Group 02
Alberto Piva
Sara Vozzella
Venakta Naga Sai Krishna Abhinay Pochiraju

Song statistics.

The data set that we analyzed represents the songs taken from the website www.AZLyrics.com. Initially the work was carried out aimed at the arrangement of data, more precisely with each song on the site, the title, the art, the URL and the lyrics were identified. The text of each song is a sentence with differents operation of stemming and stop words removal. After these operations have been performed the data-set consists of 86890 songs, which were subsequently uploaded to the MongoDB cloud.

Some simple statistical analyzes were carried out, the results of which are shown below:

- "Artist with most songs"

To evaluate the number of songs associated with each singer, a dictionary with the name of the singer as primary key and as attribute the total number of songs belonging to it.

Given the high number of singers present in the Data-set, only singers with more than 200 song are shown here. (exactly 41 singers).

The first singer turns out to be David Bowie with 211 songs, followed by Lil Wayne and Eminem that are in second place with exactly 210 songs each.

Consistent with what was expected from the analyzes conducted in this ranking we find internationally renowned artists.

The genres of music in this ranking are among the most varied, making us conclude that there are no particular correlations between the genre of music and the number of songs for each singer.

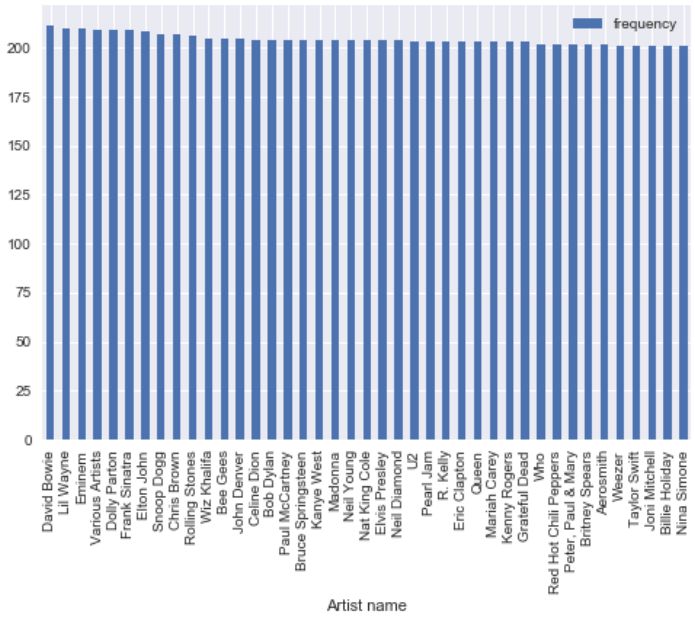
On the other hand, it is interesting to observe the historical period of the singers which appears to be for the most part contemporary, unfortunately given the lack of data concerning the production years of each single piece present in the data-set, this correlation has not been verified but simply hypothesized.

Below is the ranking of singers:

	ARTIST	CANZONI
0	David Bowie	211
1	Lil Wayne	210
2	Eminem	210
3	Various Artists	209
4	Dolly Parton	209
5	Frank Sinatra	209
6	Elton John	208
7	Snoop Dogg	207
8	Chris Brown	207
9	Rolling Stones	206
10	Wiz Khalifa	205
11	Bee Gees	205
12	John Denver	205
13	Celine Dion	204

14	Bob Dylan	204
15	Paul McCartney	204
16	Bruce Springsteen	204
17	Kanye West	204
18	Madonna	204
19	Neil Young	204
20	Nat King Cole	204
21	Elvis Presley	204
22	Neil Diamond	204
23	U2	203
24	Pearl Jam	203
25	R. Kelly	203
26	Eric Clapton	203
27	Queen	203
28	Mariah Carey	203
29	Kenny Rogers	203
30	Grateful Dead	203
31	Who	202
32	Red Hot Chili Peppers	202
33	Peter, Paul & Mary	202
34	Britney Spears	202
35	Aerosmith	202
36	Weezer	201
37	Taylor Swift	201
38	Joni Mitchell	201
39	Billie Holiday	201
40	Nina Simone	201

Through a histogram we observe the absolute frequencies of the quantity of pieces for each artist:



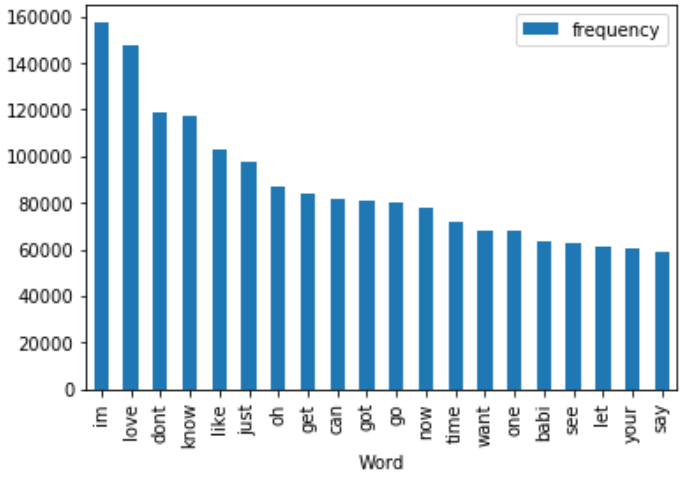
Group 02
Alberto Piva
Sara Vozzella
Venakta Naga Sai Krishna Abhinay Pochiraju

- "Identify the 20 most popular words"

For this point of the analysis were examined the absolute frequencies with which the words are repeated within all the texts of the songs present in the data-set.
The 20 most used words are shown below with their respective absolute frequency:

	WORD	FREQUENCY
0	im	157229
1	love	147335
2	dont	118717
3	know	117402
4	like	102541
5	just	97916
6	oh	87291
7	get	83594
8	can	81491
9	got	81100
10	go	80482
11	now	77688
12	time	71623
13	want	68145
14	one	67815
15	babi	63142
16	see	62995
17	let	61329
18	your	60301
19	say	58775

The most used words refer mostly to abstract concepts: nouns and adverbs such as 'Love', 'Like' are used to describe mostly emotions. As we expected, the songs are generally born to express and transmit emotions.
It is also significant to observe the absolute frequency with which these words occur, which reaches very high values such as for example in the case of 'Love' which reaches a frequency of 147335, which is equal to 1.4% of the total words used (exactly 10582208).
The histogram of the absolute frequencies is shown here above:



Group 02
Alberto Piva
Sara Vozzella
Venakta Naga Sai Krishna Abhinay Pochiraju

- "Identify the 10 most common singer names"

This analysis has been carried out considering all the names of the singers, observing the frequency with which they are repeated.
More in detail a list was created with all the names of the artists present in the data set, each name has been split in turn - so that both the name and surname of each individual artist is analyzed - and finally the stop-words have been eliminated to prevent words such as articles from appearing in the final ranking.
The following are the 10 most common names with the respective frequency.

	NAME	FREQUENCY
0	smith	7
1	james	7
2	williams	7
3	cole	7
4	john	6
5	tom	6
6	young	6
7	queen	6
8	david	6
9	perry	5

Subsequently, only singers which have at least one of the above words at least once within their name were analyzed, exactly 74 singers. Of these singers the number of songs written by each was observed.

	ARTIST	SONGS
0	Elliott Smith	122
1	Will Smith	120
2	Patti Smith	125
3	Michael W. Smith	199
4	Willow Smith	12
5	Smiths	77
6	Aerosmith	202
7	Jaden Smith	10
8	Sam Smith	24
9	James Newton Howard	2
10	James	138
11	James Taylor	200
12	Etta James	107
13	Luke James	11
14	James Blunt	12
15	Eddie James	1
16	Robbie Williams	200
17	Vanessa Williams	104
18	Hank Williams	198
19	Hank Williams Jr.	200
20	Andy Williams	153
21	Keller Williams	22

Group 02
Alberto Piva
Sara Vozzella
Venakta Naga Sai Krishna Abhinay Pochiraju

22	Pharrell Williams	41
23	J Cole	22
24	Lloyd Cole	105
25	Natalie Cole	176
26	Nat King Cole	204
27	Cole Porter	25
28	Nicole Scherzinger	14
29	Cheryl Cole	10
30	Keyshia Cole	16
31	Johnny Cash	200
32	John Denver	205
33	Daryl Hall & John Oates	200
34	John Mellencamp	200
35	Johnny Mathis	110
36	John Legend	102
37	John Waite	96
38	Elton John	208
39	Johnny Nash	10
40	Olivia Newton-John	199
41	Johnny Kid And The Pirates	1
42	Tom T. Hall	200
43	Tom Waits	200
44	Tom Jones	200
45	Tom Lehrer	47
46	Tom Petty & The Heartbreakers	200
47	Chris Tomlin	67
48	Tom DeLonge	1
49	Young Jeezy	124
50	Neil Young	204
51	Young Heretics	10
52	Fine Young Cannibals	17
53	Young Stunners	4
54	Hillsong Young & Free	2
55	Queen	203
56	Queens Of The Stone Age	88
57	Queen Adreena	44
58	Queen Latifah	64
59	Queensryche	110
60	Queen Of The Damned	15
61	Queen Ifrica	3
62	Queen & David Bowie	1
63	David Bowie	211
64	David Crowder Band	100
65	David Archuleta	11
66	David Pomeranz	21
67	David Guetta	82
68	Queen & David Bowie	1
69	Katy Perry	109
70	Perry Como	200

71	Joe Perry	11
72	Steve Perry	21
73	The Band Perry	12

Analyzing these last results there does not seem to be a significant correlation between the frequency with which the artist's name is presented and the number of songs written.

- "Create a histogram of song lengths"

Finally an analysis concerning the lyrics of the songs was conducted. The purpose of this study was to analyze the length of each song, calculated as the number of words present in it, considering this time also the stop-words.
The first 20 songs with their respective length are shown below.

	TITLE	ARTIST	SONG
0	Hazard	Richard Marx	9266
1	Faithful	Hillsong United	4851
2	These City Streets	Freestyle	4848
3	Bring It Back Home	Moe.	4507
4	Waiting For The Day	George Michael	3505
5	Hell Frezes Over (Album)	Eagles	3253
6	Secrets	Toni Braxton	3215
7	Racks On Racks	Wiz Khalifa	3024
8	Middle Of Nowhere	Hanson	2974
9	So Far So Good	Bryan Adams	2943
10	Alice's Restaurant	Arlo Guthrie	2483
11	Into The Woods - Prologue Act I	The Broadways	2457
12	Billy The Mountain	Frank Zappa	2396
13	(I've Got My) Future On Ice	Hank Williams Jr.	2190
14	Amsterdam	Yoko Ono	2168
15	Narration	W.a.s.p.	2136
16	Interlude 5	Lauryn Hill	2052
17	Mission From 'Arry	Iron Maiden	1952
18	The Murder Mystery	Velvet Underground	1950
19	Albuquerque	Weird Al Yankovic	1805

In this case the correlation between the length of the song and the genre is interesting: most of the songs listed above can be placed in the genre Hip-Hop / Rap. This phenomenon is explained by the fact that these genres are characterized by a greater speed of speech than the other genres, consequently this allows the singer to enrich the lyrics with more words at the same duration of the song itself.

The histogram of the absolute frequencies of the words in the first twenty songs of the above list is shown.

Homework – 3

Group 02
Alberto Piva
Sara Vozzella
Venakta Naga Sai Krishna Abhinay Pochiraju