ADM – Homework3

Davide Aureli(1651051)
Vigèr Durand Azimedem Tsafack(1792126)
Silvio Trampetti(1647497)

✓ **First Step**: Data collection

1. Downloading of the web pages of the songs from page https://www.azlyrics.com and storing them on the local disk: here we fixed the max number of songs to download at 31000, find the python code in the file *collect.py*

2. Parsing the downloaded pages and extracting for each the *lyric*, *artist*, *title*, and the *url* of the song: Find the python code in the file *parsing.py*

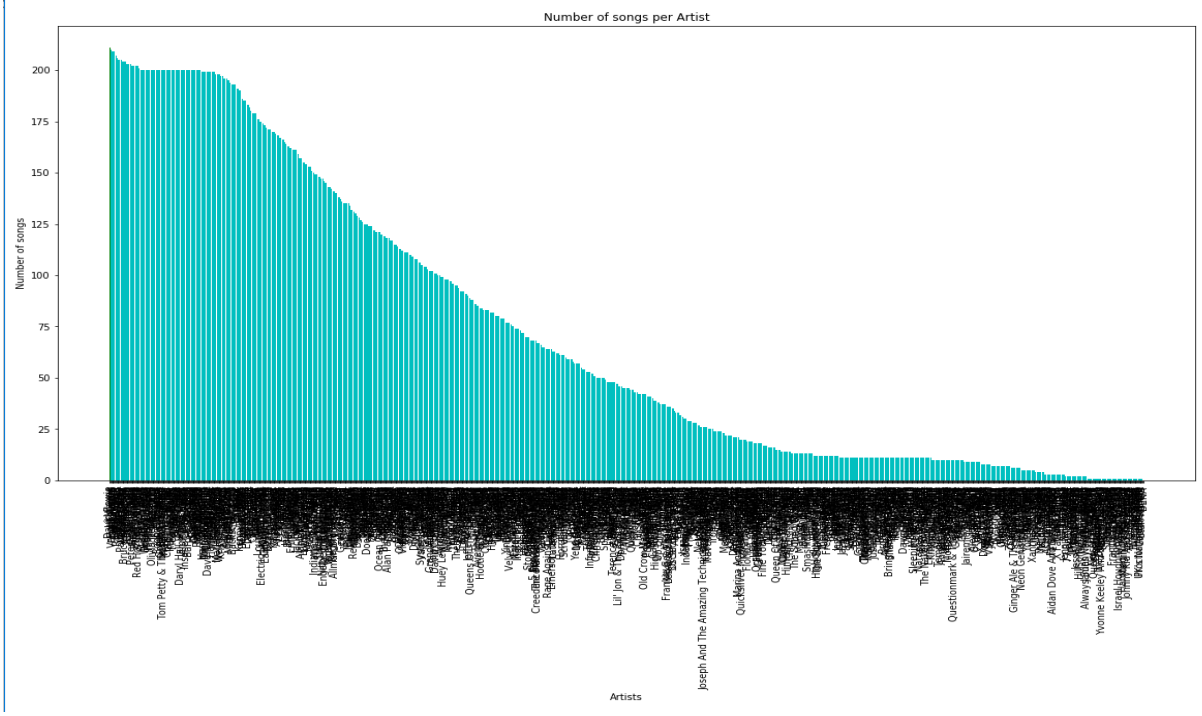
3. Store the parsed songs as documents in a MongoDB database, one document per song: Find the python code in the file *parsing.py*

```
#inserting all the songs in the mongoDB
db.lyrics.insert_many(DocList)
```

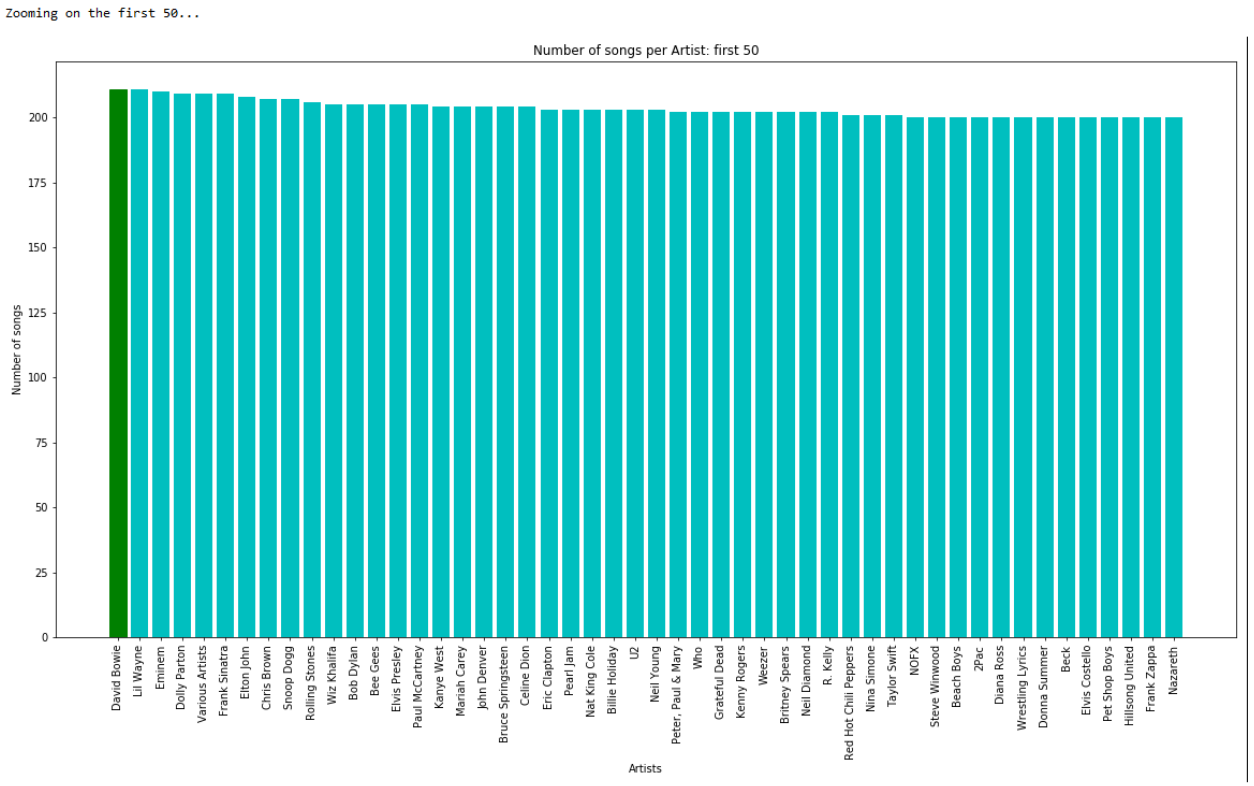
✓ **Second Step**: Song statistics

1. Identifying the artist with most songs and creating a histogram of the number of songs per Artist: The artist with most songs is Lil Wayne or David Bowie with 211 songs. Find the python code in the file *Song statistics.py*

```
Identifying the artist with most songs...
The artist with most song is:  David Bowie  with  211  songs...
Generating the bar graph of the number of songs per Artist...
```



Let’s make things more visible zooming on the first and then on the last 50 elements of the histogram:



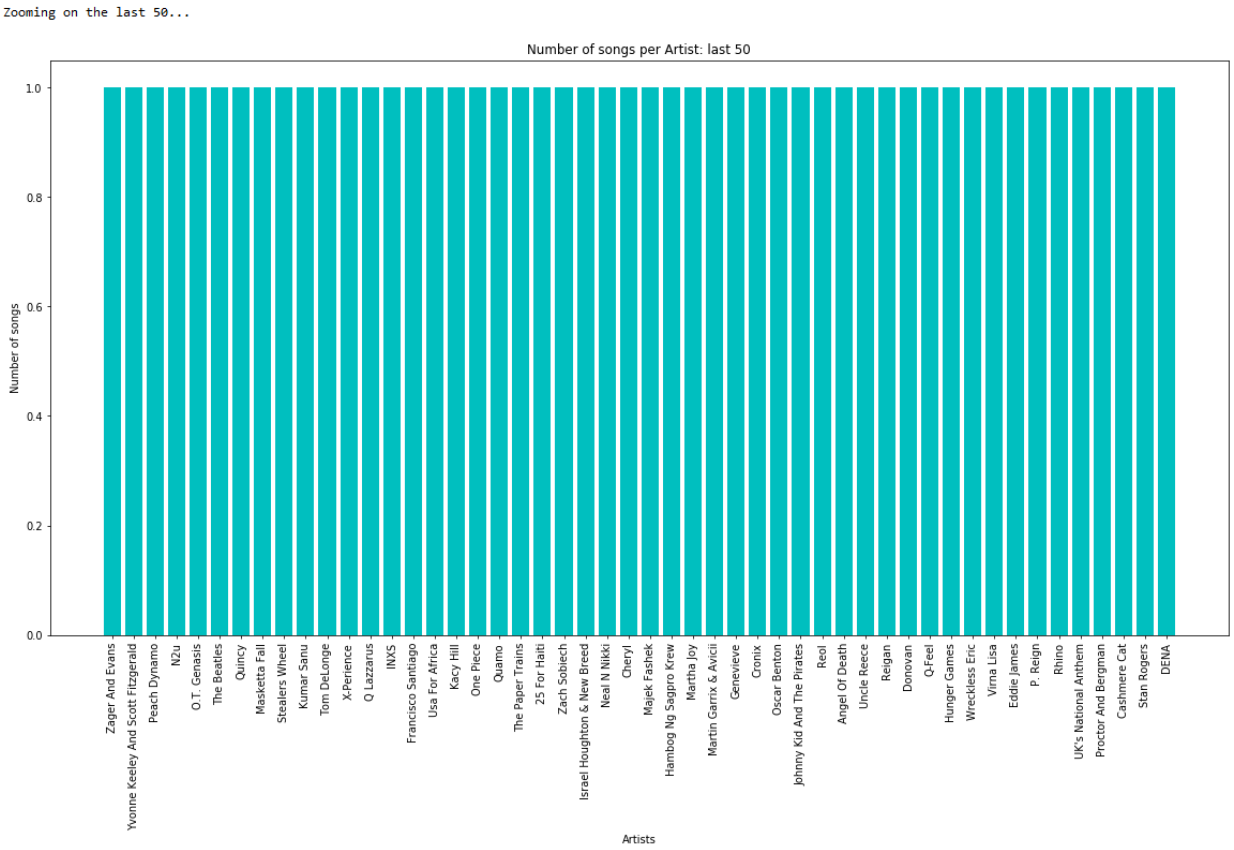
Looking at this histogram we could make some reasonings, let’s start making some research on Wikipedia about the first 15 artists with most songs.

	ARTIST	YEAR OF ACTIVITY
1	Lil Wayne	1991 - Present
2	David Bowie	1962 - 2016
3	Eminem	1992 - Present
4	Frank Sinatra	1932 - 1995
5	Dolly Parton	1959 - Present
6	Elton John	1964 - Present
7	Chris Brown	2005 - Present
8	Snoop Dogg	1992 - Present
9	Rolling Stones	1962 – Present
10	Bob Dylan	1959 – Present
11	Paul McCartney	1957 – Present
12	Elvis Presley	1953 – 1977
13	Wiz Khalifa	2004 - Present
14	Bee Gees	1960 – 2003 / 2009 - 2012
15	Mariah Carey	1990 - Present

The two first artists are Lil Wayne and David Bowie with 211 songs. Moreover, we notice that 11 over 15 are still active today, while who has finished to sing is not alive anymore.

We also notice that most of them are Rap singer. Their success coincides with the spread of Internet and new technologies, which allow everyone to listen their last hit.

It's also evident from the table above that artist with most songs also have a long career. In fact, the average of career years for these 15 first artists is about 41.5.



Let's now do some additional research about the last 50 artists.

Did they have a shorter career than the first 50?

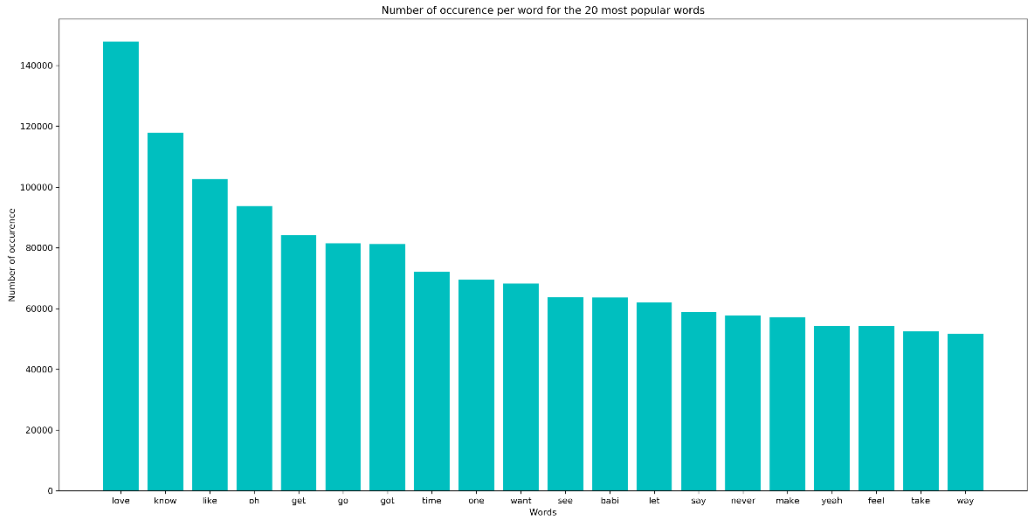
ARTIST	YEAR OF ACTIVITY
Oscar Benton	1960 - 1980
Eddie James	Not Available
Tom DeLonge	1992 - Present
Kacy Hill	2014 - Present
North of Nine	2015 – Present
Maren Morris	2005 - Present
The Beatles	1960 - 1970
Josh Kaufman	2010 - Present

The histogram shows that artists with less number of songs could be divided in 2 categories:

- **a-** Artists with a success based on few songs
- **b-** Newcomer artists.

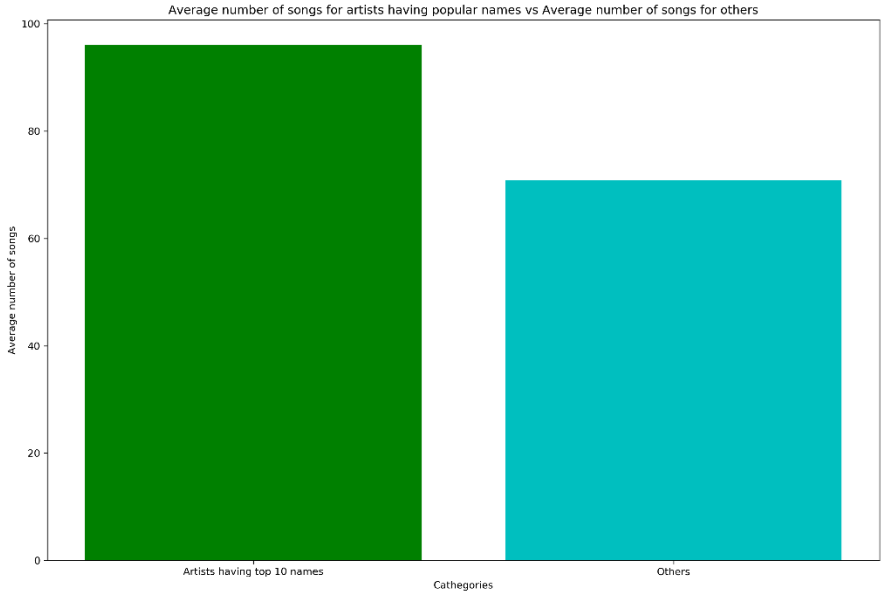
Extra research's online show that these artists are neither newcomer artists nor artists with success based on few songs. This is because the website only provides lyrics for their popular songs.

2. Identifying the 20 most popular words (exclude stopwords): find the python code in the file *Song statistics.py*.



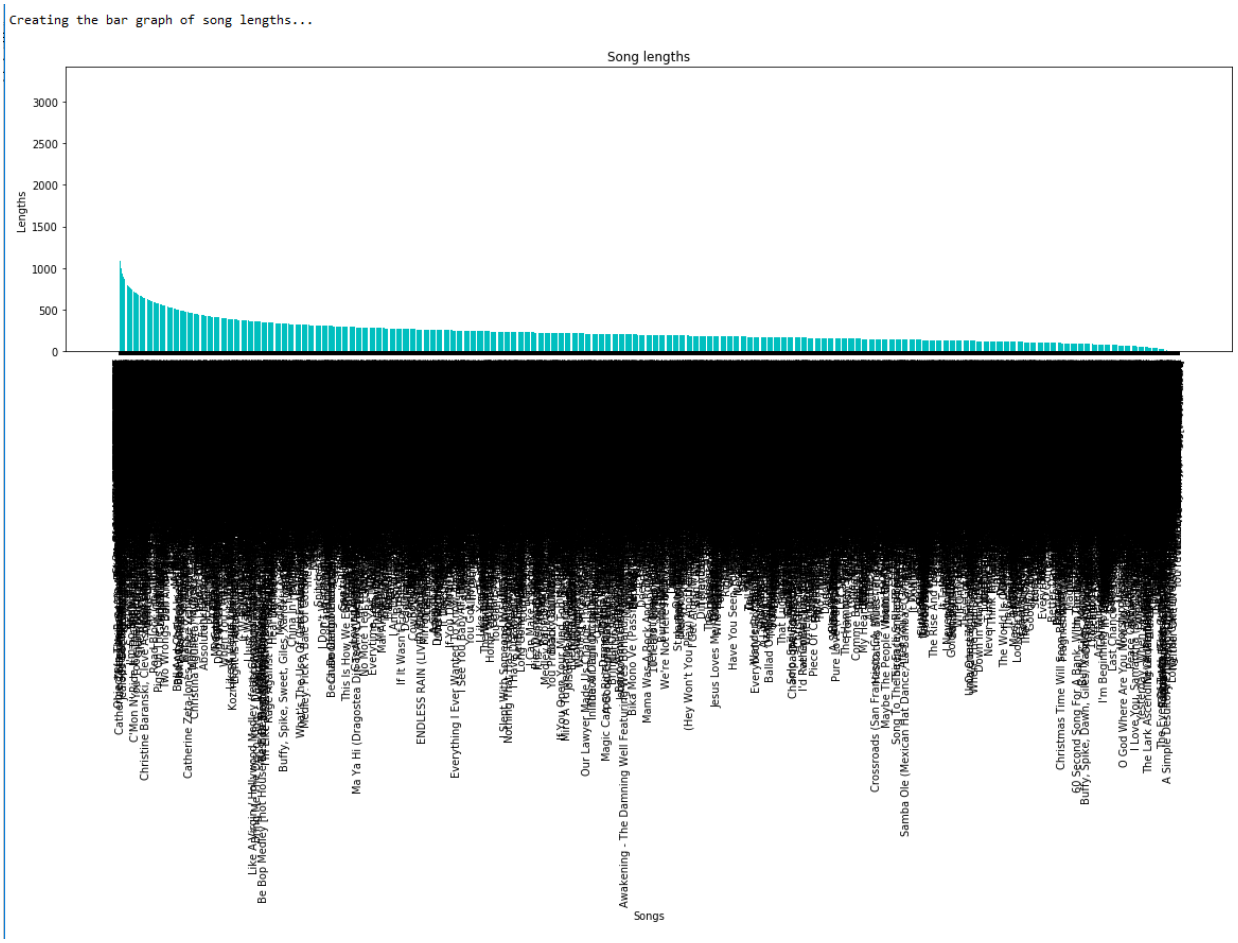
This is the bar chart of number of occurrence per word for the 20 most popular words. They seem to be almost all verbs. In particular, they mostly refer to human senses *("love", " like", " feel", "say", "see", "make", " take"*) , using this kind of words, the artists want to involve the audience both with ears and senses.
But we also have words like *"time" e "never"* which artists use to represent sadness or other negative feelings.

3. Identifying the 10 most common singer names and visualizing whether singers having popular name tend to publish more songs than others: find the python code in the file *Song statistics.py*. follows a histogram of average number of songs for artists having popular name vs average number of songs for others.

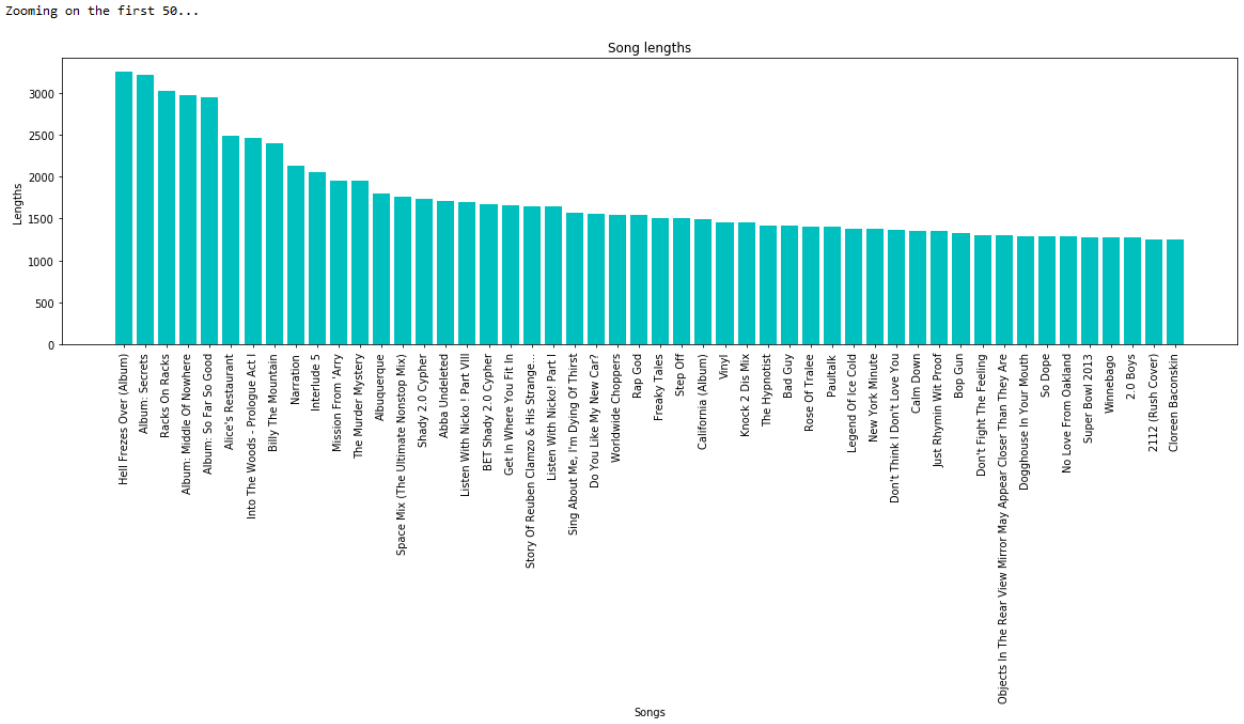


This histogram clearly shows that the artist with a popular name tend to publish more songs than the others.

4. Creating a histogram of song lengths: find the python code in the file *Song statistics.py*.



Let's make things more visible zooming on the first and then on the last 50 elements of the histogram:



✓ **Third Step**: Search engine

1. Inverted Index implementation: find the python code in the file *index.py*. here we create the inverted index as described in homework specifications and we upload it in the mongo lab database.

2. Inverted index *tf-idf* implementation: find the code in the file *indexTf-Idf.py*. here we simply replace the *tf* of the original inverted index by the tf-idf. The structure of this version of the inverted index is the following:

{term_id_1:[(document1, tf_idf{term,document1}), (document2, tf_idf{term,document2}),]}
{term_id_2:[(document1, tf_idf{term,document1}), (document2, tf_idf{term,document2}),]}

$$tf.idf = tf * idf$$

$$where$$

$$idf(t, D) = log\ \frac{N}{|\{d \in D : t \in d\}|}$$

We then upload this new inverted index in the mongo lab database because we are going to use it while searching.

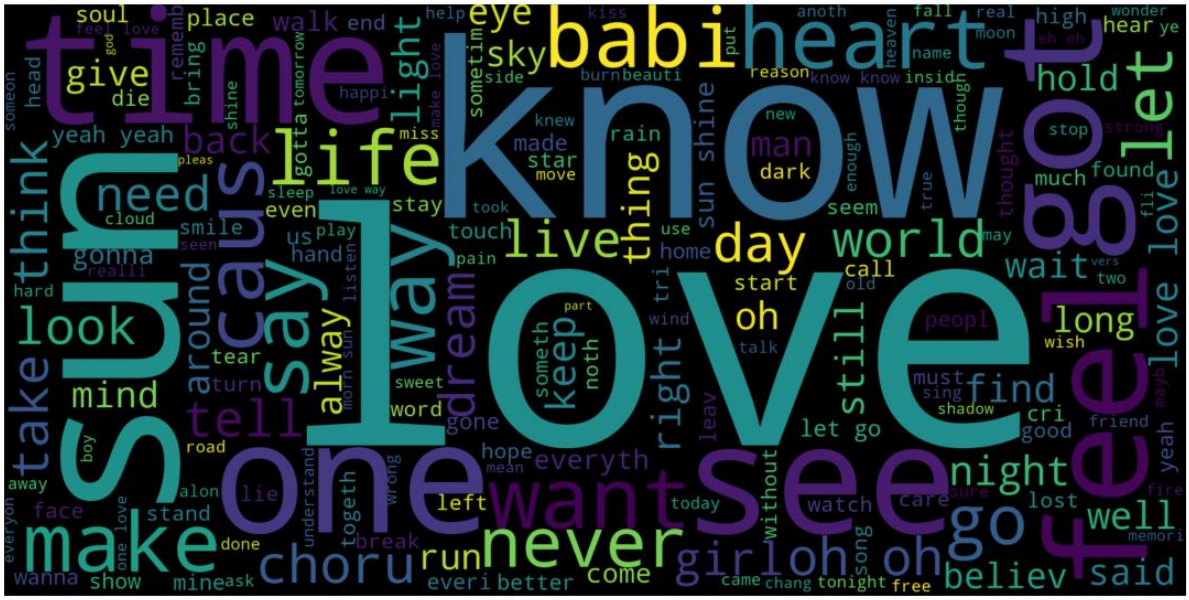
3. norms computation: find the python code in the file *norms.py*. In this file, we compute the norms of all the documents in the data set and upload them as a new collection into the database because they are going to turn useful during the computation of the cosine similarity between the documents and the set of words entered by the user during the searching phase.
Here the norm of a document is going to be the sum of the squares of *tf-idf* of all the terms inside the document all under square root.

$$norm(d) = \sqrt{\sum_{i=1}^{n} (tf.idf(t_i,\ d))^2}$$

4. Search engine implementation: find the code in the file *Search.py*. follows a test case:

```
collecting data from the database...
this operation could take a while...

What are you looking for?: love sun

what type of query do you need to perform? (1/2): 2

Number of matching songs found:  3340

insert a value k for clustering (0 < k <= 3340): 3
converting songs to vectors...
normalising vectors...
start clustering...
-->> Cluster  1 :

WordCloud of the most common words in the cluster:
```

Wordcloud of the most common words in the cluster

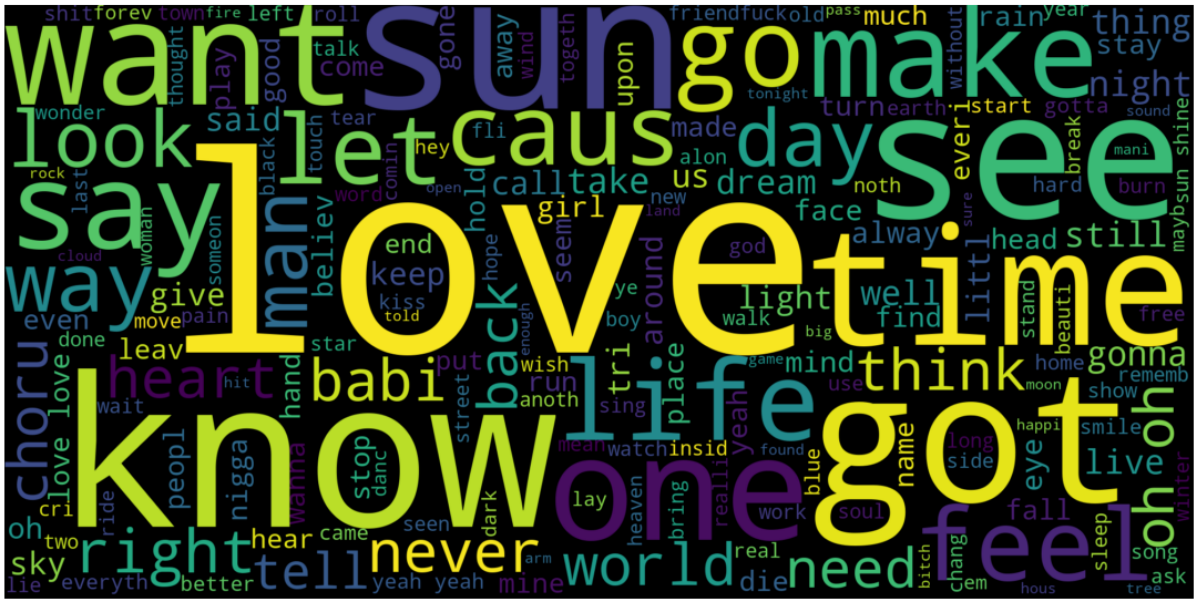


Artists and titles:
Matt Redman - 10,000 Reasons (Bless The Lord)
Matt Redman - 10,000 Reasons
Insane Clown Posse - 12
Elvis Costello - 15 Petals
Patti Smith - 1959
America - 1960
Alanis Morissette - 1974
Yg - 1Am
Snoop Dogg - 20 Dollars To My Name
Yg - 459
Prince - 5 Women
Chicago - 90 Degrees And Freezing
Out Of Eden - A Friend
Alphaville - A Handful Of Darkness
Gordon Lightfoot - A Lesson In Love
Lea Salonga - A Long, Long Time Ago
Soundtracks - A Love Before Time - CoCo Lee
Celine Dion - A Love For Me
Dan Fogelberg - A Love Like This
Michael Bolton - A Love So Beautiful
Roy Orbison - A Love So Beautiful
Soundtracks - A Man For All Seasons - Robbie Williams
Robbie Williams - A Man For All Seasons
Hanson - A Minute Without You
Celine Dion - A New Day Has Come
Various Artists - A New Day Has Come
Roy Orbison - A New Star
Barbra Streisand - A Sleepin' Bee
Hank Williams - A Stranger In The Night
Various Artists - A Thousand Years - Sting And Mariza
Sting - A Thousand Years
Deep Purple - A Twist In The Tail
Celine Dion - A World To Believe In
The Monkees - Acapulco Sun
Warren Zevon - Accidentally Like A Martyr
Tom Petty & The Heartbreakers - Accused Of Love
Vince Gill - Ace Up Your Pretty Sleeve
Beatles - Across The Universe
Soundtracks - Across The Universe
Cyndi Lauper - Across The Universe
David Bowie - Across The Universe
Fiona Apple - Across The Universe
10cc - Across The Universe
6 Cycle Mind - Across The Universe
Tina Turner - Addicted To Love
Linda Ronstadt - Adieu False Heart

-->> Cluster 2 :

WordCloud of the most common words in the cluster:

Wordcloud of the most common words in the cluster

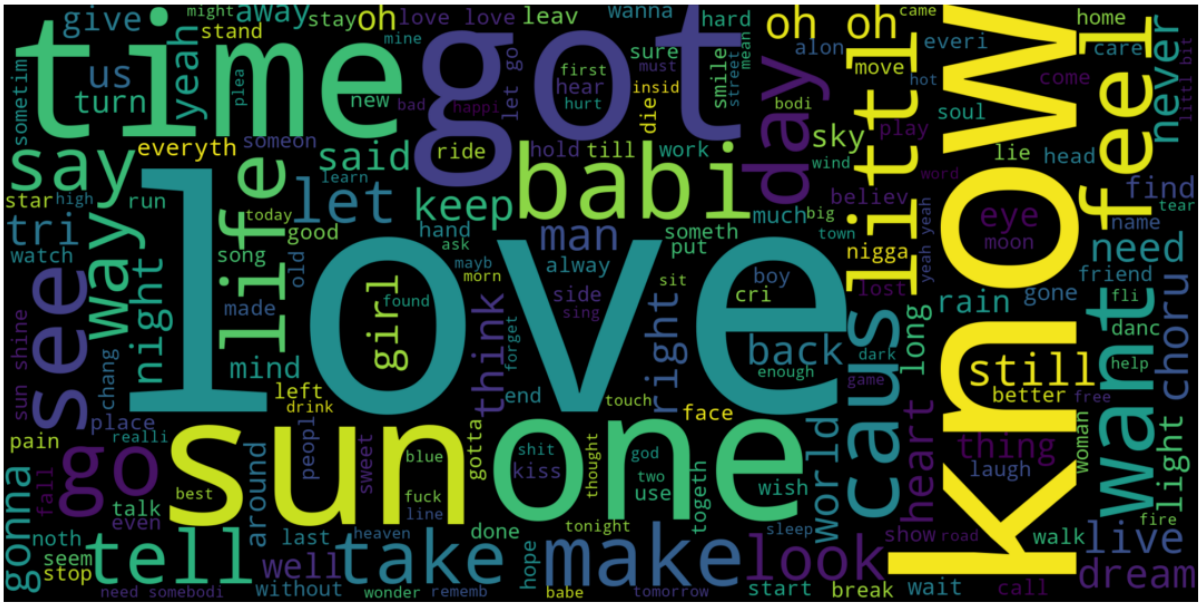


Artists and titles:
Gordon Lightfoot - 10 Degrees & Getting Colder
ASAP Rocky - 1Train
Robin Thicke - 2 The Sky
Bob Seger - 20 Years From Now
Prince - 4 The Tears In Your Eyes
Tori Amos - 500 Miles
Prince - A 1,000 Hugs And Kisses
Dream Theater - A Change Of Seasons
America - A Horse With No Name
Jimmy Buffett - A Mile High In Denver
Lenny Kravitz - A Million Miles Away
Jethro Tull - A Passion Play
Carly Simon - A Red, Red Rose
Leonard Cohen - A Singer Must Die
RJD2 - A Son's Cycle
Robin Trower - A Tale Untold
Doris Day - A Woman's Touch
Savage Garden - Affirmation
Dolly Parton - Afraid To Love Again
Stevie Wonder - Ai No, Sono
Snoop Dogg - Ain't No Fun
Diana Ross - Ain't No Mountain High Enough
Underworld - Air Towel
Unknown - Aire
Christy Moore - Aisling
Unknown - Alaska
Fun. - All Alone
Fun. - All Alright
Kenny Loggins - All I Ask
R. Kelly - All I Really Want
Kid Rock - All Summer Long
Donna Summer - All Systems Go
Fergie - All That I Got (The Make Up Song)

-->> Cluster 3 :

WordCloud of the most common words in the cluster:

Wordcloud of the most common words in the cluster



Artists and titles:
Miley Cyrus - 1 Sun
50 Cent - 21 Questions
Eric Clapton - 32-20 Blues
Willow Smith - 5
Eminem - 8 Mile
Perry Como - A Garden In The Rain
Lea Salonga - A Hundred Million Miracles
911 - A Little Bit More
John Denver - A Little Further North
U2 - A Man And A Woman
Tom T. Hall - A Picture Of Your Mother
Pm Dawn - About Nothing
Bee Gees - Above And Beyond
The Weeknd - Adaptation
Prince - Adore
Ed Sheeran - Afire Love
Boney M. - African Moon
Chuck Berry - Aimlessly Drifting
Hanson - Album: Middle Of Nowhere
Ed Sheeran - Alfire Love
Xzibit - Alkaholic
Wet Wet Wet - All I Want
Joni Mitchell - All I Want
Gary Barlow - All That I've Given Away
Rolling Stones - Always Suffering
Gloria Estefan - Always Tomorrow
Bon Jovi - Always
Love - And More
Steve Winwood - Angel Of Mercy
Natalie Grant - Another Day
Bon Jovi - Any Other Day
Kinks - Apeman
The Monkees - Apples, Peaches, Bananas & Pears

Credits to group 7:

- ➔ *Collaboration on the data collection and statistics part.*
- ➔ *Collaboration on how to use MongoDB.*