soup. After mastering it, I could easily use crawlers to obtain data in the future. I think the key to using the crawler framework lies in the understanding of HTML and crawler functions. First of all, we must know the location of the acquired data in the HTML, so that the data could be easily acquired through the corresponding function.

Database is the second technology I learned in this project. Although I have used the database MySQL before, I only know some simple queries. This is because the amount of stored data is relatively small, the data is only read and stored through the database. Therefore, this is a new knowledge for me. In addition, the database I use is MongoDB, which is a non-relational database. Before I know how to organize the database structure, what I need to do is to successfully store the data in the database first. However, when using non-relational databases, I think the storage structure of the data is very important, which could provide convenience for subsequent operations on the data. Although I successfully stored the data in the database, the data structure I designed was relatively bad, and I spent more time processing these poorly structured data when using the data.

In this project, the most important part is machine learning. For junior scholars, machine learning is very difficult. I spent a long time learning the machine learning part, and it took me a long time to sort out many concepts in machine learning. Although these concepts are of little practical use, if I don't have a deep understanding of them, I won't be able to start working on this project. Therefore, first of all, after fully understanding some concepts in machine learning, it will be easier and more flexible to use functions to build machine learning models in actual operations. I think that machine learning is not just about calling these functions, but also a clear understanding of the model used, so that the appropriate model could be used to solve practical problems. For example, some of the data in this project has not been sorted out. If such inaccurate data is used, it will actually affect the final result to a certain extent. I think I fully understand the principles of the model I use and have completed the establishment of the sentiment analysis model very well. However, it took me most of my time, and spent a lot of time on training the machine learning model and getting the results, so I will shorten this part of the time in other ways in the future.

Data visualization is the last part of computer technology. Although the technical difficulty of data visualization is much lower than machine learning, it is as important as machine learning. This is because the purpose of data visualization is to show people the final results of the project. If the project is not well displayed, even a good research and result will not be understood and recognized by people. I chose dash as the framework for data visualization to display my final results on the web. This part of the presentation is very important, but I do not have the knowledge of data visualization and I had no time to learn it. I just showed what I knew and thought about, and didn't have a specific method, so I think my data visualization results did not show the research results of my project very well. I think this is a relatively poor part of what I did, so I plan to learn more about data visualization, not just technology, but how to

make a good data visualization concept, so that my results will be more deeply rooted in the hearts of the people and it could be better.

2. PROJECT MANAGEMENT

In addition to computer knowledge, in fact, how to do a good management of the project is very important. The use of project management could let managers know the progress, goals, and costs of project development, so as to get feedback and accomplish this matter. If you do not plan and manage the project, you will be blind, and the results may be of poor quality, and you may even give up halfway and fail to complete the task. However, although I made a simple plan to plan the progress of my time and tasks, I did not use specific tools to execute and get feedback from it. I think there are two reasons for this. The first reason is that this is my first machine learning project. It is a completely unfamiliar project and I don't have any experience. Therefore, I cannot estimate how long I will spend on each step, so my plan is actually not worthy of reference. The second reason is that I did not pay much attention to the management of the project, and the execution ability in the project was relatively poor, which resulted in me not knowing the progress of the project for a long time. In short, I did poorly in project management, which also caused me to delay 4 weeks to complete my project. In addition, there are also many objective factors.

3. DATA ANALYSIS

The last part is about data analysis. Because there is no good project management, I never thought of learning how to do a good data analysis and show the results of my project research. Therefore, my data analysis part is actually some simple analysis and summary of the data from my personal perspective. In this critical part, my analysis is too simple. It only uses simple data as the support for the results. Such an analysis is a bit simple. Therefore, I did a poor job in this part. I plan to learn the knowledge of data analysis, because it is very important for the research and presentation of the project. It can help me make a good adjustment to the structure of the data, and it can also help me better analyze and display the results of the project research, making my research more reasonable and logical. For example, for music prediction projects, good data analysis can make the analysis more reasonable, which can improve the accuracy of the prediction and achieve the purpose of project research.

Reference:

Sandstrom, G.M., Russo, F.A., 2010. Music hath charms: the effectsof valence and arousal on recovery following an acute stressor.Music Med. 2, 137—143.

Abraham, S. et al. 2015. Trait anger but not anxiety predicts incident type 2 diabetes: The Multi-Ethnic Study of Atherosclerosis (MESA). Psychoneuroendocrinology 60, pp. 105–113. doi: 10.1016/j.psyneuen.2015.06.007.

Chamberlin, D.D. and Boyce, R.F. 1976. SEQUEL: A structured English query language. In: Proceedings of the 1976 ACM SIGFIDET (now SIGMOD) workshop on Data description, access and control - FIDET '76. Not Known: ACM Press, pp. 249–264. Available at: http://portal.acm.org/citation.cfm?doid=800296.811515 [Accessed: 22 October 2021].

Dhenakaran, S.S. and Sambanthan, K.T. [no date]. WEB CRAWLER - AN OVERVIEW., p. 3.

Feldman, R. 2013. Techniques and applications for sentiment analysis. Communications of the ACM 56(4), pp. 82–89. doi: 10.1145/2436256.2436274.

Pagán, J.E. et al. 2015. A repository for scalable model management. Software & Systems Modeling 14(1), pp. 219–239. doi: 10.1007/s10270-013-0326-8.

Chamberlin, Donald D; Boyce, Raymond F, 1974. "SEQUEL: A Structured English Query Language"(PDF). Proceedings of the 1974 ACM SIGFIDET Workshop on Data Description, Access and Control. Association for Computing Machinery: 249–64.

Pagán, J.E. et al. 2015. A repository for scalable model management. Software & Systems Modeling 14(1), pp. 219–239. doi: 10.1007/s10270-013-0326-8

El Naqa I., Murphy M.J., 2015. What Is Machine Learning?. In: El Naqa I., Li R., Murphy M. (eds) Machine Learning in Radiation Oncology. Springer, Cham. https://doi.org/10.1007/978-3-319-18305-3_1

Liu B., 2011. Supervised Learning. In: Web Data Mining. Data-Centric Systems and Applications. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-19460-3_3