Mining YouTube Data

This report contains the Mining YouTube Data using Python project as part of my data science task . There are two parts in this project: The first part - About YouTube is an introduction of YouTube, its history and some interesting statistics. The second part - Mining YouTube Data using Python contains the Python code to work with YouTube Data API in order to extract data and perform various analysis on YouTube videos.

Why YouTube?

With a huge collection of videos attracting billions of views from users each month, the resulting data generated by YouTube is enormous. Data like view count, like count, dislike count, user comments, etc are all valuable data that can be extracted and analyzed to uncover insights about user preferences and sentiment towards a particular video or a particular cause e.g. the Ice Bucket Challenge, sometimes called the ALS Ice Bucket Challenge that went viral on YouTube few years ago. It also presents valuable information to marketers in their decision-making process of promoting a particular product or service. A fun example would be a movie studio, having uploaded a new movie trailer on their YouTube channel and would like to know about viewers' response towards the upcoming movie. Statistics on the videos such as view count, like count and user comments can help marketers to gauge the market response to the videos and allocate their marketing budget accordingly. This is the primary motivation behind this project. In this project, I have extracted and analyzed some interesting statistics about popular videos from MBC MASR .

The Process

In this project, I have used YouTube Data API v3 to extract data about videos and retrieve their statistics such as number of views, likes, dislikes, comments, etc. Then I converted these statistics into pandas DataFrame for further analysis. Analysis performed included ranking the most popular videos by view count and like count, and analyzing view count by day of the week. In the last part of this project, I have also extracted the comments from a YouTube video and performed wordcloud visualization on the most popular words, and followed by sentiment analysis using NLTK (the Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language).

The notebook is contain two main parts :

1- By Keyword: which retrieves the parameter for the query term that you are interested, in this example we will search for "mbc masr".

2- By Channel: There are times when we want to specifically extract data for a particular YouTube Channel that we are interested to analyze.

In this project, I have used YouTube Data API v3 to extract data about videos and retrieve their statistics such as number of views, likes, dislikes, comments, etc. Then I converted these statistics into pandas DataFrame for further analysis. Analysis performed included ranking the most popular videos by view count and like count, and analyzing view count by day of the week. In the last part of this project, I have also extracted the comments from a YouTube video and performed wordcloud visualization on the most popular words, and followed by sentiment analysis using NLTK. The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language.

Note: To use the YouTube Data API, you will need a API key which you can obtain in the Google APIs Console. You will need to create a new project, and enable the YouTube Data API for the project. To simplify things, I have used Google API python client which is a python client to interact with the Google APIs in an easier way. In order to use the API, you have to build a resource object for that API.

Alternatively, you can also manually call the YouTube Data API to retrieve the data, by setting the query parameters in the API endpoint to perform search on query term that you are interested.

- After importing libraries we set API Endpoints .

- Set YouTube Search parameters: There are some parameters that you need to set:
  - set SERVICE_NAME as "youtube" and the VERSION as "v3
  - set DEVELOPER_KEY to the API key that you obtain from Google APIs Console. You will need to create a new project, and enable the YouTube Data API for the project
  - set QUERY parameter for the query term that you are interested, in this example we will search for "mbc masr"
  - set PART parameter to id,snippet (it specifies a comma-separated list of one or more search resource properties that the API response will include)
  - set MAXRESULTS parameter as 50 as limited by Google (it specifies the maximum number of items that should be returned in the result set)
  - set PAGETOKEN if you are extracting the next page videos
  - the default ORDER is by relevance, but it can be changed to "date, viewCount", "rating", "title", "videoCount".

```
In [5]: # set parameters
YOUTUBE_API_SERVICE_NAME = "youtube"
YOUTUBE_API_VERSION = "v3"
DEVELOPER_KEY = "AIzaSyB3vkxBSyKUnEtlEi7Gv-BzpM7oZMR-0Ig"
QUERY = "mbc+masr"
PART = "id,snippet"
MAXRESULTS = "50"
PAGETOKEN = ""
ORDER = "relevance" # default=relevance, can be "viewCount", "rating", "title", "videoCount"

url = "https://www.googleapis.com/{0}/{1}/search?&key={2}&q={3}&part={4}&maxResults={5}&pageToken={6}&order={7}"
.format(YOUTUBE_API_SERVICE_NAME, YOUTUBE_API_VERSION, DEVELOPER_KEY, QUERY, PART, MAXRESULTS, PAGETOKEN, ORDER)

search_response = requests.get(url).json()

search_response
```

```
Out[5]: {'kind': 'youtube#searchListResponse',
 'etag': '"Fznwjl6JEQdo1MGvHOGaz_YanRU/EVvnabDc3MRGuAVe6frPfOxxAFM"',
 'nextPageToken': 'CDIQAA',
 'regionCode': 'EG',
 'pageInfo': {'totalResults': 327042, 'resultsPerPage': 50},
 'items': [{'kind': 'youtube#searchResult',
   'etag': '"Fznwjl6JEQdo1MGvHOGaz_YanRU/-BlpZATLdoL18hpw4XHEghYV0hA"',
   'id': {'kind': 'youtube#channel', 'channelId': 'UCnFKsbAof9fRv614I4wJX_w'},
   'snippet': {'publishedAt': '2012-10-14T17:38:53.000Z',
    'channelId': 'UCnFKsbAof9fRv614I4wJX_w',
    'title': 'MBC مصر',
    'description': 'خرجت MBC مصر إلى النور في شهر نوفمبر/ تشرين الثاني من عام 2012 ، لتقدم تغطية متميزة للشأن المصري بكافة أبعاده....',
    'thumbnails': {'default': {'url': 'https://yt3.ggpht.com/-IyYrDIDdtpM/AAAAAAAAAAI/AAAAAAAAAAA/cWkAtG0ntQQ/s88-c-k-no-mo-rj-c0xffffff/photo.jpg'},
     'medium': {'url': 'https://yt3.ggpht.com/-IyYrDIDdtpM/AAAAAAAAAAI/AAAAAAAAAAA/cWkAtG0ntQQ/s240-c-k-no-mo-rj-c0xffffff/photo.jpg'},
     'high': {'url': 'https://yt3.ggpht.com/-IyYrDIDdtpM/AAAAAAAAAAI/AAAAAAAAAAA/cWkAtG0ntQQ/s800-c-k-no-mo-rj-c0xffffff/photo.jpg'}},
    'channelTitle': 'MBC مصر',
    'liveBroadcastContent': 'upcoming'}},
```

Google API Python Client

Another simpler method is to use Google API Python Client or `apiclient` which is a python client to interact with the Google APIs in an easier way. In order to use the API, we need to build a resource object for the API. We passed the service name, its version and our Developer(or API) key to build method of apiclient.discovery module.

```
In [6]: # arguments to be passed to build function
        YOUTUBE_API_SERVICE_NAME = "youtube"
        YOUTUBE_API_VERSION = "v3"
        DEVELOPER_KEY = "AIzaSyB3vkxBSyKUnEtlEi7Gv-BzpM7oZMR-0Ig"

        # create youtube resource object for interacting with API

        youtube = build(YOUTUBE_API_SERVICE_NAME, YOUTUBE_API_VERSION,
                        developerKey=DEVELOPER_KEY)
```

We will use search.list function and pass in the query parameters. YouTube API will return a collection of search results that match the query parameters specified in the API request. By default, a search result set identifies matching video, channel, and playlist resources, but you can also configure queries to only retrieve a specific type of resource. More info at https://developers.google.com/youtube/v3/docs/search/list

```
In [7]: # query parameters
        query = "mbc masr"
        part = "id,snippet"
        maxresults = "50"
        order = "relevance" # default=relevance, can be "viewCount", "rating", "title", "videoCount"
        channelid = ""
        pagetoken = ""

        # calling the search.list method to retrieve youtube search results
        search_response = youtube.search().list(q = query,
                                                part = part,
                                                maxResults = maxresults,).execute()
```

```
In [8]: # take a look at the JSON object returned by YouTube, it should be the same as calling the endpoint manually
        search_response
```


```
Out[8]: {'kind': 'youtube#searchListResponse',
         'etag': '"Fznwjl6JEQdo1MGvHOGaz_YanRU/6s7bpw1AzapiAm6zFoes7ozAVXA"',
         'nextPageToken': 'CDIQAA',
         'regionCode': 'EG',
         'pageInfo': {'totalResults': 327042, 'resultsPerPage': 50},
         'items': [{'kind': 'youtube#searchResult',
           'etag': '"Fznwjl6JEQdo1MGvHOGaz_YanRU/-BlpZATLdoL18hpw4XHEghYV0hA"',
           'id': {'kind': 'youtube#channel', 'channelId': 'UCnFKsbAof9fRv614I4wJX_w'},
           'snippet': {'publishedAt': '2012-10-14T17:38:53.000Z',
            'channelId': 'UCnFKsbAof9fRv614I4wJX_w',
            'title': 'MBC مصر',
            'description': 'مصر إلى النور في شهر نوفمبر/ تشرين الثاني من عام 2012 ، لتقدم تغطية متميزة للشأن المصري بكافة أبعاده MBC خرجت....',
            'thumbnails': {'default': {'url': 'https://yt3.ggpht.com/-IyYrDIDdtpM/AAAAAAAAAAI/AAAAAAAAAAA/cWkAtG0ntQQ/s88-c-k-no-mo-rj-c0xffffff/photo.jpg'},
             'medium': {'url': 'https://yt3.ggpht.com/-IyYrDIDdtpM/AAAAAAAAAAI/AAAAAAAAAAA/cWkAtG0ntQQ/s240-c-k-no-mo-rj-c0xffffff/photo.jpg'},
             'high': {'url': 'https://yt3.ggpht.com/-IyYrDIDdtpM/AAAAAAAAAAI/AAAAAAAAAAA/cWkAtG0ntQQ/s800-c-k-no-mo-rj-c0xffffff/photo.jpg'}},
            'channelTitle': 'MBC مصر',
            'liveBroadcastContent': 'upcoming'}},
```


extract the results from search response:

```
In [9]: results = search_response.get("items", [])

        results
```


```
Out[9]: [{'kind': 'youtube#searchResult',
          'etag': '"Fznwjl6JEQdo1MGvHOGaz_YanRU/-BlpZATLdoL18hpw4XHEghYV0hA"',
          'id': {'kind': 'youtube#channel', 'channelId': 'UCnFKsbAof9fRv614I4wJX_w'},
          'snippet': {'publishedAt': '2012-10-14T17:38:53.000Z',
           'channelId': 'UCnFKsbAof9fRv614I4wJX_w',
           'title': 'MBC مصر',
           'description': 'مصر إلى النور في شهر نوفمبر/ تشرين الثاني من عام 2012 ، لتقدم تغطية متميزة للشأن المصري بكافة أبعاده MBC خرجت....',
           'thumbnails': {'default': {'url': 'https://yt3.ggpht.com/-IyYrDIDdtpM/AAAAAAAAAAI/AAAAAAAAAAA/cWkAtG0ntQQ/s88-c-k-no-mo-rj-c0xffffff/photo.jpg'},
            'medium': {'url': 'https://yt3.ggpht.com/-IyYrDIDdtpM/AAAAAAAAAAI/AAAAAAAAAAA/cWkAtG0ntQQ/s240-c-k-no-mo-rj-c0xffffff/photo.jpg'},
            'high': {'url': 'https://yt3.ggpht.com/-IyYrDIDdtpM/AAAAAAAAAAI/AAAAAAAAAAA/cWkAtG0ntQQ/s800-c-k-no-mo-rj-c0xffffff/photo.jpg'}},
           'channelTitle': 'MBC مصر',
           'liveBroadcastContent': 'upcoming'}},
         {'kind': 'youtube#searchResult',
          'etag': '"Fznwjl6JEQdo1MGvHOGaz_YanRU/wOYaxvk-V3wc-erj2UYSqZQSFN0"',
          'id': {'kind': 'youtube#video', 'videoId': 'vhpq4F5B4HM'},
          'snippet': {'publishedAt': '2020-02-18T19:20:29.000Z',
           'channelId': 'UCnFKsbAof9fRv614I4wJX_w',
```

create 3 separate empty lists to store video, playlist and channel metadata:

```
In [10]:
# create 3 separate empty lists to store video, playlist and channel metadata
videos = []
playlists = []
channels = []

# extract the required info from each result object
for result in results:
    # video result object
    if result['id']['kind'] == "youtube#video":
        video = {}
        video['id'] = result["id"]["videoId"]
        video['title'] = result['snippet']['title']
        video['descr'] = result['snippet']['description']
        video['thumbnail'] = result['snippet']['thumbnails']['default']['url']
        videos.append(video)
    # playlist result object
    elif result['id']['kind'] == "youtube#playlist":
        playlist = {}
        playlist['id'] = result['id']['playlistId']
        playlist['title'] = result['snippet']['title']
        playlist['descr'] = result['snippet']['description']
        playlist['thumbnail'] = result['snippet']['thumbnails']['default']['url']
        playlists.append(playlist)
    # channel result object
    elif result['id']['kind'] == "youtube#channel":
        channel = {}
        channel['id'] = result['id']['channelId']
        channel['title'] = result['snippet']['title']
        channel['descr'] = result['snippet']['description']
        channel['thumbnail'] = result['snippet']['thumbnails']['default']['url']
        channels.append(channel)
```

```
In [11]:
# take a look at the videos list
print("There are {} videos in the result".format(len(videos)))
videos
```

The format of a single sample video will look like the result below, with the description, video id, thumbnail URL and title of the video:

```
In [12]:
# take a look at the format of a single video
videos[0]
```

```
Out[12]: {'id': 'vhpq4F5B4HM',
 'title': 'قصتي بكل تنزل في الحلقة الأخيرة بوس وك قصة',
 'descr': 'MBCMASR #MBCMASR2 #MBC #SHAHID #قصة_وك_بوس » Subscribe for more: http://onmbc.net/60521Yixo » Watch Full Episodes Fre
e: ...',
 'thumbnail': 'https://i.ytimg.com/vi/vhpq4F5B4HM/default.jpg'}
```

```
In [13]:
# take a look at the playlist list
print("There are {} playlists in the result".format(len(playlists)))
playlists
```

```
There are 4 playlists in the result
```

```
Out[13]: [{'id': 'PLNkaIgADZ9OBrMoItuKQK-sSIkKfBBII7',
  'title': 'Popular Videos - MBC Masr & Zeina',
  'descr': '',
  'thumbnail': 'https://i.ytimg.com/vi/HQqTharHvPc/default.jpg'},
 {'id': 'PLNkaIgADZ9OCHsvWHmTUoknXNIK4KvakP',
  'title': 'Popular Videos - MBC Masr & Game',
  'descr': '',
  'thumbnail': 'https://i.ytimg.com/vi/x8NW7TZwTek/default.jpg'},
 {'id': 'PLNkaIgADZ9OBIreMkI11UvYbCBJPCDPra',
  'title': 'Popular Videos - MBC Masr',
  'descr': '',
  'thumbnail': 'https://i.ytimg.com/vi/mg_nFaZsKZM/default.jpg'},
 {'id': 'PLNkaIgADZ9OBEpIlzSelAg3vm30S0ysa7',
  'title': 'Popular Videos - MBC Masr & Hassan Shakosh',
  'descr': '',
  'thumbnail': 'https://i.ytimg.com/vi/vOh7Ftz1j-M/default.jpg'}]
```

```
In [14]:
# take a look at the channel list
print("There are {} channel in the result".format(len(channels)))
channels
```

```
There are 3 channel in the result
```

```
Out[14]: [{'id': 'UCnFKsbAoF9fRv614I4wJX_w',
  'title': 'MBC مصر',
  'descr': 'مصر إلى النور في شهر نوفمبر/ تشرين الثاني من عام 2012 ، لتقدم تجربة متميزة للشأن المصري بكافة أبعاده MBC خرجت مصر....',
  'thumbnail': 'https://yt3.ggpht.com/-IyYrDIDdtpM/AAAAAAAAAAI/AAAAAAAAAAA/cwkAtG8ntQQ/s88-c-k-no-mo-rj-c0xffffff/photo.jpg'},
 {'id': 'UC1F59TWMyxxVKI-BsWaI5mA',
  'title': 'الممكلة',
  'descr': 'MBC MASR برنامج جديد بطل من حلقاته الإعلامي الكبير عمرو أديب على جمهوره من الجمعة إلى الاثنين من كل أسبو الممكلة"...',
  'thumbnail': 'https://yt3.ggpht.com/-0_7AhbVVxIA/AAAAAAAAAAI/AAAAAAAAAAA/OtLWlHECq9U/s88-c-k-no-mo-rj-c0xffffff/photo.jpg'},
 {'id': 'UCngKM1KHxBoSiKTekOBBamA',
  'title': 'MBC MASR 2',
  'descr': 'بعد النجاح الكبير الذي حققته قناة "MBCمصر"، وتصدرها المواقع الأولى في السوق المصري بعد عامين على إطلاقها...',
  'thumbnail': 'https://yt3.ggpht.com/-q-toUsFZmWM/AAAAAAAAAAI/AAAAAAAAAAA/eoZVax2F1yw/s88-c-k-no-mo-rj-c0xffffff/photo.jpg'}]
```

*Note: * There are 50 videos returned as results which contain the video id, thumbnail, title and description. For a quicker result and easy viewing, we can also extract the data into a dictionary format with video ID and title

Extract Video Statistics

The title and description of the videos often do not offer much information or insights, what is more useful is the statistics of those videos for further analysis. With YouTube Data API, we can get the statistics for the videos such as like count, dislike count, view count and so on. After populating the statistics into a dictionary, we can then convert the dictionary into a pandas DataFrame for further analysis and visualization. Also, by default, a search result set identifies matching video, channel, and playlist resources, but since we want to analyze the statistics for videos only, we will configure queries to only retrieve a specific type of resource

```
In [22]: # query parameters
query = "Elhekayashow"
part = "id,snippet"
maxresults = "50"
order = "relevance" # default=relevance, can be "viewCount", "rating", "title", "videoCount"
channelid = ""
pagetoken = ""
type = "video"

# calling the search.list method to retrieve youtube search results
search_response = youtube.search().list(q = query,
                                        part = part,
                                        maxResults = maxresults,
                                        order = order,
                                        pageToken = pagetoken,
                                        type = type).execute()
```

```
In [23]: # take a look at the JSON object returned by YouTube

search_response
```

```
Out[23]: {'kind': 'youtube#searchListResponse',
 'etag': '"Fznwjl6JEQdo1MGvHOGaz_YanRU/VV3TeFlNsbFr_D66d0LIdWgz-lw"',
 'nextPageToken': 'CDIQAA',
 'regionCode': 'EG',
```

convert the list of dictionary into a pandas DataFrame:

```
dataframe = pd.DataFrame.from_dict(res)
dataframe
```

	channelTitle	commentCount	dislikeCount	favoriteCount	likeCount	publishedAt	v_id	v_title	viewCount
0	الحكاية	27	20	0	242	2020-02-14	l1AAZ2ggB9E	...وزير الخارجية سامح شكرى يرد على سؤال عمرو أديب	24109
1	الحكاية	149	70	0	1190	2020-02-15	a24mt_S9TDw	... ..عمرو أديب: إحنا بنتكلم علزمالك على سن ورمح	54528
2	الحكاية	81	44	0	656	2020-02-15	EeWYNKnS8Fg	...عمرو دياب يرد على سؤال عمرو أديب: دينا الشربين	78052
3	الحكاية	47	24	0	251	2020-02-16	eLwL9pfEHc0	...حسن شاكوش: ورحمة أبويا مضيت تعهد في نقابة المو	23392
4	الحكاية	149	69	0	1184	2020-02-14	h2lHaAC2QWM	...عمرو أديب يشرح تفاصيل غسيل أموال بـ٢ مليار جني	107093
5	الحكاية	6	2	0	89	2020-02-17	cfFe1pdqRcQ	...عمرو أديب: الطيران البريطاني رجع شرب الشيخ عقب	4802
6	الحكاية	6	4	0	124	2020-02-16	yzqtOBZC2Hc	...حسن شاكوش: أنا كنت لعيب كرة وناس متعلمة مش متس	8283
7	الحكاية	109	55	0	620	2020-02-15	w9YlMexUjTo	... عمرو أديب: كوريا الشمالية ضربوا مسؤول بالرصاص	58971
8	الحكاية	17	8	0	133	2020-02-16	uROlTwGf4yM	...هو السبب في أزمة حذف لـ حسن شاكوش: الرجل بتاع ال	11305
9	الحكاية	4	2	0	70	2020-02-15	E87KGCkxmPw	.....عمرو دياب: أنا بقالي 35 سنة وشفت حاجات كتيرة	4441
10	الحكاية	16	1	0	62	2020-02-17	0D9OlFoUDxs	... رئيس لجنة السوبر المصرى بالإمارات يكشف تفاصيل	4894
11	الحكاية	12	5	0	52	2020-02-17	ahkntl00f7Y	...يتابع من مطروح إنهاء العزل الصحي للـ (الحكاية)	3735
12	الحكاية	72	15	0	212	2020-02-14	hAQbRa9VXEY	...عمرو أديب عن جرائم اغتصاب طارق رمضان: عنتيل قر	13501
13	الحكاية	5	12	0	113	2020-02-16	yVoZ87fJdSU	...حسن شاكوش: أنا راضي بأي قرار باخده القفان هاني	11729
14	الحكاية	9	10	0	74	2020-02-16	NBE__69AadY	...المنتجة سارة الطباخ تكشف لأول مرة تفاصيل الصلح	13547
15	الحكاية	29	12	0	121	2020-02-16	GaIXkDTvb28	...عمرو أديب يشرح تفاصيل مكالمته مع وزير الاتصالا	14082
16	الحكاية	17	3	0	97	2020-02-14	XRdf6XQQP-M	...عمرو أديب: أنا قلت الموضوع لو في فلوس وهناخد ٥	10457
17	الحكاية	33	15	0	302	2020-02-16	wTZc1SP8tOg	...حلمي بكر: دلوقتي عالكمبيوتر ممكن نعيب صوت حمار	26115
18	الحكاية	9	6	0	65	2020-02-14	e4FV3MsYJmo	...وزير الخارجية سامح شكرى يشرح تفاصيل وتطورات جو	5952

Google only allow us to extract maximum 50 results per call. What if we want more results to work with? Fortunately, Google allows us to query the second page using a Token and extract another 50 results. We can then combine the results into a single DataFrame with 100 results. You can continue to query the next page until 1,000,000 results. However, search results are constrained to a maximum of 500 videos if your request specifies a value for the channelId parameter and sets the type parameter value to video.

After we have extracted the statistics for the next page videos, we need to merge the two DataFrames into a single DataFrame for further analysis:

```
In [33]: # merge the next page result with the first page result
         df = pd.concat([dataframe, next_page_df]).reset_index(drop=True)
         df
```

Out[33]:

	channelTitle	commentCount	dislikeCount	favoriteCount	likeCount	publishedAt	v_id	v_title	viewCount
0	الحكاية	27	20	0	242	2020-02-14	l1AAZ2ggB9E	...وزير الخارجية سامح شكري يرد على سؤال عمرو أديب	24109
1	الحكاية	149	70	0	1190	2020-02-15	a24mt_S9TDw	... ...عمرو أديب: إحنا بنتكلم عالزمالك على سن ورمح	54528
2	الحكاية	81	44	0	656	2020-02-15	EeWYNKnS8Fg	...عمرو دياب يرد على سؤال عمرو أديب: دينا الشربين	78052
3	الحكاية	47	24	0	251	2020-02-16	eLwL9pfEHc0	...حسن شاكوش: ورحمة أبويا مضيت تعهد في نقابة المو	23392
4	الحكاية	149	69	0	1184	2020-02-14	h2lHaAC2QWM	...عمرو أديب بشرح تفاصيل غسيل أموال بـ٢ مليار جني	107093
5	الحكاية	6	2	0	89	2020-02-17	cfFe1pdqRoQ	...عمرو أديب: الطيران البريطاني رجع شرم الشيخ عقب	4802
6	الحكاية	6	4	0	124	2020-02-16	yzqtOBZC2Hc	...حسن شاكوش: أنا كنت لعيب كرة وناس متعلمة مش متس	8283
7	الحكاية	109	55	0	620	2020-02-15	w9YlMexUjTo	... عمرو أديب: كوريا الشمالية ضربوا مسؤول بالرصاص	58971
8	الحكاية	17	8	0	133	2020-02-16	uROlTwGf4yM	...هو السبب في أزمة حذف لـ حسن شاكوش: الرجل بتاع ال	11305
9	الحكاية	4	2	0	70	2020-02-15	E87KGCkxmPw	.....عمرو دياب: أنا بقالي 35 سنة وشفت حاجات كتيرة	4441
10	الحكاية	16	1	0	62	2020-02-17	0D9OlFoUDxs	... رئيس لجنة السوبر المصري بالإمارات يكشف تفاصيل	4894
11	الحكاية	12	5	0	52	2020-02-17	ahknti00f7Y	...يتابع من مطروح إنهاء العزل الصحي للم (الحكاية)	3735
12	الحكاية	72	15	0	212	2020-02-14	hAQbRa9VXEY	...عمرو أديب عن جرائم اغتصاب طارق رمضان: عنتيل قر	13501
13	الحكاية	5	12	0	113	2020-02-16	yVoZ87fJdSU	...حسن شاكوش: أنا راضي بأي قرار ياخده اللقان هاني	11729
14	الحكاية	9	10	0	74	2020-02-16	NBE__69AadY	...المتهمة سارة الطباخ تكشف لأول مرة تفاصيل الصلح	13547
15	الحكاية	29	12	0	121	2020-02-16	GalXkDTvb28	...عمرو أديب بشرح تفاصيل مكالمته مع وزير الاتصالا	14082
16	الحكاية	17	3	0	97	2020-02-14	XRdf6XQQP-M	...عمرو أديب: أنا قلت الموضوع لو في فلوس وهناخد ٥	10457
17	الحكاية	33	15	0	302	2020-02-16	wTZc1SP8tOg	...حلمي بكر: دلوقتي عالكمبيوتر ممكن نعيب صوت حمار	26115

Remove Duplicates

We can check for any duplicate values in the dataframe using function `.duplicated()` and remove them accordingly using function `drop_duplicates()`.

```
: print("There are {} duplicated values.".format(df.duplicated().sum()))
  df[df.duplicated(keep=False)].head(10)
```

```
There are 0 duplicated values.
```

:

channelTitle	commentCount	dislikeCount	favoriteCount	likeCount	publishedAt	v_id	v_title	viewCount

There are no duplicates in our DataFrame so we can further analyze. Note that sometimes we may encounter missing value for the column dislikeCount and likeCount, this is likely because the author has disabled showing the like and dislike counts, the same can also happen for the comments column.

Pre-Processing

From the `info()` function above, we can see that all of the columns of the DataFrame are in object format. We need to convert those column format into the correct format before we can perform analysis and visualization. For example, we need to convert the columns for "commentCount", "dislikeCount", "favoriteCount", "likeCount", "viewCount" into numeric data format. For "publishedAt" column, we need to convert it into datetime format. After that, we will sort the DataFrame by viewCount and likeCount for easy comprehension at a glance.

In [37]:
```
# convert string into numeric and datetime columns
# sort table by viewCount and likeCount
numeric_columns = ["commentCount", "dislikeCount", "favoriteCount", "likeCount", "viewCount"]
df[numeric_columns] = df[numeric_columns].apply(pd.to_numeric)
df["publishedAt"] = pd.to_datetime(df["publishedAt"])
df_final = df.sort_values(by=['viewCount', 'likeCount'], ascending=False).reset_index(drop=True)
df_final
```

Out[37]:

	channelTitle	commentCount	dislikeCount	favoriteCount	likeCount	publishedAt	v_id	v_title	viewCount
0	الحكاية	558	211	0	5254	2020-02-16	DJY0KgLgVa0	... أول مداخلة للبضعة عمرو دياب بعد ألبومه الجديد	262884
1	الحكاية	887	820	0	5274	2020-02-15	UsSEMxgSK_s	... عمرو أديب: وهو الرجل المطلق كان محروق أوى كده	256459
2	الحكاية	196	134	0	2496	2019-12-16	h4WuTw7lI6w	عمرو أديب يوجه رسالة للأهلاوية	241253
3	الحكاية	938	437	0	2912	2019-12-15	V5P6y3oZ4sk	عمرو أديب يوجه رسالة للرئيس السيسي	164201
4	الحكاية	140	137	0	2404	2020-02-14	Feh-8UjWJ2I	المفتي عبر كمال يكشف أجره بعد نجاح بنت الجيران	161457
5	الحكاية	111	35	0	883	2020-01-27	UaSampnakZc	زينة عاطفة الولاد بأبوهو كويسة جدا وتخصيهد هو	157596
6	الحكاية	82	65	0	1622	2020-02-16	T1mW_98wO_c	... تعليق حلمي بكر على مداخلة البضعة عمرو دياب في	135635
7	الحكاية	633	214	0	2132	2020-02-16	vKLbTQkUe6k	...هاني شاكر: استحالة أي مطرب مهرجانات يغني في أي	120011
8	الحكاية	159	71	0	1162	2020-02-10	HBtD7w48h7c	تعرف على تفاصيل حادث حدائق الأهرام الغامض	118767
9	الحكاية	50	21	0	582	2020-01-26	Ib3z1sQsbYY	زينة : قايلت ظلم كتير ماحدش يتحمله#	110510
10	الحكاية	149	69	0	1184	2020-02-14	h2lHaAC2QWM	...عمرو أديب يشرح تفاصيل غسيل أموال بـ٢ مليار جني	107093
11	الحكاية	214	52	0	1245	2020-01-26	q1xqwiTqVSU	يتابع أعمال التطوير بميدان التحرير (الحكاية)	82595
12	الحكاية	171	91	0	920	2020-02-15	QcltXdJlvVw	حلمي بكر: إحنا في عصر الجهل باسم المهرجانات	79980
13	الحكاية	81	44	0	656	2020-02-15	EeWYNKnS8Fg	...عمرو دياب يرد على سؤال عمرو أديب: مينا الشربين	78052

By Channels

Analyzing Particular Channel

There are times when we want to specifically extract data for a particular YouTube Channel that we are interested to analyze. In such cases, we can add in a new parameter called `channelId` and set the channel ID accordingly. You can find out the channel ID from the Channel Page website address, for example https://www.youtube.com/channel/UCnFKsbAof9fRv614I4wJX_w, the channel ID is the last part of the URL i.e. "UCnFKsbAof9fRv614I4wJX_w". To simplify things, we will modify the earlier Function and assign it with a new name.

```
test_lead = youtube_search_channelId("UCnFKsbAof9fRv614I4wJX_w", max_results = 50)
```

```
# take a look at the top 10 videos from the result
test_lead.head(10)
```

	channelTitle	commentCount	dislikeCount	favoriteCount	likeCount	publishedAt	v_id	v_title	viewCount
0	MBC مصر	5271	9220	0	249170	2017-12-30	5LLP-KMIaI0	محمد أسامه يمدح في عطاء عز الحبايب	35684953
1	MBC مصر	0	9037	0	151125	2015-10-24	wHELCNX8Ppo	...تفاعل الجمهور مع حامد بشكل أكثر من رائع .. شا	27646924
2	MBC مصر	0	9558	0	210162	2015-11-14	q6qM32PyiTI	...حامد يقف أصوات الفنانين بتأثر بداية من على ال	26321033
3	MBC مصر	0	6746	0	110851	2016-06-24	CcHDI_rPypg	...أول لقاء بين ناصر و نمارا بعد ما رفضته تفكروا	25674543
4	MBC مصر	2754	6087	0	144860	2018-07-28	FZBslmHPSJE	...على رسمي بخرج عن النص ويوجه رسالة كوميدية لـ ج	20078117
5	MBC مصر	0	4951	0	86813	2016-07-02	SpiDRuOMLg0	...ناصر لـ نمارا : انتي كسرتيني انتي واهنك وخلتيني	17632115
6	MBC مصر	2325	5624	0	113488	2018-02-09	SY25AhPLn5A	...إسراء بتقلش على جوزها حمدي ميرغني.. شوف رد على	16356782
7	MBC مصر	0	4307	0	68352	2016-12-09	uY6MrKB6wME	نقيب رائع للفنان عبد الباسط حمودة في #مسرح_مصر	15356514
8	MBC مصر	0	4248	0	71904	2015-10-31	jA0qxUCtEe8	...مسرح مصر - تعرف على توافق برج الكلب مع الأبراج	15187380
9	MBC مصر	0	3232	0	84405	2015-11-28	vV8jKuBSiEg	... مسرح مصر	حمدي المرغني في دور النحلتي .. هل#	14874385

Visualization

From the results above, the YouTube channel "mbs masr" only has 43 videos, we will plot a bar chart to visualize all the 43 videos for this channel

**Note: ** From the charts, it seems that the view count and like count are concentrated on one video, titled "elhekayashow" with over 3.5*e7 views

Futher Analysis

We can also analyze the data further by breaking down the view count by the day of the week. To do so, we need to extract the day of the week from the "publishedAt" column. Since the column is already in datetime format, we can use the `dt.dayofweek` property to extract the day. By default, this property returns the day of the week with Monday=0, Sunday=6. To make it more intuitive, we will add 1 to the result, which will become Monday=1, Sunday=7.

```
In [47]: # create a new column and extract the day of the week + 1
test_lead["weekday"] = test_lead["publishedAt"].dt.dayofweek + 1
test_lead
```

Out[47]:

	channelTitle	commentCount	dislikeCount	favoriteCount	likeCount	publishedAt	v_id	v_title	viewCount	weekday
0	MBC مصر	5271	9220	0	249170	2017-12-30	5LLP-KMIaI0	محمد أسامه يبدع في غناء عز الحبايب	35684953	6
1	MBC مصر	0	9037	0	151125	2015-10-24	wHELCNX8Ppo	تفاعل الجمهور مع حامد بشكل أكثر من رائع ..... شا	27646924	6
2	MBC مصر	0	9558	0	210162	2015-11-14	q8qM32PyiTI	حامد يقلد أصوات الفنانين بتأثق بداية من ...علي ال	26321033	6
3	MBC مصر	0	6746	0	110851	2016-06-24	CcHDI_rPypg	أول لقاء بين ناصر و تمارا بعد ما رفضته ...تفكروا	25674543	5
4	MBC مصر	2754	6087	0	144860	2018-07-28	FZBslmHPSJE	علي ربيع يخرج عن النص ويوجه رسالة ...كواسية لـ ج	20078117	6
5	MBC مصر	0	4951	0	86813	2016-07-02	SpiDRuOMLg0	ناصر لـ تمارا : انتي كسرتني انتي واهلك ...وخلتين	17632115	6
6	MBC مصر	2325	5624	0	113488	2018-02-09	SY25AhPLn5A	إسراء بتقلش على جوزها حمدي امرغني .. شوف رد علي	16356782	5
7	MBC مصر	0	4307	0	68352	2016-12-09	uY6MrK86wME	تقليد رائع للفنان عبد الباسط حمودة في #نصر_مصر	15366514	5

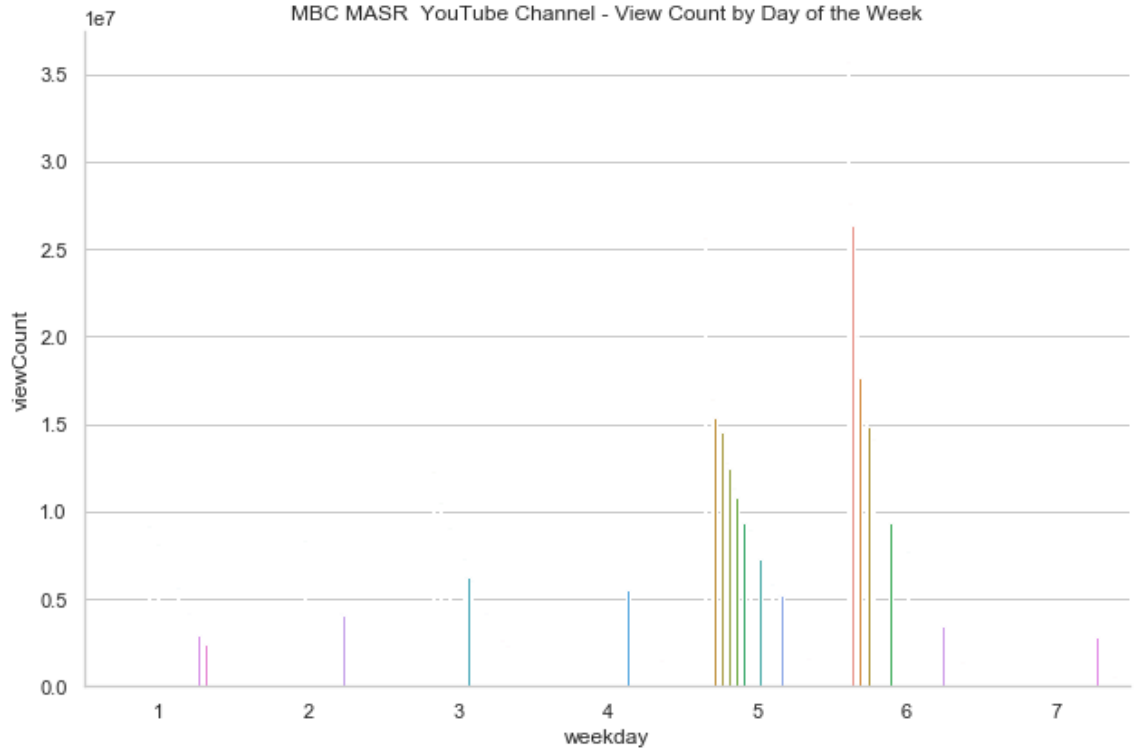
Note: It seems like wednesday has the highest average and total view count.

```
In [48]: # group the "weekday" column by view count
         test_lead.groupby("weekday").agg({"viewCount": [lambda x: x.count(), np.sum, np.mean]}).rename(columns={'<lambda>': 'Count',
                                                                                                                 'sum': 'Sum',
                                                                                                                 'mean': 'Average'})
```

Out[48]:

	viewCount		
	Count	Sum	Average
weekday			
1	6	32420322	5.403387e+06
2	3	16495383	5.498461e+06
3	10	66880725	6.688072e+06
4	2	7011289	3.505644e+06
5	14	161020438	1.150146e+07
6	12	193431860	1.611932e+07
7	3	9301888	3.100629e+06

plot a bar chart to see the distribution of view count by videos:



Note: It seems that the reason wednesday has the highest average and total view count is due to an extreme value i.e. the video with the highest view count was published on wednesday with over 500 views. But this does not mean that videos published on wednesday will always have high view count, in this case it is just because of the effect of one extreme value. A closer look at the viewCount by weekday table above, the other days with the highest average view count aretday, thursday and monday. So perhaps, it is a wise move to publish videos during weekends in order to get more views

Extracting Video Comments

YouTube Data API can also return the comments of a video. There are times when we want to extract the comments for a particular video to analyze futher, for example to do sentiment analysis on the comments to gauge the users' reaction to the video. To do this, we can use YouTube's CommentThreads. A commentThread resource contains information about a YouTube comment thread, which comprises a top-level comment and replies, if any exist, to that comment. Again, to make things simpler, we will create a new function by modifying the earlier function. The maximum number of items that will be returned in the result per call is 100. More info at https://developers.google.com/youtube/v3/docs/commentThreads

```
In [50]: # call the function to extract the comments
         comments = youtube_video_comments("zYqcOpSQ69s", max_results = 100)
```

```
In [48]: # take a look at the results
         print("No of comments: {}".format(len(comments)))
         comments

           No of comments: 17

Out[48]: [{'id': 'UgyYBJ0KurcIUQJ_vMJ4AaABAg', 'text': 'ايه ده محمد فؤاد', 'likes': 0},
          {'id': 'UgzWAmsK7Vmd_yy2Es94AaABAg',
           'text': 'دمعت والله كلمات رائعه وصوت اروع... ربنا حيعوضك لانه هو الرزاق وأنت صوت يستحق',
           'likes': 0},
          {'id': 'UgxaWumbz2B51SxRqC94AaABAg',
           'text': 'جميل جدا أنا احترمته والله',
           'likes': 1},
          {'id': 'UgwrWq0KPKUTupjenPR4AaABAg',
           'text': 'صوته روعه😍💛💛💛💛💛💛💛💛💛💛💛💛💛💛💛💛💛💛💛💛💛',
           'likes': 1},
```

perform tokenization:

the w+ allows us to capture the word as a whole, and giving us a much cleaner result compared to only using split()

```
tokens = re.findall(r'\w+', text)
print("Number of tokens = "+"{:,}\n".format(len(tokens)))
print(tokens[:50])

  Number of tokens = 258

  ['ايه', 'ده', 'محمد', 'فؤاد', 'دمعت', 'والله', 'كلمات', 'رائعه', 'وصوت', 'اروع', 'ربنا', 'حيعوضك', 'لانه', 'هو', 'الرزاق', 'وأنت', 'صوت', 'يستحق',
  'جميل', 'جدا', 'أنا', 'احترمته', 'والله', 'صوته', 'روعه', 'حسبنا', 'الله', 'ونعم', 'الوكيل', 'فيك', 'يا', 'هاني', 'يا', 'شاكر', 'وصلوني', '١٠٠٠', 'مشتر',
  'الشيخ', 'الروحاني', 'ابو', 'محمود', '٦٦00963955455366', 'ك', 'ادخل', 'شوف', 'قناتي', 'والفيديوهات', 'ربنا', 'يرزقك', 'ويفرحك', 'ديما']
```

Sentiment Analysis using NLTK

We will perform sentiment analysis using NLTK’s VADER (a Python module) to classify comments as positive, negative or neutral. NLTK comes with an inbuilt sentiment analyser module – nltk.sentiment.vader—that can analyse a piece of text and classify the sentences under positive, negative and neutral polarity of sentiments. The "compound" value conveys the overall positive or negative user experience.

```
In [58]: # classify the comments
summary = {"positive":0,"neutral":0,"negative":0}

for x in all_text:
    ss = sid.polarity_scores(x)
    if ss["compound"] == 0.0:
        summary["neutral"] +=1
    elif ss["compound"] > 0.0:
        summary["positive"] +=1
    else:
        summary["negative"] +=1

summary

Out[58]: {'positive': 0, 'neutral': 17, 'negative': 0}
```

```
In [59]: # take a look at the different sentiment proportions for individual comments
for sentence in all_text:
    print(sentence)
    ss = sid.polarity_scores(sentence)
    for k in sorted(ss):
        print('{0}: {1}, '.format(k, ss[k]), end='')
    print("\n")

ايه ده محمد فؤاد
compound: 0.0, neg: 0.0, neu: 1.0, pos: 0.0,

دمعت والله كلمات رائعه وصوت اروع... ربنا حيعوضك لانه هو الرزاق وأنت صوت يستحق
compound: 0.0, neg: 0.0, neu: 1.0, pos: 0.0,

جميل جدا أنا احترمته والله
compound: 0.0, neg: 0.0, neu: 1.0, pos: 0.0,

صوته روعه😍💛💛💛💛💛💛💛💛💛💛💛💛💛💛💛💛💛💛💛💛
compound: 0.0, neg: 0.0, neu: 1.0, pos: 0.0,
```

Findings

Out of the 100 comments, 0 are classified as positive, 17 as neutral and 0 as negative. This indicates that the viewers' responses are mixed, although one could argue that the sentiment is positive as there are more positive comments than negative ones. Given the small sample size of 100, whether this positive sentiment could translate into good box office result is anyone's guess. But it does indicate to the channel studio that they need to beef up its marketing effort to promote the channel.