Automatic Finding Sales Lead for Concert Ticket

Kexin (Colleen) Su

05/13/2020

Abstract

Successfully finding good sales leads can drastically increase the company's conversion rate, thus implementing an efficient and accurate sales-lead-finding process has become more and more important to companies. In this project, I will design a system that automatically helps ticket sellers promote their concert tickets by first selecting concerts that should be promoted, and then recommending potential customers corresponding to those concerts. The system will utilize methods such as sentiment analysis, search query index, knowledge graph, etc. to analyze twitter user's information, and determine if the user is a potential buyer. In the end, the system will return a list of potential customers with their twitter user information.

Introduction

Sales leads, in simple words, is just people or businesses that are potential buyers or potential customers of a product. A high-quality and revenue-producing sales leads can increase the sales of the product. Therefore, an accurate sales-lead-finding process is critical to a company because it enables the seller to spend less money on marketing, yet have a higher conversion rate.

However, the current situation, according to Hubspot's report in 2018, is that about 60% of marketers think it is a great challenge to generate traffic and leads. In reality, a significant amount of salespeople complain that they encounter and waste a lot of time on bad leads by

marketing to someone who is not interested in their product. Therefore, in this project, I want to design a system that automatically helps ticket sellers promote their concert tickets by recommending potential customers; this system will ideally make the sales-lead-finding process more accurate and efficient.

The system will be composed of two parts: automatic concert selection and automatic sales leads finding. In the first part, the system will take in all concert information, and use natural language processing methods on search query data and twitter data to determine which concerts will have the most significant impact when being advertised. Then in the second part, the system will create a knowledge graph for each performer using search query data, then analyze twitter data to find people that are more likely to purchase the concert ticket. In the end, the system will return the user information to the seller so that the marketing team can reach out to them.

Related Work

Marketing automation is a popular method for creating a personalized customer experience. The Strategic IC (2017) has reported that the revenues of the companies with lead management automation increase 10% or more in just 6 to 9 months. As personal information and preference become more accessible to others through internet, a lot of companies are implementing recommendation systems for their customers. Recommendation system can make sure that their products target their existing users more accurately, and thus increase the chance of existing users continue to purchase their products. For example, in Zhang et al.'s research paper (2019), they implemented a personalized web-based movie recommendation system that utilizes publicly

available movie datasets and collaborative filtering to improve performance metrics. They also opened up a real website and collected user-feedback on recommendations to evaluate the accuracy of their system. By personalizing recommendation, the system increases the chance that the user is interested in the film, and thus increase the likelihood of user purchasing the ticket.

However, this type of recommendation system requires historical user data, and are more targeted to maintaining existing user instead of attracting new customers. It also focuses more on the buyers' side, tries to analyze what is more intriguing to the customer. If the company wants to generate new sales leads and find new customers, the mindset needs to switch to the seller's side—we don't need to know what other things this person likes, all we want to know is if this person will be interest in our specific product. In Zeng et al.'s research paper (2018), they designed a leads-generation robot that intelligently understands the requirements of leads and automatically mine the leads from web data to recommend them to salesman. This structure is very similar to the second part of my system; however, in addition to this part, my system also includes a first part that helps the seller decide which product should be promoted by combining seller's concert information with current trends.

After reading through some journal articles, I find out that some researchers utilize search query data to discover new trends and public interests. In a research done by Caputi et al. (2017), they used Google search query data to estimate and predict the ascending public interest in heat-not-burn tobacco product by averaging monthly searches for related products. Coogan, Sui, and Raubenheimer (2018) also found that monthly trends in search query data can be used to represent and predict dietary behavior at the population level. Both studies have utilized search

query data to accurately monitored public interest, so I decide to combine this approach with concert information to more accurately decide the impact of advertising each concert.

Data

Concert Data

For the concert dataset, in order to mimic the concert information provided by the seller using this system, I scraped the concert information from SeatGeek website using their official developer API. This dataset contains the concert information for concerts happening in New York in the upcoming three months. The original data I got using the API was in JSON form, so I saved it into a CSV file for later use. The resulting dataset has 20 features, including concert name, performer name, performer genre, and date, etc., and has more than 2500 entries in total.

Search Query Data

For acquiring search query data, I used the platform Google Trend. It allows users to access search index popularity of a keyword within a specified timeframe as well as its related topics and related queries. I choose Google Trend because Google is the largest search engine used within the US so that it can provide a more accurate and holistic representation of the public interest of the US population. Since Google Trend does not have an official API that is publicly available, I used a pseudo API for Python called Pytrend to scrape the information from Google Trend. Specifically, I used Google Trend to get the related topics for music genre and selected

performers. This method will give me a dataframe containing 20 related topics ranked in their search popularity.

Twitter Data

For collecting Twitter data, I used the API called Tweepy. It allows me to input a keyword or a keyword list, and it will return all tweets containing those keywords from the past week. For each performer, 300 tweets are collected and saved into a CSV file, and all of the tweets are cleaned (removing URLs, emojis, hashtags, etc.) using Tweet-Preprocessing API and a self-written method.

Methods

Phase 1

As introduced in previous sections, the system is composed of two main parts: automatic concert selection, where the system takes in concerts information and return selected performer, and automatic sales leads finding, where the system finds potential buyers corresponding to each performer and return their Twitter information.

For automatic concert selection, I want to combine concert information with current trends to more comprehensively decide which concerts will have a more substantial impact and potentially generate more profit when being advertised. For grasping current trends, my first thought was to extract current music trends from entertainment news, then promote concerts related to those trends and topics. With this thought, I gathered entertainment news data using the API on Webhose.io, which includes news title, location, content, etc. Then, I moved on to analyzing the

news data to extract the current trend. I wanted to use LDA topic modeling on the title and content of the news to potentially find music-related topics. However, after examining the output, I realized that since LDA is a clustering method, it will only give you the top keywords that are related to each other. Therefore, it requires the analyst to have a wide range of factual knowledge to be able to identify the topic that includes those keywords. Looking at the resulting keywords from each topic, I found out that most of the keyword-sets do not point to a clear topic that I can identify (See Figure 1). Given this result, it became very challenging to further match the topics with existing concert information and decide on a concert to promote.

```
Topics found via LDA:

Topic #0:
craig daniel bond tom brothel james character odom nevada star

Topic #1:
apos chicago tribune album new adele cubs campaign 25 studio

Topic #2:
star wars george interview political brand clooney crisis sandra bullock

Topic #3:
movie chris war review rock rihanna paris bridge mountain spies

Topic #4:
center clinton murphy sanders eddie afghan trial drugs start working

Topic #5:
makes county return news indian talk new jersey alia bhatt
```

Figure 1

After topic modeling failed, I need to use another method to understand the current entertainment/music trend. After reading through some journal articles, I find out that some researchers utilize search query data and search engine index to discover new trends and public interests, so I decide to use this method. I choose to use the platform Google Trend to acquire search query data. The system will first search for topics related to the keyword “Music Genre”, and select the music genre that is ranked the highest, which means that this genre is currently the

most searched genre. Then the system will use music genre as the first selection criteria and only keep those concerts who performer's genre matches the selected genre.

However, even after we filtered on music genre, there are still hundreds of singers within this genre, so to further narrow down the scope, I used Twitter data. Within the selected music genre, each performer's name will be passed into the system, and it will count the number of tweets containing the performer's name within the past week. The number of tweets related to a performer act as a measurement of the singer's current popularity; if there are less than 200 tweets, it means that the performer is not popular enough, then his/her concert will not be promoted because not many people will be interested in it. For those performers that have more than 200 related tweets, the system will scrape up to 300 tweets for each singer, and clean those tweets to remove useless information such as URLs, hashtags, etc.

Then the tweets for each singer will be combined into a single paragraph, and sentiment analysis will be conducted using Textblob's sentiment.analysis function. The function will calculate the polarity score for each paragraph ranging from -1 to 1, with -1 being complete negative sentiment and 1 being complete positive sentiment. The polarity score for each paragraph represents the overall Twitter sentiment towards each performer. Performers with higher overall sentiment are preferred because it means that more twitter users like this performer, so more people are likely to purchase the tickets. At the end of phase 1, all of the selected singers will be ranked according to their polarity score from the highest to the lowest, and the company can decide how many singers they want to promote. The selection procedure for phase 1 is shown in figure 2.

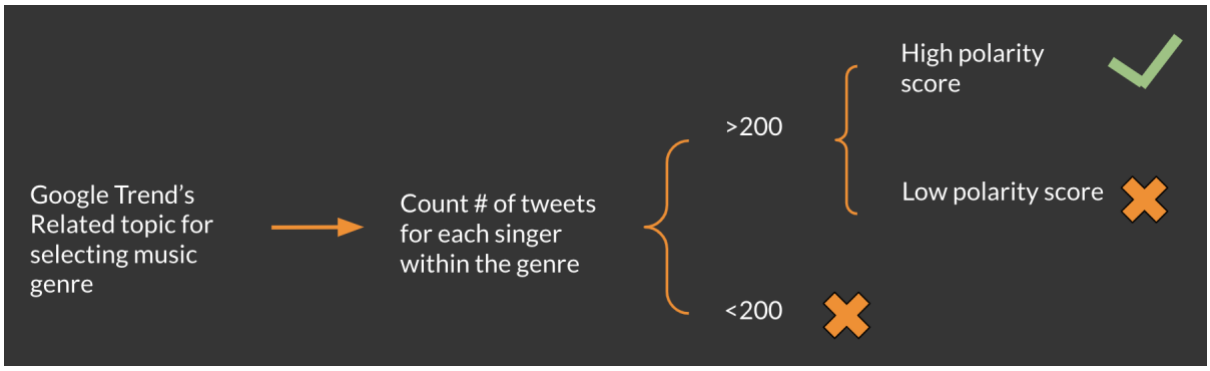


Figure 2

Phase 2

After the concert is decided, the system will move on to part 2, which is automatic finding sales leads. The system will first pass in each selected performer's name into Google Trend's "related topic" function and use related singers, bands, and supergroups to create a knowledge graph around the performer (figure 3). Initially, I also wanted to include song or album name as part of the knowledge graph, but since the name of songs and albums are often general phrases or words, the system will get many irrelevant tweets when collecting tweets related to those keywords. So, I decided to only use singer and band names as the branches of the knowledge graph.

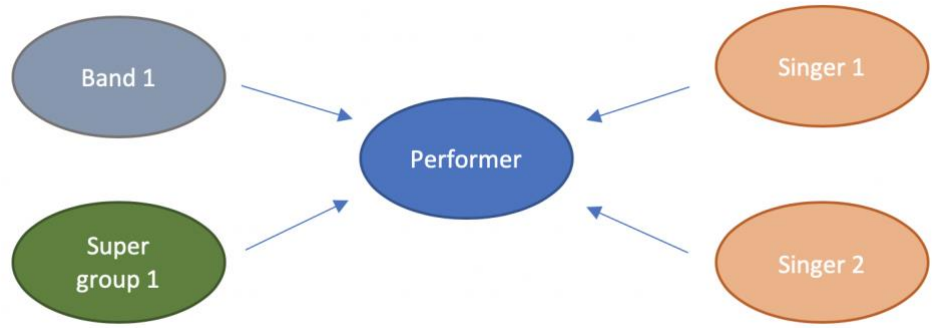


Figure 3

Then, tweets containing the performer's name or containing those related keywords will be collected, and each tweet's polarity score will be calculated. If a tweet related to the performer contain negative sentiment, then the Twitter user sending that tweet will not be considered as a potential buyer. For those users that sent positive tweets, their user location will then be matched to the concert's location. Only those users that posted positive tweets, and whose location is the same as the concert's location will be saved as potential sales leads, and the user's Twitter information such as username, Twitter ID, and location will be collected. In the end, the system will output a dataframe containing the performer's name, and its corresponding sales leads' Twitter information. The procedure for phase 2 is displayed in figure 4.

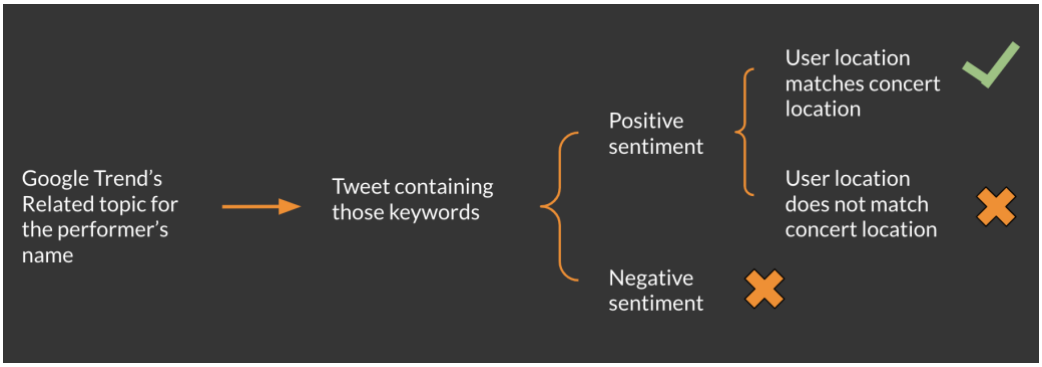


Figure 4

Experiments

After the design of the system is complete, I used the New York upcoming concert data to test if the system works as planned. For phase 1, the system first used Google Trend to search for the

most popular music genre, and from table 1, the genre that is ranked the highest is “Rock”, so the system will only keep those performers whose genre is rock.

	value	topic_title	topic_type
0	100	Music	Topic
1	99	Genre	Topic
2	69	Music genre	Topic
3	5	Rock	Musical genre
4	5	Crossword	Topic
5	4	Musical ensemble	Topic
6	4	Popular music	Musical style
7	4	Musician	Topic
8	3	Pop music	Musical genre
9	3	Artist	Topic
10	2	Hip hop music	Musical genre

Table 1

After filtering by music genre, there are still 80 performers left, so the system will use Twitter data to make further selections. The system then counts the number of tweets related to each performer from the past week, and there are only 31 singers with more than 200 tweets. After this, up to 300 tweets related to each selected performer are collected and cleaned (figure 5), and polarity scores are calculated for each paragraph.

	Performer_name	Tweets
2	PalayeRoyale	Should I name my hermit crab after @RemingtonLeith from @PalayeRoyale ?'
3	PalayeRoyale	b"@PalayeRoyale You're the best! Where it would be?"
4	PalayeRoyale	hi i\xe2\x80\x99m back again with #palayemakeup \nthis time i took inspiration from @SebastianDanzig in the dying in a hot tub\xe2\x80\xa
5	PalayeRoyale	@Demii_louise92 @SebastianDanzig @PalayeRoyale Me right now https://t.co/y9sJUxNOFn'
6	PalayeRoyale	I miss @PalayeRoyale daily tweet https://t.co/bpX8dFs6nl'
7	PalayeRoyale	i just straightend my brothers hair with rem\xf0\x9f\x98\x8e\xf0\x9f\x98\x8e\xf0\x9f\x98\x8e @remingtonleith @palayeroyale https://t.co/
8	PalayeRoyale	Everyday a inspired makeup look by Palaye Royale. \nFashion-art rock \xf0\x9f\x96\xa4\nRemington\xe2\x80\x99s look day 5 \xf0\x9f\x98\
9	PalayeRoyale	@SarahxGibson91 @SebastianDanzig We don\xe2\x80\x99t want that. Just need to wait patiently till this all blows over and we ca\xe2\x80\
10	PalayeRoyale	@PalayeRoyale I hope the US shows go on as planned, I miss you all and can\xe2\x80\x99t wait to see you guys again \xe2\x9d\xa4\xef\xb8'
11	PalayeRoyale	My dumb head decided to translate more @PalayeRoyale songs into German versions... I mean Lonely was easy but why d\xe2\x80\xa6 http
12	PalayeRoyale	@LOVETHEV0ID @SebastianDanzig @PalayeRoyale @EmersonBarrett @RemingtonLeith While listening to Lonely*'
13	PalayeRoyale	petition for @PalayeRoyale to do @MichaelBolten\xe2\x80\x99s makeup on one of the streams \xf0\x9f\x99\x8f \n@RemingtonLeith @Seb
14	PalayeRoyale	b"I haven't watched this in a while, but I just wanna say that this n lonely r only 2 of my fav songs by\xe2\x80\xa6 https://t.co/i92xybmzQL"
15	PalayeRoyale	i don\xe2\x80\x99t have acnh but still. i am a @PalayeRoyale stan even in the game https://t.co/nCqrqebHtA'
16	PalayeRoyale	okay I know I\xe2\x80\x99ve already tweeted these pics but the king in b&w hits different, sorry I don\xe2\x80\x99t make the rules her
17	PalayeRoyale	@PalayeRoyale On twitch or YouTube x'

'This min has an instudio visit withwriterartist We talked Hear the full min interview Talked more came up in conversation withwriterartist Heres the full min convo w writer artist Kelly Brack in Ep219 We talk I really feel for the class of yall been through it Talked Writer artist Kelly Brack is a guest of Episode Wetalked Writer artist is a guest of Episode In the full min convo wetalked This min has an instudio visit withwriterartist We talked Hear the full min interview Talked Heres the full min convo w writer artist Kelly Brack in Ep219 We talk more came up in conversation withwriterartist Heres the full min convo w writer artist Kelly Brack in Ep219 We talk I washed it yesterday lolTalked Writer artist Kelly Brack is a guest of Episode Wetalked Hear the full min interview Talked This min has an instudio visit withwriterartist We talked more came up in conversation withwriterartist more came up in conversation withwriterartist Heres the full min convo w writer artist Kelly Brack in Ep219 We talk Nah he some

Figure 5

Figure 6 shows the distribution of the polarity scores for those selected performers. From the distribution, we can see that it is center around 0.1, and only a few performers have scores above 0.2. Therefore, if the ticket seller wants to decide on which performers to promote, it would be more effective to choose those that have a polarity score above 0.2. Table 2 displays the performers ranked according to their polarity score, and we can see that out of 35 singers, only 5 has polarity scores above 0.2, so it would be a good choice to promote those five performers.

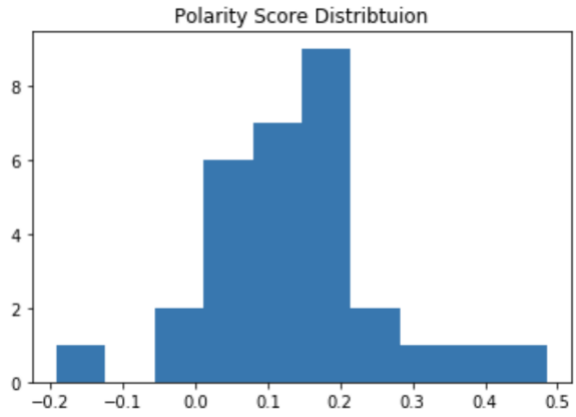


Figure 6

	Singer Name	Polarity Score
0	Palaye Royale	0.486007
1	Alter Bridge	0.391436
2	Nile Rodgers	0.323203
3	The Temptations	0.264479
4	Judas Priest	0.234733
5	Starset	0.197671

Table 2

After the performers are settled, the system will pass each singer's name into Google Trend and get their related singers, bands, and supergroups. For example, in table 3, the band Alter Bridge was searched, and out of the 20 related topics, there are three that are related bands and

supergroups, so the system will search for tweets containing the performer's name or these related keywords. Then the system will calculate the polarity score for each tweet and match the user's location to the concert location (in the case of my concert dataset, the location would be New York). In the end, it will output a table containing the performer's name, and the sales leads' information (table 4).

	value	topic_title	topic_type
0	194800	The Last Hero	Studio album by Alter Bridge
1	85800	My Champion	Song by Alter Bridge
2	69450	Royal Albert Hall	Concert hall in London, England
3	67650	Show Me a Leader	Topic
4	23750	Saint Asonia	Supergroup
13	150	Tremonti	Band
14	130	The End Is Here	Song by Alter Bridge
15	130	Arena	Topic
16	100	Iron Maiden	Band

Table 3

Performer	UserID	Name	ScreenName	Location
Alter Bridge	34318891	The IWC Messiah	JDfromNY206	Bronx, NY
Alter Bridge	1253393944517521411	PoetessIntheDeep	ABpoetess	New York, USA
Alter Bridge	40136182	Tom Travers	tommytravers	Long Island, New York
Alter Bridge	234878051	Loudwire	Loudwire	New York, NY
Alter Bridge	550643117	SlamminTunes	SlamminTunes	New York, NY
Alter Bridge	376796401	Debbie H 🙏🇺🇸🙏🇮🇹🙏🌎	deb_h101	Long Island New York USA🇺🇸🗽

Table 4

System Overview

To build the web application, I used Python Flask together with HTML and CSS. The web application has three tabs. Figure 7.1 shows the home page; it includes a brief introduction to the purpose of the application and detailed explanations of what the other two tabs do.

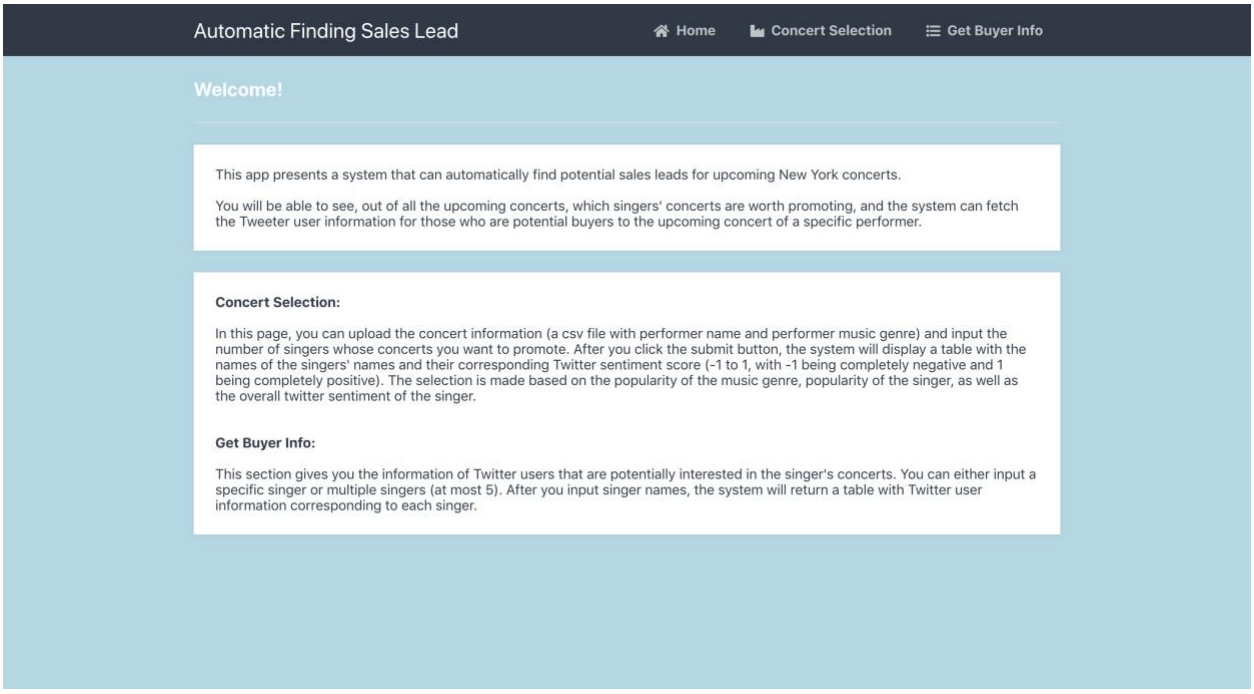


Figure 7.1

Figure 7.2 shows the “Concert Selection” page, which corresponds to phase 1 of the system. The user can upload the concert information (has to be a CSV file and it has to contain performers’ names and performers’ genres) and the number of performers that the user wants to promote, then the system will return a ranked table containing the singers’ names and their polarity scores.

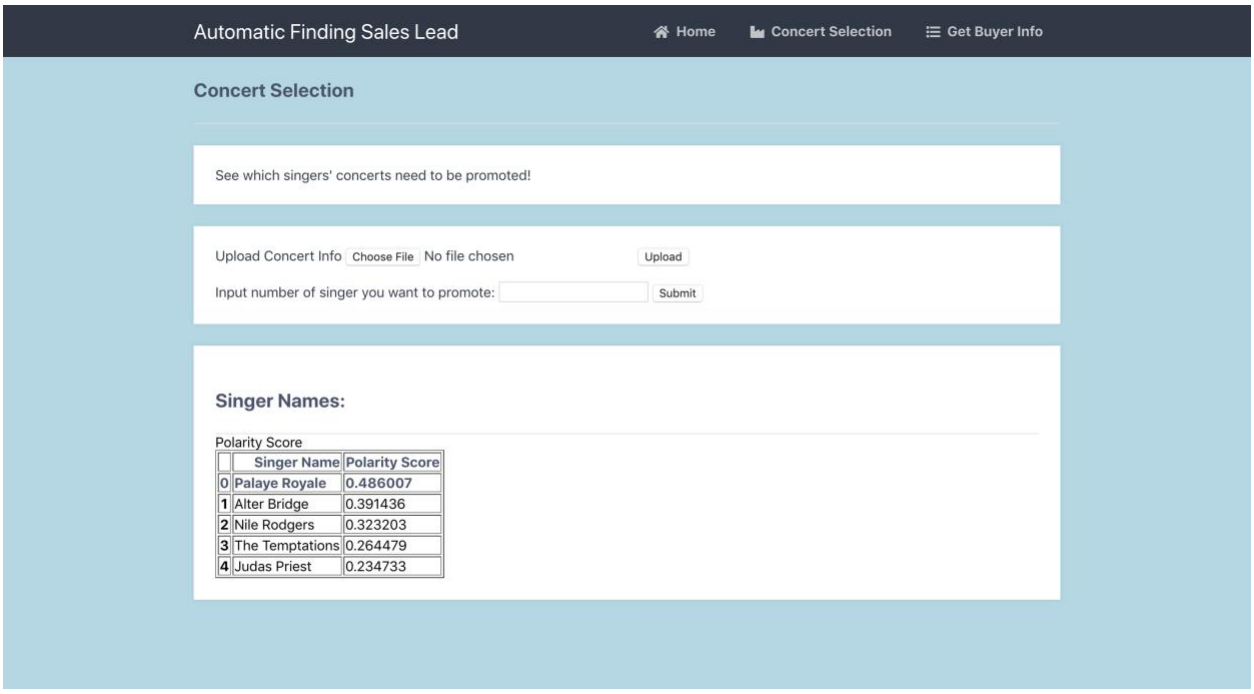


Figure 7.2

Figure 7.3 and 7.4 shows the "Get Buyer Info" page, which corresponds to the second phase of the system. It allows the user to input up to 5 performers at a time, and the system will return a table containing the performer's name and the sales leads' Twitter information.

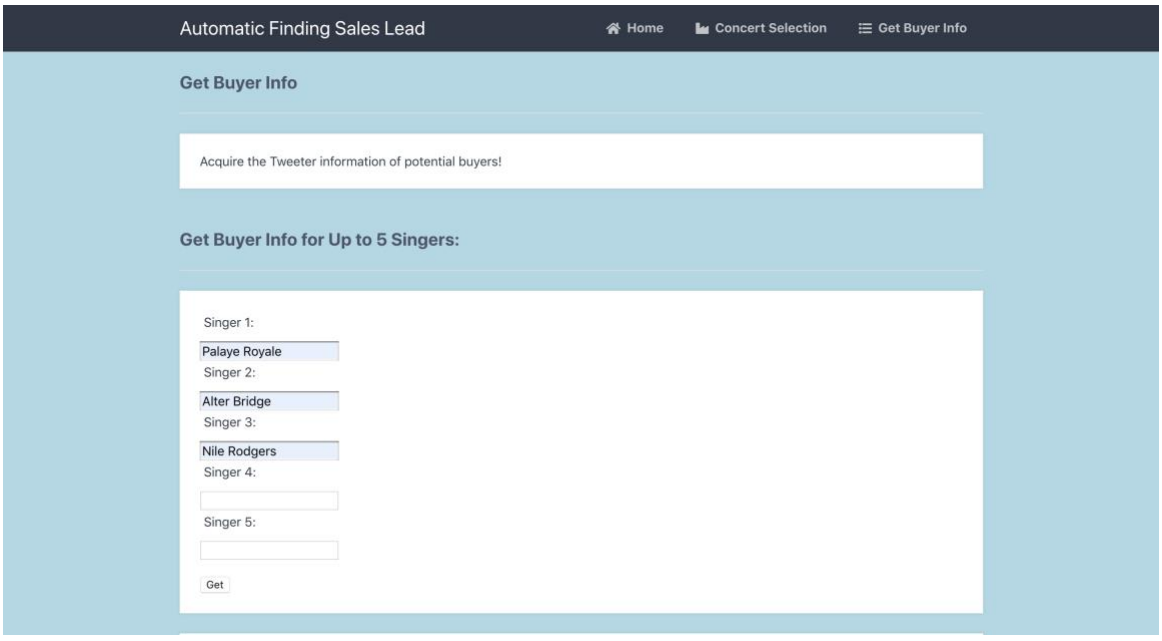
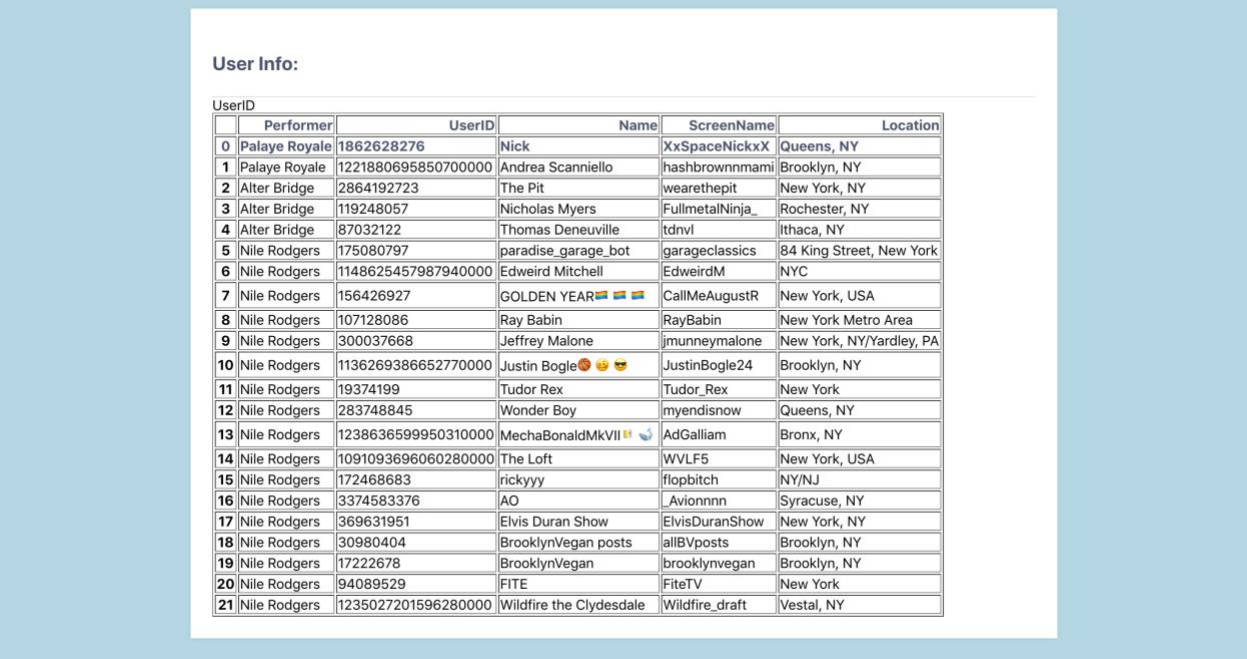


Figure 7.3



User Info:

UserID

	Performer	UserID	Name	ScreenName	Location
0	Palaye Royale	1862628276	Nick	XxSpaceNickxX	Queens, NY
1	Palaye Royale	1221880695850700000	Andrea Scanniello	hashbrownnmami	Brooklyn, NY
2	Alter Bridge	2864192723	The Pit	wearethepit	New York, NY
3	Alter Bridge	119248057	Nicholas Myers	FullmetalNinja_	Rochester, NY
4	Alter Bridge	87032122	Thomas Deneuville	tdnvl	Ithaca, NY
5	Nile Rodgers	175080797	paradise_garage_bot	garageclassics	84 King Street, New York
6	Nile Rodgers	1148625457987940000	Edweird Mitchell	EdweirdM	NYC
7	Nile Rodgers	156426927	GOLDEN YEAR	CallMeAugustR	New York, USA
8	Nile Rodgers	107128086	Ray Babin	RayBabin	New York Metro Area
9	Nile Rodgers	300037668	Jeffrey Malone	jmunneymalone	New York, NY/Yardley, PA
10	Nile Rodgers	1136269386652770000	Justin Bogle	JustinBogle24	Brooklyn, NY
11	Nile Rodgers	19374199	Tudor Rex	Tudor_Rex	New York
12	Nile Rodgers	283748845	Wonder Boy	myendisnow	Queens, NY
13	Nile Rodgers	1238636599950310000	MechaBonaldMkVII	AdGalliam	Bronx, NY
14	Nile Rodgers	1091093696060280000	The Loft	WVLF5	New York, USA
15	Nile Rodgers	172468683	rickyyy	flopbitch	NY/NJ
16	Nile Rodgers	3374583376	AO	_Avionnnn	Syracuse, NY
17	Nile Rodgers	369631951	Elvis Duran Show	ElvisDuranShow	New York, NY
18	Nile Rodgers	30980404	BrooklynVegan posts	allBVposts	Brooklyn, NY
19	Nile Rodgers	17222678	BrooklynVegan	brooklynvegan	Brooklyn, NY
20	Nile Rodgers	94089529	FITE	FiteTV	New York
21	Nile Rodgers	1235027201596280000	Wildfire the Clydesdale	Wildfire_draft	Vestal, NY

Figure 7.4

Conclusion

In this project, I wanted to create a system that helps ticket sellers automatically select concerts that need to be promoted, and then automatically find potential sales leads for those concerts. Throughout this process, I found out that search query data and social media data are actually very popular and useful tools to grasp public interest and the population's attitude. It is widely used not only in the US but also in China. When I was looking for related journal articles on Baidu, I found out that a lot of Chinese platforms such as Baidu and Sougou all have their own search engine index system. Social media such as Weibo also use search query data to analyze public trends and opinions, and use it as a feature on their social media. It made me realize again how many information are available online and that if we use them properly, we are able to achieve significant results.

Even though my system only focuses on promoting concert tickets, the idea behind it can be applied to any kind of event tickets or even general products because finding good sales leads does not only benefit ticket sellers; it is a crucial step for all product-oriented companies. Furthermore, the current system I have right now still has a lot of room for improvement. The algorithm for phase 1 and phase 2 can be expanded into more complex forms with more data and more time. For example, if I can have access to tweets not just from the past week, I can potentially count the number of tweets a user sent related to a specific singer. A user that posted multiple tweets about a singer will be more likely to purchase the ticket than a person who has only posted once. With these additional criteria, the system will become more accurate and more helpful to the company.

Works Cited

Caputi, Theodore L., et al. “They’Re Heating up: Internet Search Query Trends Reveal Significant Public Interest in Heat-Not-Burn Tobacco Products.” *Plos One*, vol. 12, no. 10, 2017, doi:10.1371/journal.pone.0185735.

Coogan, S; Sui, Z; Raubenheimer, D. “Gluttony and Guilt: Monthly Trends in Internet Search Query Data Are Comparable with National-Level Energy Intake and Dieting Behavior.” 2018, doi:10.1057/s41599-017-0055-7.

Dapergolas, Gerry. “40 Statistics To Inform & Enhance B2B Lead Generation Strategies.” *Link to Our Blog*, 2017, blog.strategic-ic.co.uk/lead-generation-statistics.

HubSpot. “2020 Marketing Statistics, Trends & Data - The Ultimate List of Digital Marketing Stats.” *2020 Marketing Statistics, Trends & Data - The Ultimate List of Digital Marketing Stats*, www.hubspot.com/marketing-statistics.

J. Zeng, J. Che, C. Xing, and L. Zhang. "LeadsRobot: A Sales Leads Generation Robot Based on Big Data Analytics." In *International Conference on Big Data*, pp. 277-290. Springer, Cham, 2018.

J. Zhang, Y. Wang, Z. Yuan and Q. Jin, "Personalized real-time movie recommendation system: Practical prototype and evaluation" in *Tsinghua Science and Technology*, vol. 25, no. 2,

pp. 180-191, April 2020.