

PROFESSIONAL TRAINING REPORT
at
Sathyabama Institute of Science and Technology
(Deemed to be University)

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Technology Degree in

Information Technology

By

SOWMIYA K (Reg. No. 37120072)



DEPARTMENT OF INFORMATION TECHNOLOGY
SCHOOL OF COMPUTING
SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY
JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI – 600119. TAMIL NADU

JANUARY 2021



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with "A" Grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai - 600 119.

Phone: 044 - 2450 3150 / 51 / 52 / 54 / 55 Fax: 044 - 2450 2344

www.sathyabama.ac.in



SCHOOL OF COMPUTING

BONAFIDE CERTIFICATE

This is to certify that this Professional Training Report is the bonafide work of SOWMIYA.K (Reg. No.37120072) who underwent the professional training in "**E-DESIGN INTERIOR SERVICES**" under our supervision from July 2019 to September 2019.

Internal Guide

Dr. P. JEYANTHI, ME., Ph.D.,

Head of the Department

Dr. R. SUBHASHINI, M.E., Ph.D.,

Submitted for Viva voce Examination held on_____

Internal Examiner

External Examiner

DECLARATION

I, SOWMIYA (Reg. No. 37120072) hereby declare that the Professional Training Report on “**SOCIAL MEDIA DATA ANALYTICS**” done by me under the guidance of Dr.P.JEYANTHI, ME., Ph.D., at Sathyabama Institute of Science and Technology is submitted in partial fulfillment of the requirements for the award of Bachelor of technology degree in **Information Technology**.

DATE:

PLACE:

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to Board of Management of SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to Dr. T.Sasikala M.E., Ph.D, Dean, School of Computing , Dr.SUBHASHINI M.E., Ph.D. , Head of the Department of Information and Technology for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide Dr.P.JEYANTHI M.E., Ph.D.,for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of Information Technology who were helpful in many ways for the completion of the project.

ABSTRACT

Analysis of structured data has seen tremendous success in the past. However, analysis of large-scale unstructured data in the form of video format remains a challenging area. YouTube, a Google company, has over a billion users and generates billions of views. Since YouTube data is getting created in a very huge amount and with an equally great speed, there is a huge demand to store, process and carefully study this large amount of data to make it usable. The project utilizes the YouTube Data API (Application Programming Interface) that allows the applications/websites to incorporate functions that are used by YouTube application to fetch and view information. The Google Developers Console is used to generate a unique access key which is further required to fetch YouTube public channel data. This project extracts the meaningful output of which can be used by the management for analysis.

CHAPTER	TITLE	PAGE NO
	ABSTRACT	5
	LIST OF TABLES	6
	LIST OF FIGURES	9
1	INTRODUCTION	10
	1.1 Synopsis	10
	1.2 Objective	11
	1.2 Problem definition	11
2	SOCIAL MEDIA ANALYTICS	12
	2.1. Analysis	12
	2.2. The Era Of Big Data And Social Media Analytics	12
3	CHALLENGES OF SOCIAL MEDIA ANALYTICS	14
	3.1. Getting social media metrics to the right people	14
	3.2. Visualization	15
	3.3. Unstructured data is challenging	15
	3.4. Increasing signal to noise	15
	3.5. A wait and see attitude	16

4	TERMINOLOGY	17
	4.1 Natural language processing	17
	4.2 News analytics	17
	4.3 Opinion mining	17
	4.4 Scraping	17
	4.5 Sentiment analysis	17
	4.6 Text analytics	18
	4.7 Figures	18
	4.7.1 I/O Diagram	18
	4.7.2 Usecase Diagram	19
	4.7.3 Activity Diagram	19
	4.7.4 Data Diagram	20
	4.7.5 Sequential Diagram	20
	4.7.6 Architecture diagram	21
5	Social Media Overview	22
	5.1 Social media data	22
	5.2 Social media programmatic access	22
	5.3 Data sources, services and tools	22
	5.4 Data feeds via APIS	22
	5.5 Social media platforms	23

6	SOCIAL MEDIA DATA	24
	6.1 Types of data	24
	6.2 Text data formats	24
	6.3 Open-source databases	26
	6.4 Data access via tools	26
	6.4.1 Freely accessible sources	26
	6.4.2 Commercial sources	27
	6.4.3 Data feed access via APIs	28
	6.4.4 Wiki media	28
	6.4.5 Social networking media	29
7	Fetching YouTube Data Using API	30
	7.1 Google API	30
	7.2 Getting API and scrapping data	30
8	YOUTUBE DATA ANALYSIS	35
	8.1 Occurrence of words in Title	35
	8.2 Trending videos by channel	35
	APPENDIX	36
	CONCLUSION	40
	REFERENCE	42

LIST OF FIGURES

FIGURES	TITLE	PAGE NO
4	TERMINOLOGIES	13
	4.7 Figures	14
	4.7.1 I/O Diagram	14
	4.7.2 Use case Diagram	14
	4.7.3 Activity Diagram	15
	4.7.4 Data Diagram	16
	4.7.5 Sequential Diagram	16
	4.7.6 Architecture diagram	17

LIST OF TABLES

FIGURES	TITLE	PAGE NO
8	YOUTUBE DATA ANALYSIS	35
	8.1 Occurrence of words in title	35
	8.2 Trending Videos by Channels	35

INTRODUCTION

Social media analytics is the process of gathering and analyzing data from social networks such as YouTube, Facebook, Instagram, LinkedIn and Twitter. It is commonly used by marketers to track online conversations about products and companies. One author defined it as "the art and science of extracting valuable hidden insights from vast amounts of semi-structured and unstructured social media data to enable informed and insightful decision making.

Similarly, in this project going to scarp and analysis the YouTube data using google API. YouTube as a user generated content is one of the largest and most popular video sharing websites. It hosts over four billion views a day. YouTube provides public statistics regarding its uploaded videos, most notably the number of views, which shows the aggregate number of times a video has been watched up to that point. Naturally, the number of views for a video indicates the level of popularity of that video; and it takes a varying amount of time for a video to become popular (if it becomes popular). Meanwhile, there are some videos that can attract users' attention in a relatively short time.

YouTube also supports a feature called trending, which represents content that has the potential of becoming popular in a relatively short time. Consequently, although trending videos are usually not popular (yet) when declared as trending by YouTube, they have the potential of becoming popular (eventually). For example, some videos are labeled trending while having only few hundreds in viewership numbers. From another perspective, through trending videos, YouTube tries to highlight emerging trends developing within different viewership communities.

Meanwhile, the general attributes of the viewership of trending videos have not been studied thoroughly. To the best of our knowledge, basic statistics about YouTube trending videos have not been studied, analyzed, or even received any adequate attention. Considering the fact that more than one billion unique users visit YouTube each month and they upload 72 hours of video every minute, YouTube is the best place for e.g. brand engagement or advertising, but it is genuinely difficult and competitive to get the attention of users.

Therefore, when a video becomes popular, it is exposed to millions of users for free and has the opportunity of keeping their attention for a while. Finding these trends are significantly important that many different websites have been emerged just to pick up on the latest trends on the web, such as “BuzzFeed” or “WhatTheTrend”. There are other kinds of websites such as “channel Factory” that tries to build an audience on YouTube for content owners or advertisers.

Understanding of YouTube trending videos and their statistics, and a deeper insight about their lifecycles, can greatly affect the strategies for marketing, target advertising, recommendation systems and search engines, as was suggested by prior YouTube measurement studies. This represents a key motivation for our effort.

In this paper, we present our findings for measuring, analyzing, and comparing key aspects of YouTube trending videos. Our study is based on collecting and monitoring high-resolution time-series of the viewership and related statistics of more than 8,000 YouTube videos over an aggregate period. To put this number into perspective, YouTube declares only tens of videos as trending on a daily basis. This number is highly selective when compared to the thousands of videos uploaded on YouTube every minute.

Furthermore, and from the onset of our study, our goal has been to collect data viewership time-series of trending videos to achieve robust analysis of their entire lifecycles. Since trending videos are declared as such just several hours after they are uploaded, we are able to analyze trending videos across their lifecycle; this provides an invaluable insight into their viewership time-series over a critical period of their lifetime.

Social Media Monitoring is the second term most often confused for Social Media Analytics. It's also thought to be synonymous with Social Listening, but the two are very different. Social Monitoring focuses on following social audiences to be alerted to spikes in activity that present either an opportunity you wouldn't want to miss, or a potential disaster you want to avoid. It's about seeing posts like this in time to respond and avoid a viral crisis.

CHAPTER 2

SOCIAL MEDIA ANALYTICS

Social Media Analytics is an emerging interdisciplinary research field that aims on combining, extending, and adapting methods for analysis of social media data. On the one hand it can support IS and other research disciplines to answer their research questions and on the other hand it helps to provide architectural designs as well as solution frameworks for new social media-based applications and information systems.

2.1. Analysis

This is where you go beyond simply reporting the numbers to analyse what the data actually means. For example, rather than simply reporting the number of engagements, you might want to analyse the number of engagements per post or per follower. And rather than simply reporting the number of clicks you drove to your website; you might want to analyse how much of total website traffic is coming from social campaigns.

Again, make sure to provide context by comparing your results to previous campaigns, or to what you know about competitor campaigns. If you can, try to analyse your data to determine why you got the results you did. The facts are interesting, sure, but the reasons behind the data are what can help you tweak your strategy and set meaningful social media goals.

2.2.The Era Of Big Data And Social Media Analytics

The era of Big Data is underway, computer scientists, physicists, economists, mathematicians, political scientists, bio-informaticists, sociologists, and other scholars are clamoring for access to the massive quantities of information produced by and about people, things, and their interactions (Boyd et al., 2012). Parliamentary office of science and technology in its journal Houses of parliament, number 460 March 2014 write an article and brought some truths about social media and big data: 57% of over-16s in the UK use social media, generating vast amounts of accessible data. Analyzing social media data can help organizations understand behaviors and target products and services more effectively. Key applications include profiling voters and

complementing traditional polling, targeting adverts at consumers, credit scoring and informing policing decisions. There is a debate about how to analyze social media data, including which methods to use and how to control for biases.

Personal data can be shared or sold with 146 users' consent as long as they are anonymized. There are concerns that users are not fully aware of how their data are being used and that it is often possible to identify individuals from linking anonymized datasets. Analyzing large quantities of readily available data from social media has created new opportunities to understand and influence how people think and act. The rate of unstructured data production on social media makes it difficult to analyze using traditional methods that rely on human analysts. Social media analytics is a new field of study that is developing automated or semi-automated methods for analyzing data. Some advocates of big data argue that the sheer size of the datasets reduces, or even eliminates, the need for established statistical methods such as random sampling, because all the data can be analyzed.

However, in the case of social media data, it only contains data about people that use social media. In the UK, around 49% of the population use Facebook and 24% use Twitter and not all users create content. There are concerns that social media data may not represent vulnerable groups in society, such as the elderly or those from lower income backgrounds. This means that there are significant gaps in the data, and there are not yet accepted methods for controlling for biases. This paper presents an overview on studies associated with big data in social media. The study uses Scopus database as a primary search engine and analyzes the data.

CHAPTER 3

CHALLENGES OF SOCIAL MEDIA ANALYTICS

3.1. Getting social media metrics to the right people

Often, social media is treated like the ugly stepchild within the marketing department and real-time social media analytics are either absent or ignored. Real-time social media analytics creates serious challenges for many organizations. Often, organizations are married to an old paradigm — a vestige of by-gone days when data was hard to get, taking months of data gathering and analysis. These organizations didn't integrate data gathering into tactical and strategic decision-making because they couldn't. Incorporating real-time analytics just isn't possible with their existing bureaucracy.

For one thing, real-time analytics requires moving analysts closer to decision-makers and enabling decision-makers with analytic skills for ad hoc data analysis. But, that's not what most businesses look like. Many decision-makers lack analytics skills necessary for ad hoc analysis.

New arrival, Uber, which runs a ride-sharing program that competes with taxi companies and car services, uses real-time analytics to show how people move around a city at any give time, allowing Uber to optimize their customer service. Placing cars nearby reduces competition with local cab companies and real-time analytics provide insights necessary to do that. To do this, Uber uses real-time data to incentivize more drivers to provide services by raising the price of a ride.

Others such as Samsung and NASCAR do a great job of providing real-time social media analytics to guide decision-makers. Take a look at this command centre NASCAR uses to monitor chatter surrounding their events.

3.2. Visualization

Visualizing real-time social media analytics is another key element involved in developing insights that matter. Face it, human beings don't do a great job of processing long tables of numbers. Notice on NASCAR's command centre, much of the data is displayed visually.

Simply displaying values graphically helps in making the kinds of fast interpretations necessary for making decisions with real-time data, but adding more complex algorithms and using models provides deeper insights, especially when visualized.

3.3. Unstructured data is challenging

Unlike the survey data firms are used to dealing with, most (IBM estimates 80%) is unstructured — meaning it consists of words rather than numbers. And, text analytics lags seriously behind numeric analysis.

While unstructured data tends to muck-up any kind of analysis, it's especially challenging in the context of real-time analytics, because you want interpretations IN REAL TIME. Handling text in real time often means using computer-generated translations of the written word. However, no computer can effectively categorize much of what's written in social media where “bad” might mean bad or it might mean good, depending on context, relationship, and other variables.

3.4. Increasing signal to noise

Social media data is inherently noisy. Reducing noise to even detect signal is challenging — especially in real time. Sure, with enough time, new analytics tools can ferret out the few meaningful comments across various social networks, but few can handle this in realtime.

3.5. A wait and see attitude

Again, businesses are used to a certain operational model that makes real-time social media analytics challenging. For instance, we listened to a presentation by an analyst from NPR. He showed complex A/B testing used to determine the effectiveness of headlines, even whole articles online. As a statistician, he's concerned about achieving statistical significance in his testing before making decisions.

That's great if you're talking about putting \$100 million into building and marketing a product, but doesn't make much sense in the fast-paced world of social media. Real-time analytics require real-time decisions. Period.

If it'll take you several days to gather enough data for statistical significance, forget it. Especially if you're only trying to determine which headline does better, by the time you have a statistically significant answer, no one cares anymore. The news trend has moved to another topic.

CHAPTER 4

TERMINOLOGY

Start with definitions of some of the key techniques related to analyzing unstructured textual data:

4.1 Natural language processing

(NLP) is a field of computer science, artificial intelligence and linguistics concerned with the interactions between computers and human (natural) languages. Specifically, it is the process of a computer extracting meaningful information from natural language input and/or producing natural language output.

4.2 News analytics

The measurement of the various qualitative and quantitative attributes of textual (unstructured data) news stories. Some of these attributes are: sentiment, relevance and novelty.

4.3 Opinion mining

Opinion mining (sentiment mining, opinion/sentiment extraction) is the area of research that attempts to make automatic systems to determine human opinion from text written in natural language.

4.4 Scraping

Collecting online data from social media and other Web sites in the form of unstructured text and also known as site scraping, web harvesting and web data extraction.

4.5 Sentiment analysis

Sentiment analysis refers to the application of natural language processing, computational linguistics and text analytics to identify and extract subjective information in source materials.

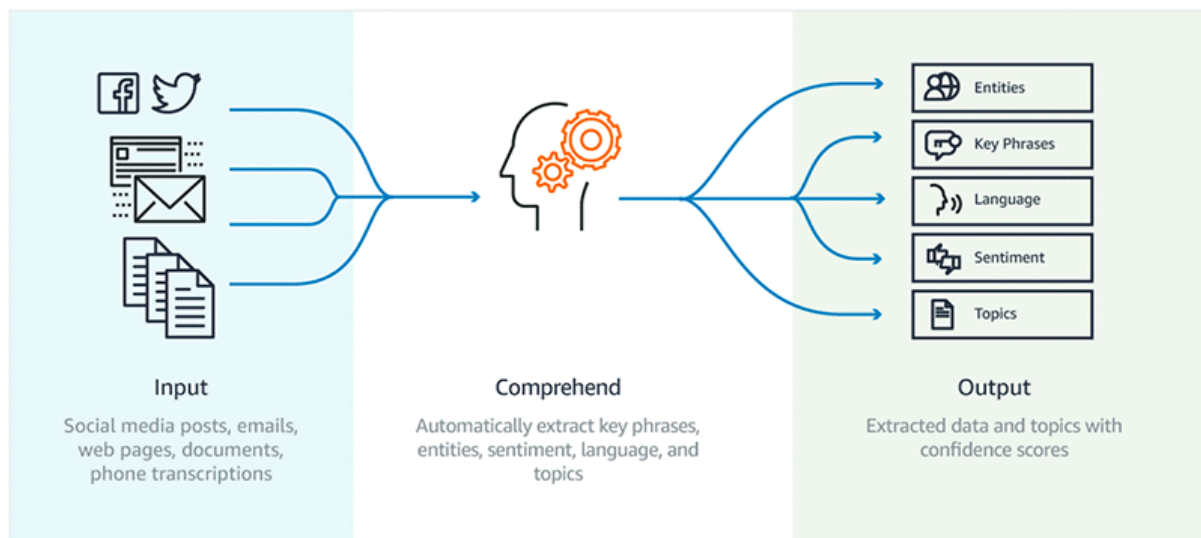
4.6 Text analytics

involves information retrieval (IR), lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization and predictive analytics

4.7 Figures

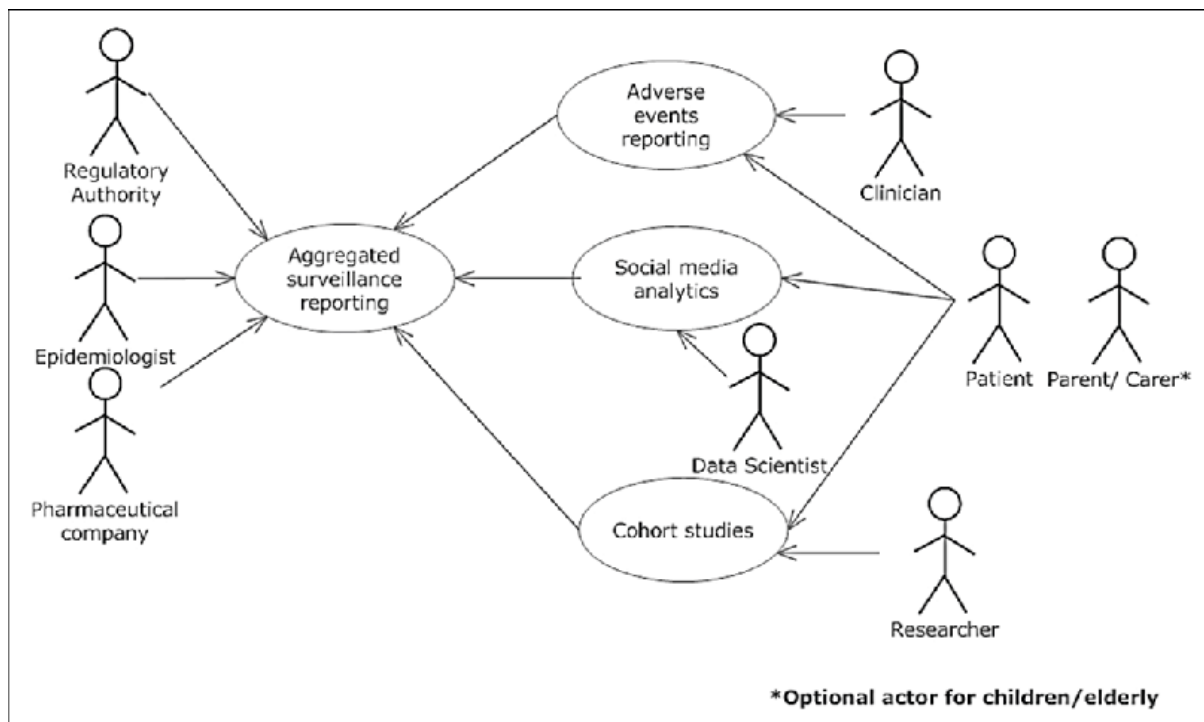
4.7.1 I/O Diagram

Input design is the process of converting user-originated inputs to a computer-based format. Input design is one of the most expensive phases of the operation of computerized system and is often the major problem of a system



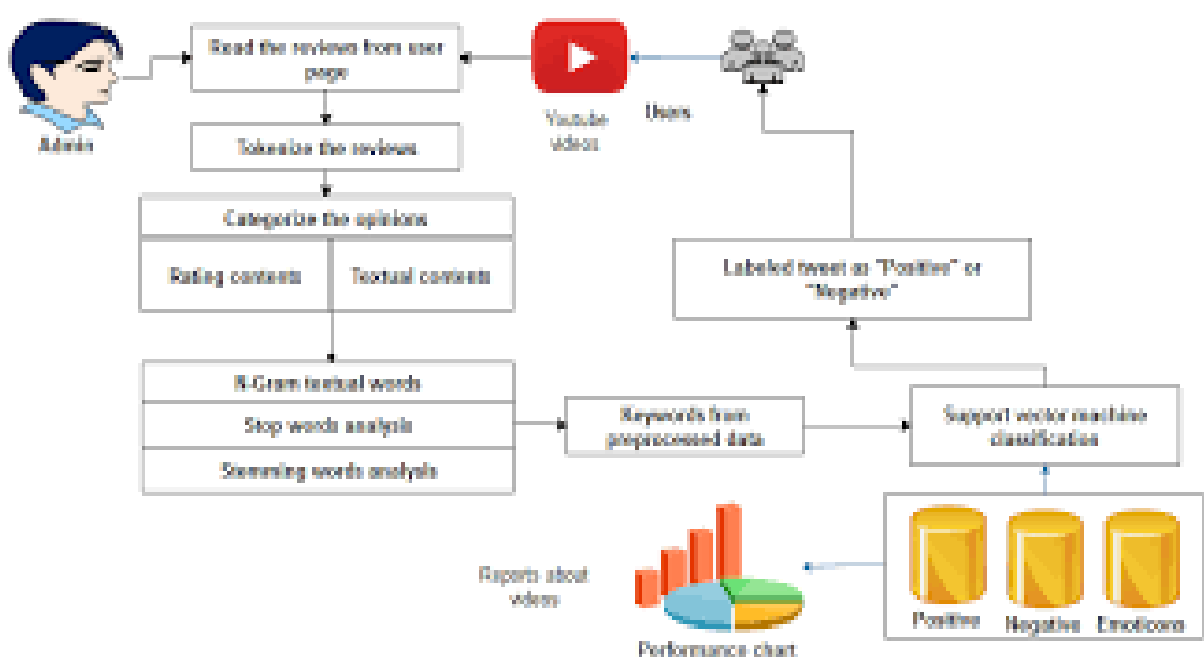
4.7.2 Usecase Diagram

A use case diagram at its simplest is a representation of a user's interaction with the system that shows the relationship between the user and the different use cases in which the user is involved. A use case diagram can identify the different types of users of a system and the different use cases and will often be accompanied by other types of diagrams as well.



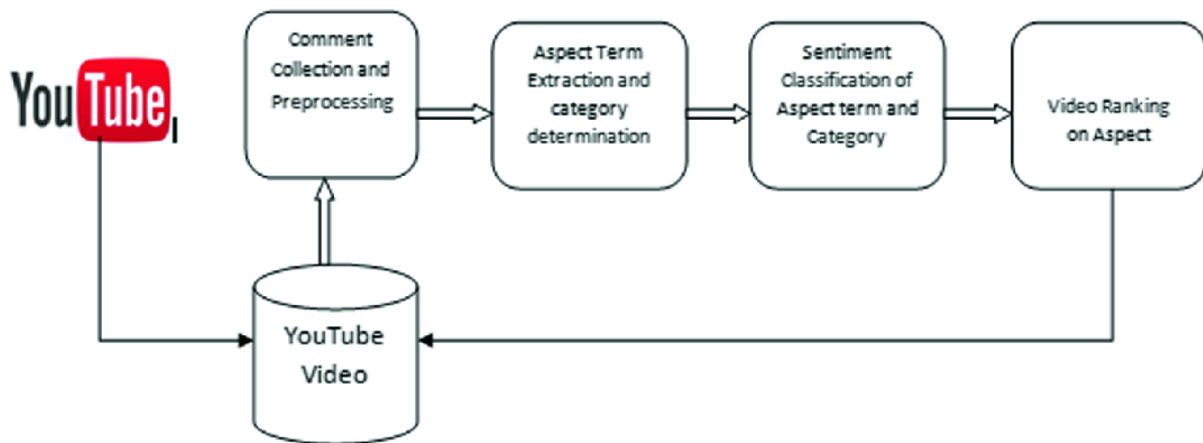
4.7.3 Activity Diagram

Activity diagram is another important diagram in UML to describe the dynamic aspects of the system. Activity diagram is basically a flowchart to represent the flow from one activity to another activity. The activity can be described as an operation of the system. The control flow is drawn from one operation to another.



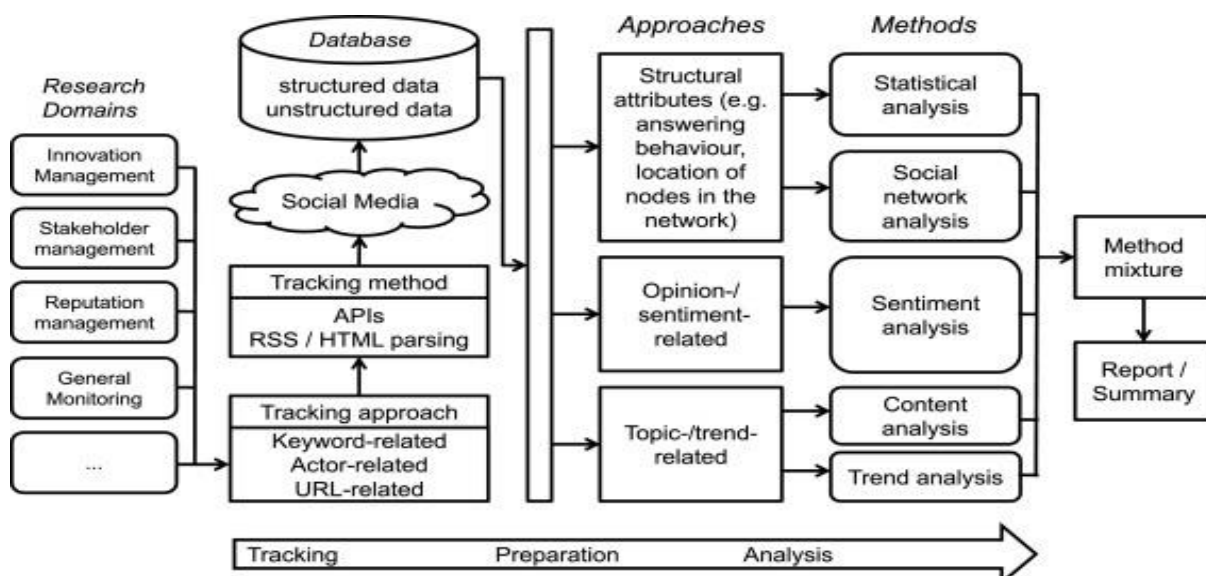
4.7.4 data Diagram

Data design is the first design activity, which results in less complex, modular and efficient program structure. The information domain model developed during analysis phase is transformed into data structures needed for implementing the software.



4.7.5 Sequential Diagram

Sequence diagram is the most common kind of interaction diagram, which focuses on the message interchange between a number of lifelines. Sequence diagram describes an interaction by focusing on the sequence of messages that are exchanged, along with their corresponding occurrence specifications on the lifelines. UML sequence diagrams are used to show how objects interact in a given situation.

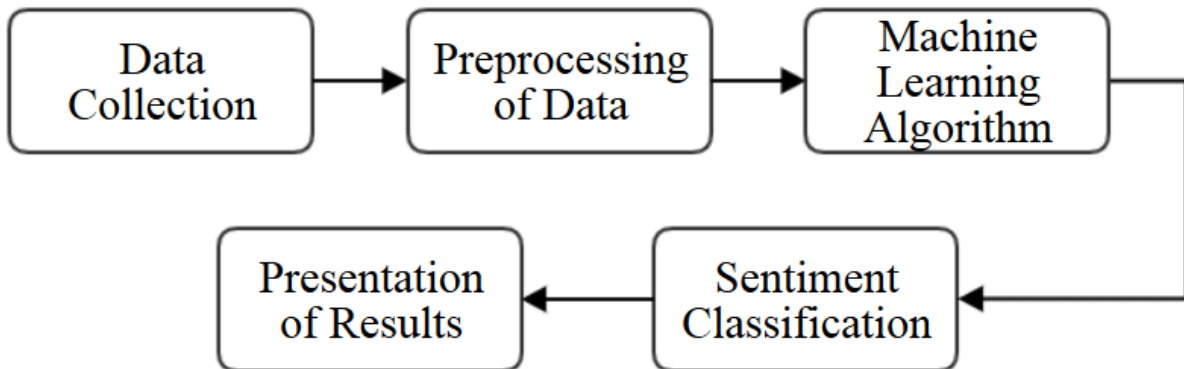


4.7.6 Architecture Diagram

The design process for identifying the sub-systems making up a system and the framework for sub-system control and communication is architectural design. The output of this design process is a description of the software architecture.

- It defines an abstraction level at which the designers can specify the functional and performance behaviors of the system.
- It evaluates all top-level designs
- It develops and documents top-level designs for the external and internal interface
- It develops all preliminary versions of users' documentation
- It defines and documents preliminary test requirements and the schedule for software integration.

System architecture is a conceptual model that defines the structure, behaviors, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structure and behavior of the system.



CHAPTER 5

Social Media Overview

For this paper, we group social media tools into:

5.1 Social media data

Social media data types (e.g., social network media, wikis, blogs, RSS feeds and news, etc.) and formats (e.g., XML and JSON). This includes data sets and increasingly important real-time data feeds, such as financial data, customer transaction data, telecoms and spatial data.

5.2 Social media programmatic access

Data services and tools for sourcing and scraping (textual) data from social networking media, wikis, RSS feeds, news, etc. These can be usefully subdivided into:

5.3 Data sources, services and tools

Where data is accessed by tools which protect the raw data or provide simple analytics. Examples include: Google Trends, Social Mention, Social Pointer and Social Seek, which provide a stream of information that aggregates various social media feeds.

5.4 Data feeds via APIS

Where data sets and feeds are accessible via programmable HTTP-based APIs and return tagged data using XML or JSON, etc. Examples include Wikipedia, Twitter and Facebook.

- **Text cleaning and storage tools**

Tools for cleaning and storing textual data. Google Refine and Data Wrangler are examples for data cleaning.

- **Text analysis tools**

Individual or libraries of tools for analyzing social media data once it has been scraped and cleaned. These are mainly natural language processing, analysis and classification tools, which are explained below.

- **Transformation tools**

Simple tools that can transform textual input data into tables, maps, charts (line, pie, scatter, bar, etc.), timeline or even motion (animation over timeline), such as Google Fusion Tables, Zoho Reports, Tableau Public or IBM's Many Eyes.

- **Analysis tools**

More advanced analytics tools for analyzing social data, identifying connections and building networks, such as Gephi (open source) or the Excel plug-in NodeXL.

5.5 Social media platforms

Environments that provide comprehensive social media data and libraries of tools for analytics. Examples include: Thomson Reuters Machine Readable News, Radian 6 and Lexalytics. **Social network media platforms:** Platforms that provide data mining and analytics on Twitter, Facebook and a wide range of other social network media sources. • News platforms—platforms such as Thomson Reuters providing commercial news archives/feeds and associated analytics.

CHAPTER 6

SOCIAL MEDIA DATA

Clearly, there is a large and increasing number of (commercial) services providing access to social networking media (e.g., Twitter, Facebook and Wikipedia) and news services (e.g., Thomson Reuters Machine Readable News). Equivalent major academic services are scarce. We start by discussing types of data and formats produced by these services.

6.1 Types of data

Although we focus on social media, as discussed, researchers are continually finding new and innovative sources of data to bring together and analyze. So, when considering textual data analysis, we should consider multiple sources (e.g., social networking media, RSS feeds, blogs and news) supplemented by numeric (financial) data, telecoms data, geospatial data and potentially speech and video data. Using multiple data sources is certainly the future of analytics. Broadly, data subdivides into:

- Historic data sets—previously accumulated and stored social/news, financial and economic data.
- Real-time feeds—live data feeds from streamed social media, news services, financial exchanges, telecoms services, GPS devices and speech. And into:
 - Raw data—unprocessed computer data straight from source that may contain errors or may be unanalyzed.
 - Cleaned data—correction or removal of erroneous (dirty) data caused by disparities, keying mistakes, missing bits, outliers, etc.
 - Value-added data—data that has been cleaned, analyzed, tagged and augmented with knowledge.

6.2 Text data formats

The four most common formats used to markup text are: HTML, XML, JSON and CSV.

- HTML—Hypertext Markup Language (HTML) as well-known is the markup language for web pages and other information that can be viewed in a web browser. HTML consists of HTML elements, which include tags enclosed in angle brackets (e.g., `<div>`), within the content of the web page.

- XML—Extensible Markup Language (XML)—the markup language for structuring textual data using `<tag[...<tag]` to define elements.
- JSON—JavaScript Object Notation (JSON) is a textbased open standard designed for human-readable data interchange and is derived from JavaScript.
- CSV—a comma-separated values (CSV) file contains the values in a table as a series of ASCII text lines organized such that each column value is separated by a comma from the next column's value and each row starts a new line. For completeness, HTML and XML are so-called markup languages (markup and content) that define a set of simple syntactic rules for encoding documents in a format both human readable and machine readable. A markup comprises start-tags (e.g., `<tag`), content text and endtags (e.g., `>tag`).

Many feeds use JavaScript Object Notation (JSON), the lightweight data-interchange format, based on a subset of the JavaScript Programming Language. JSON is a language-independent text format that uses conventions that are familiar to programmers of the C-family of languages, including C, C++, C#, Java, JavaScript, Perl, Python, and many others. JSON's basic types are: Number, String, Boolean, Array (an ordered sequence of values, comma separated and enclosed in square brackets) and Object (an unordered collection of key: value pairs). The JSON format is illustrated in Fig. 1 for a query on the Twitter API on the string 'UCL,' which returns two 'text' results from the Twitter user 'uclnews.' Comma-separated values are not a single, well-defined format but rather refer to any text file that:

- (a) is plain text using a character set such as ASCII, Unicode or EBCDIC.
- (b) consists of text records (e.g., one record per line)
- (c) with records divided into fields separated by delimiters

4 Social media providers

Social media data resources broadly subdivide into those providing:

- Freely available databases—repositories that can be freely downloaded, e.g., Wikipedia (<http://dumps.wikimedia.org>) and the Enron e-mail data set available via <http://www.cs.cmu.edu/~enron/>.

- Data access via tools—sources that provide controlled access to their social media data via dedicated tools, both to facilitate easy interrogation and also to stop users 'sucking' all the data from the repository. An example is Google's Trends. These further subdivided into:

(1) Free sources—repositories that are freely accessible, but the tools protect or may limit access to the ‘raw’ data in the repository, such as the range of tools provided by Google.

(2) Commercial sources—data resellers that charge for access to their social media data. Gnip and DataSift provide commercial access to Twitter data through a partnership, and Thomson Reuters to news data.

- Data access via APIs—social media data repositories providing programmable HTTP-based access to the data via APIs (e.g., Twitter, Facebook and Wikipedia).

6.3 Open-source databases

A major open source of social media is Wikipedia, which offers free copies of all available content to interested users (Wikimedia Foundation 2014). These databases can be used for mirroring, database queries and social media analytics. They include dumps from any Wikimedia Foundation project: <http://dumps.wikimedia.org/>, English Wikipedia dumps in SQL and XML: <http://dumps.wikimedia.org/enwiki/>, etc.

Another example of freely available data for research is the World Bank data, i.e., the World Bank Databank ([http:// databank.worldbank.org/data/databases.aspx](http://databank.worldbank.org/data/databases.aspx)), which provides over 40 databases, such as Gender Statistics, Health Nutrition and Population Statistics, Global Economic Prospects, World Development Indicators and Global Development Finance, and many others. Most of the databases can be filtered by country/region, series/topics or time (years and quarters). In addition, tools are provided to allow reports to be customized and displayed in table, chart or map formats.

6.4 Data access via tools

As discussed, most commercial services provide access to social media data via online tools, both to control access to the raw data and increasingly to monetize the data.

6.4.1 Freely accessible sources

Google with tools such as Trends and Insights is a good example of this category. Google is the largest ‘scraper’ in the world, but they do their best to

'discourage' scraping of their own pages. (For an introduction of how-to surreptitious scrape Google—and avoid being 'banned'—see <http://google-scraper.squabbel.com>.) Google's strategy is to provide a wide range of packages, such as Google Analytics, rather than from a researchers' viewpoint the more useful programmable HTTP-based APIs.

Trends displays a particular search term, in this case 'libor.' Using Google Trends, you can compare up to five topics at a time and also see how often those topics have been mentioned and in which geographic regions the topics have been searched for the most.

6.4.2 Commercial sources

There is an increasing number of commercial services that scrape social networking media and then provide paid-for access via simple analytics tools. (The more comprehensive platforms with extensive analytics are reviewed in Sect. 8.) In addition, companies such as Twitter are both restricting free access to their data and licensing their data to commercial data resellers, such as Gnip and DataSift. Gnip is the world's largest provider of social data.

GNIP was the first to partner with Twitter to make their social data available, and since then, it was the first to work with Tumblr, Foursquare, WordPress, Disqus, StockTwits and other leading social platforms. Gnip delivers social data to customers in more than 40 countries, and Gnip's customers deliver social media analytics to more than 95 % of the Fortune 500. Real-time data from Gnip can be delivered as a 'Firehose' of every single activity or via PowerTrack, a proprietary filtering tool that allows users to build queries around only the data they need.

PowerTrack rules can filter data streams based on keywords, geo boundaries, phrase matches and even the type of content or media in the activity. The company then offers enrichments to these data streams such as Profile Geo (to add significantly more usable geo data for Twitter), URL expansion and language detection to further enhance the value of the data delivered. In addition to real-time data access, the company also offers Historical PowerTrack and Search API access for Twitter which give customers the ability to pull any Tweet since the first message on March 21, 2006.

Gnip provides access to premium (Gnip's 'Complete Access' sources are publishers that have an agreement with Gnip to resell their data) and free data feeds (Gnip's 'Managed Public API Access' sources provide access to normalized and consolidated free data from their APIs, although it requires Gnip's paid services for the Data Collectors) via its dashboard.

Different types of feeds serve different types of use cases and correspond to different types of queries and API endpoints on the publisher's source API. After selecting the feed, the user is assisted by Gnip to configure it with any required parameters before it begins collecting data. This includes adding at least one rule. Under 'Get Data' –['Advanced Settings' you can also configure how often your feed queries the source API for data (the 'query rate'). Choose between the publisher's native data format and Gnip's Activity Streams format (XML for Enterprise Data Collector feeds).

6.4.3 Data feed access via APIs

For researchers, arguably the most useful sources of social media data are those that provide programmable access. Google Trends AI & Soc (2015) 30:89–116 95 123 APIs, typically using HTTP-based protocols. Given their importance to academics, here, we review individually wikis, social networking media, RSS feeds, news, etc.

6.4.4 Wiki media

Wikipedia (and wikis in general) provides academics with large open-source repositories of user-generated (crowdsourced) content. What is not widely known is that Wikipedia provides HTTP-based APIs that allows programmable access and searching (i.e., scraping) that returns data in a variety of formats including XML. In fact, the API is not unique to Wikipedia but part of MediaWiki's (<http://www.mediawiki.org/>) open-source toolkit and hence can be used with any MediaWiki-based wikis. The wiki HTTP-based API works by accepting requests containing one or more input arguments and returning strings, often in XML format, that can be parsed and used by the requesting client. Other formats supported include JSON, WDDX, YAML, or PHP serialized. Details can be found at: <http://en.wikipedia.org/w/api.php?action=>

query&list=allcategories&acprop=size&acprefix=hollywood &format=xml. The HTTP request must contain:

- a) the requested 'action,' such as query, edit or delete operation;
- b) an authentication request; and

c) any other supported actions. For example, the above request returns an XML string listing the first 10 Wikipedia categories with the prefix 'hollywood.' Vaswani (2011) provides a detailed description of how to scrape Wikipedia using an Apache/PHP development environment and an HTTP client capable of transmitting GET and PUT requests and handling responses.

6.4.5 Social networking media

As with Wikipedia, popular social networks, such as Facebook, Twitter and Foursquare, make a proportion of their data accessible via APIs. Although many social networking media sites provide APIs, not all sites (e.g., Bing, LinkedIn and Skype) provide API access for scraping data. While more and more social Fig. 3 Gnip Dashboard, Publishers and Feeds 96 AI & Soc (2015) 30:89–116 123 networks are shifting to publicly available content, many leading networks are restricting free access, even to academics. For example, Foursquare announced in December 2013 that it will no longer allow private check-ins on iOS 7, and has now partnered with Gnip to provide a continuous stream of anonymized check-in data. The data is available in two packages: the full Firehose access level and a filtered version via Gnip's PowerTrack service. Here, we briefly discuss the APIs provided by Twitter and Facebook.

CHAPTER 7

FETCHING YOUTUBE DATA USING API

7.1 Google API

Google APIs are application programming interfaces (APIs) developed by Google which allow communication with Google Services and their integration to other services. Examples of these include Search, Gmail, Translate or Google Maps. Third-party apps can use these APIs to take advantage of or extend the functionality of the existing services.

The APIs provide functionality like analytics, machine learning as a service (the Prediction API) or access to user data (when permission to read the data is given). Another important example is an embedded Google map on a website, which can be achieved using the Static maps API, Places API or Google Earth API.

Google APIs are application programming interfaces developed by Google which allow communication with Google Services and their integration to other services. Examples of these include Search, Gmail, Translate or Google Maps.

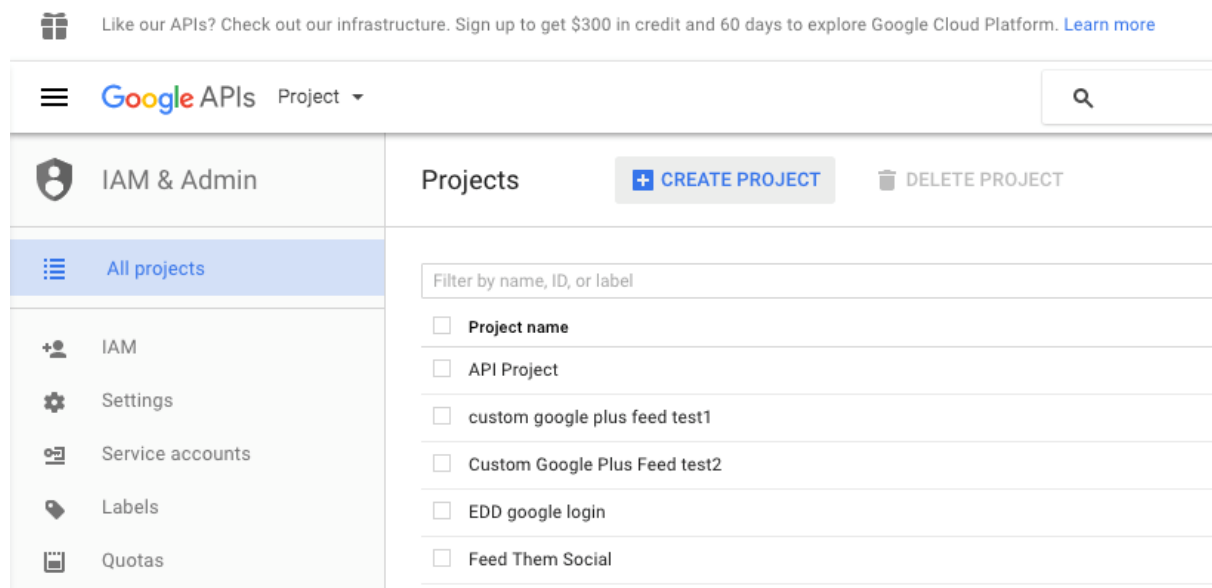
Google Apps Script is a cloud-based JavaScript platform which allows developers to write scripts only owner can manipulate API services such as Calendar, Docs, Drive, Gmail, and Sheets and easily create Add-Ons for these services with chromium-based applications.

Some **Google APIs** charge for usage, and you need to enable billing before you can start using these **APIs**. Some **APIs** allow **free** usage up to a courtesy usage limit, and in some cases this **free** limit is increased when you enable billing.

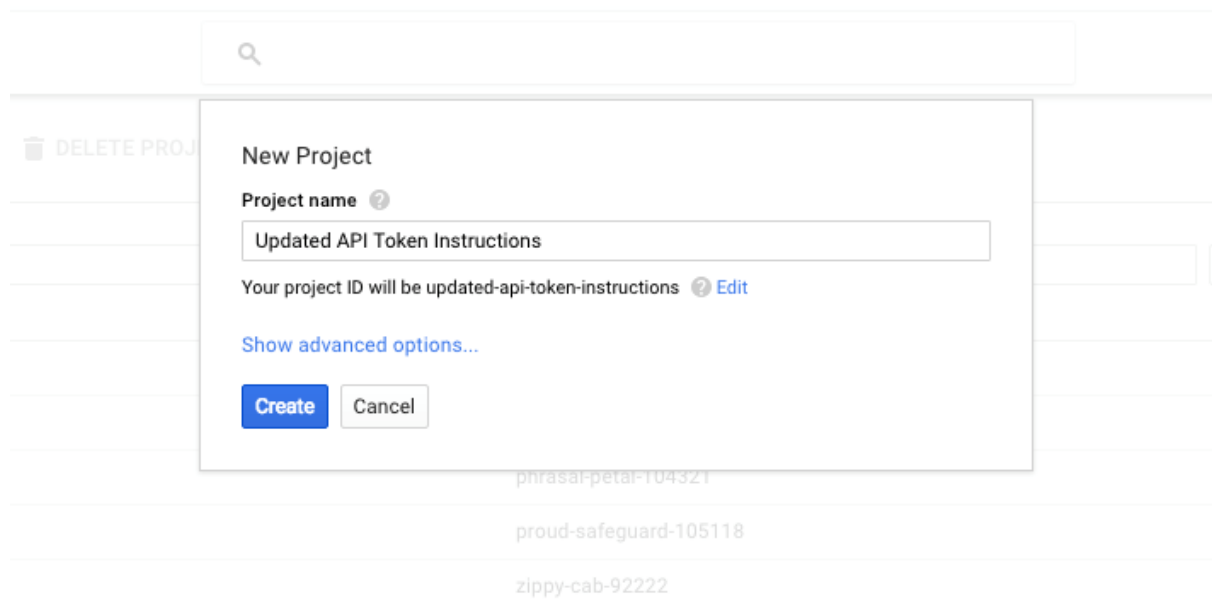
7.2 Getting API and Scrapping Data

- 1.Go to <https://developers.google.com/> and log in or create an account, if necessary.

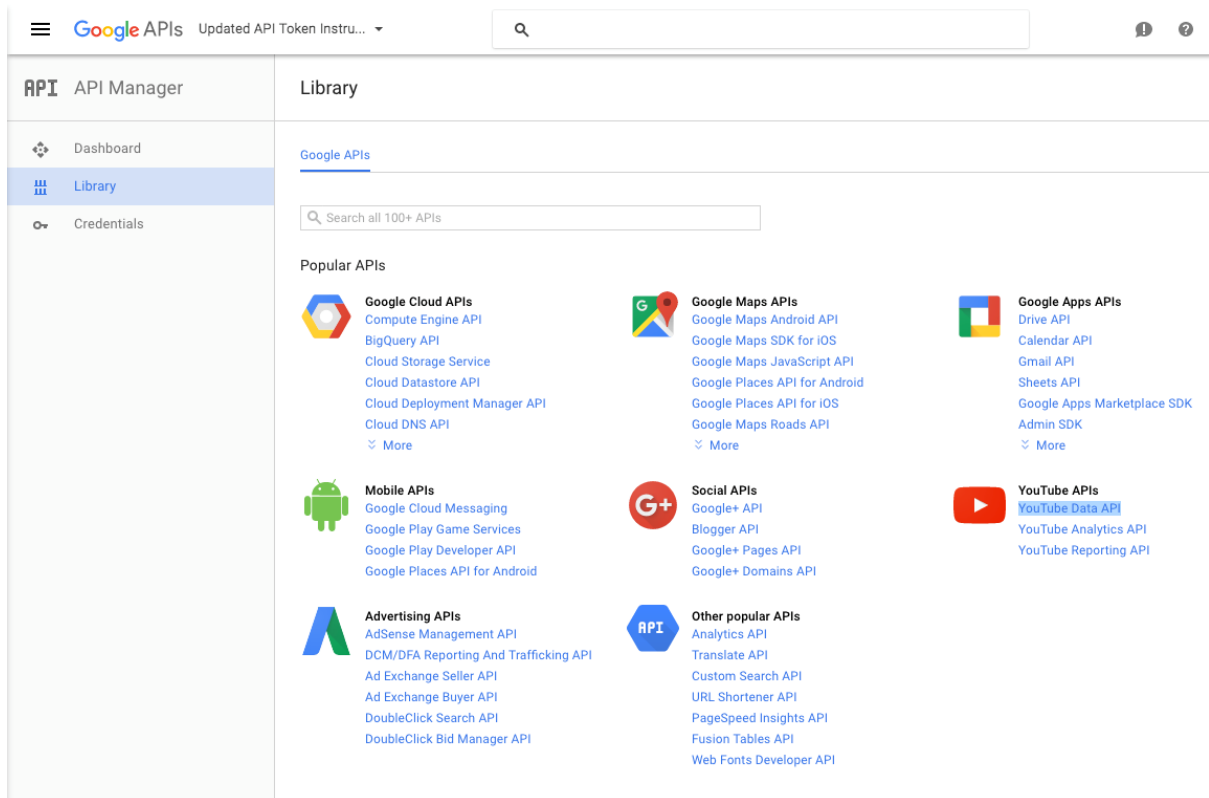
- 2.After logging in go to this link <https://console.developers.google.com/project> and click on the blue CREATE PROJECT button as depicted in the photo below. Wait a moment as google prepares your project.



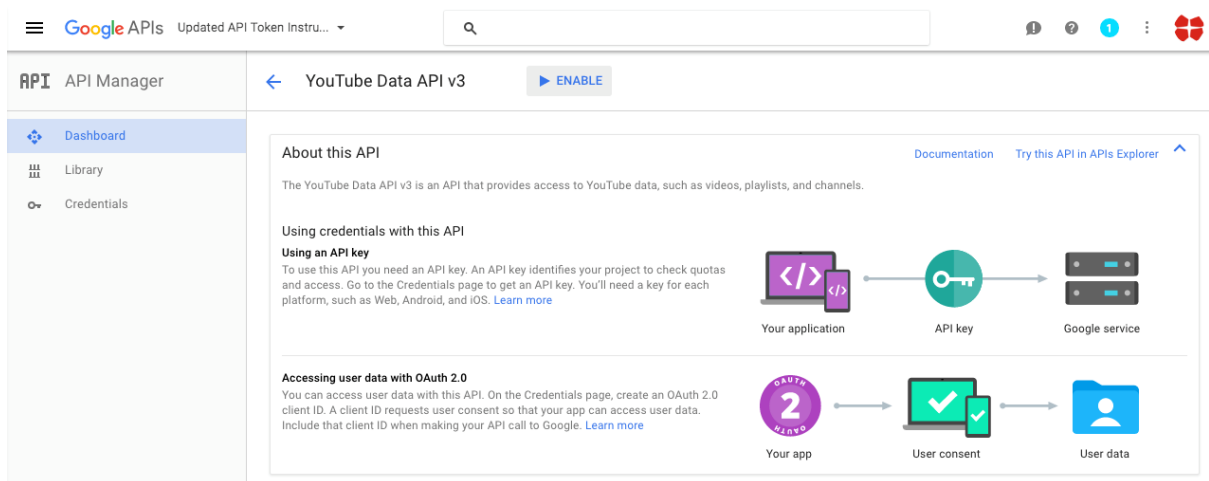
3. Fill in whatever Project Name you want.



4. Then click GoogleAPIs link in the top left corner and then click the link option called "YouTube Data API." It's under YouTube API's. You can see it highlighted in the photo below, bottom right.



5. Now click on the “ENABLE” button.



6. Next click on the blue ‘Go to Credentials’ button to the right.

Google APIs Updated API Token Instru...

API Manager

- Dashboard
- Library
- Credentials

YouTube Data API v3 ■ DISABLE

⚠ This API is enabled, but you can't use it in your project until you create credentials. Click "Go to Credentials" to do this now (strongly recommended). [Go to Credentials](#)

[Overview](#) [Quotas](#)

About this API [Documentation](#) [Try this API in APIs Explorer](#) ▼

All API versions ▼ All API credentials ▼ Selected 30 methods ▼

1 hour 6 hours 12 hours 1 day 2 days 4 days 7 days 14 days **30 days**

Traffic By response code ▼

Requests/sec (5 min average)

There is no data for this API in this time span

7. Choose the select option YouTube Data API v3 for the first select option and Web server(e.g. node js. Tomcat) for the second selection. Then choose Public data. Now click the blue button, "What credentials do I need?."

Google APIs Updated API Token Instru...

API Manager

- Dashboard
- Library
- Credentials**

Credentials

Add credentials to your project

1 Find out what kind of credentials you need

We'll help you set up the correct credentials
If you wish you can skip this step and create an [API key](#), [client ID](#), or [service account](#)

Which API are you using?
Determines what kind of credentials you need.

YouTube Data API v3 ▼

Where will you be calling the API from?
Determines which settings you'll need to configure.

Web server (e.g. node.js, Tomcat) ▼

What data will you be accessing?

☒ **Public data**
Access publicly available data provided by the API

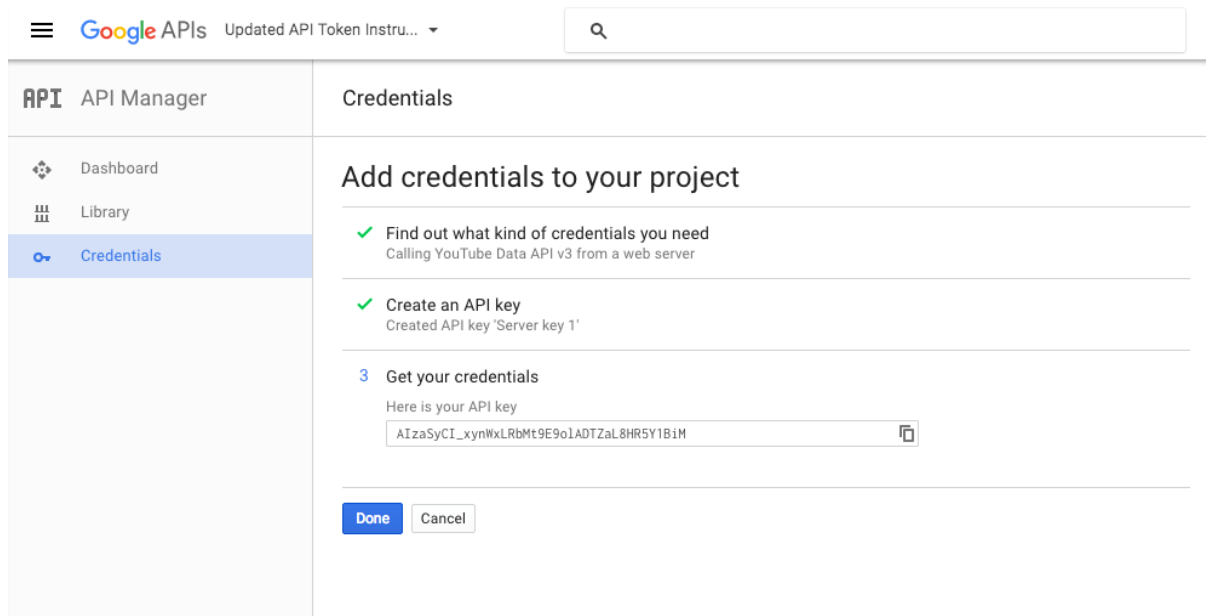
☐ **User data**
Access data belonging to a Google user, with their permission

[What credentials do I need?](#)

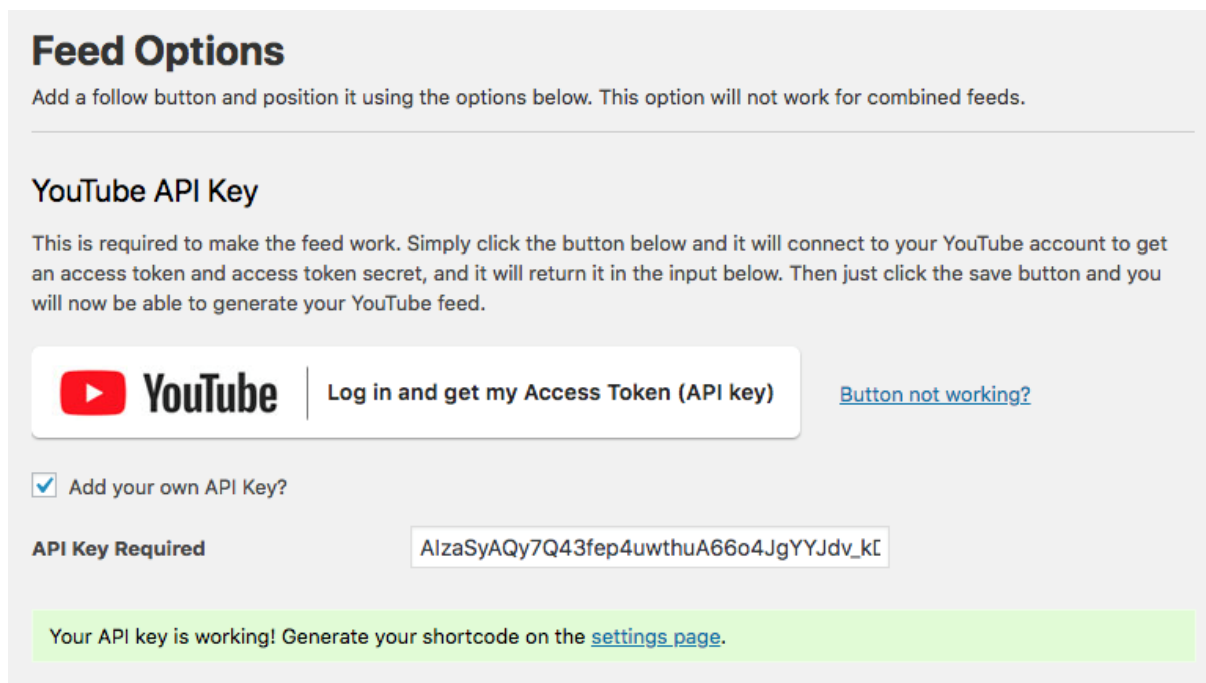
2 Get your credentials

[Cancel](#)

8. Almost done, wait for google to create your new project and you should see the screen below where you can copy your API Key.



9. Paste the API Key in our YouTube Options menu page as depicted below.



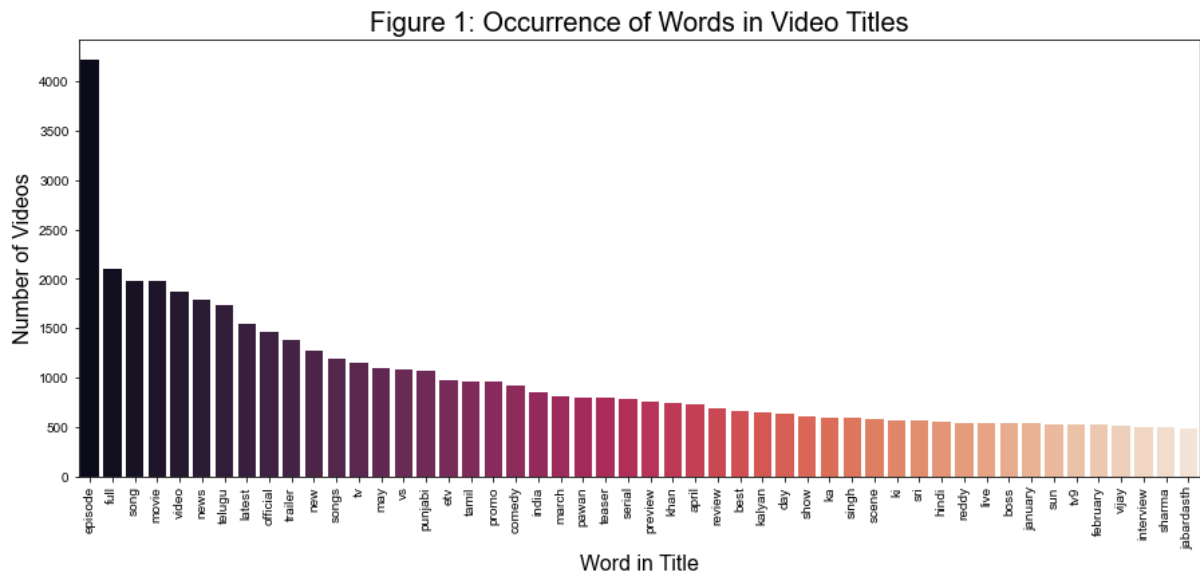
10 Store it in a secure file (api.text). Call the text file in the python program then run the program now we get the some of the countries YouTube trending videos data. For getting these data, first ensure the country code no saved in separate text file.

CHAPTER 8

YOUTUBE DATA ANALYSIS

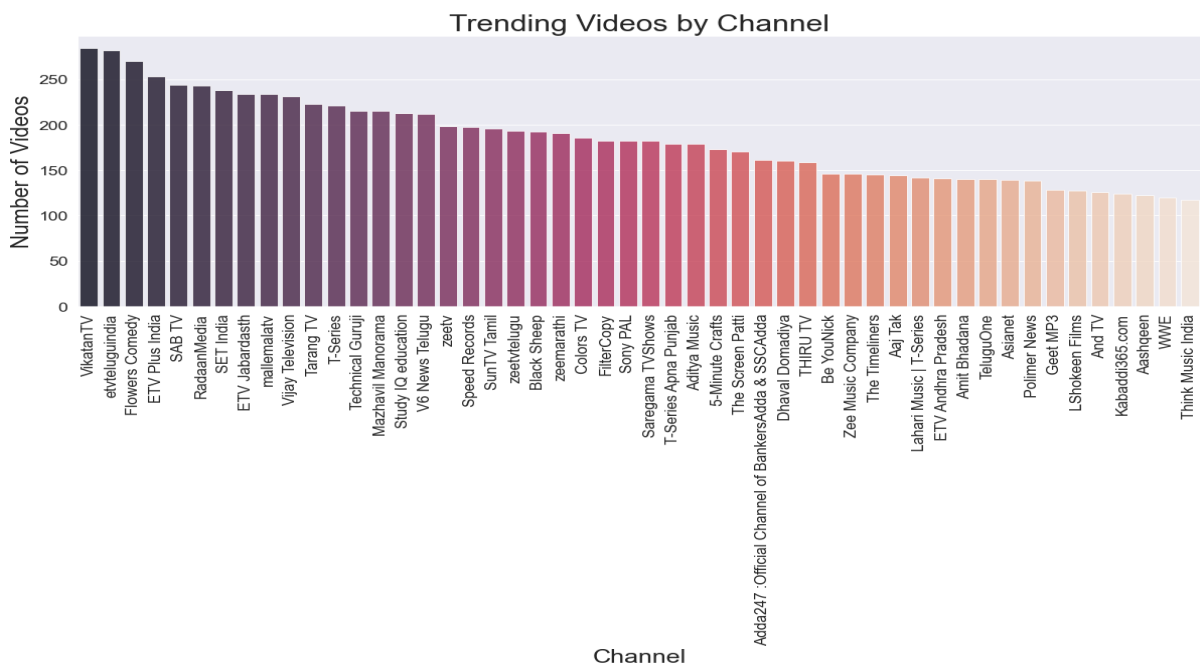
Analysis the YouTube trending videos data (India) by using jupyter notebook.

8.1 Occurrence of Words in Video Titles



8.2 Trending video by channel

Histogram for number of channels vs number of trending videos below 20



SOURCE CODE

```
import requests, sys, time, os, argparse

# List of simple to collect features
snippet_features = ["title",
                    "publishedAt",
                    "channelId",
                    "channelTitle",
                    "categoryId"]

# Any characters to exclude, generally these are things that become problematic in CSV files
unsafe_characters = ['\n', '"]']

# Used to identify columns, currently hardcoded order
header = ["video_id"] + snippet_features + ["trending_date", "tags", "view_count", "likes", "dislikes",
                                             "comment_count", "thumbnail_link", "comments_disabled",
                                             "ratings_disabled", "description"]

def setup(api_path, code_path):
    with open(api_path, 'r') as file:
        api_key = file.readline()

    with open(code_path) as file:
        country_codes = [x.rstrip() for x in file]

    return api_key, country_codes

def prepare_feature(feature):
    # Removes any character from the unsafe characters list and surrounds the whole item in quotes
    for ch in unsafe_characters:
        feature = str(feature).replace(ch, "")
    return f'"{feature}"'

def api_request(page_token, country_code):
    # Builds the URL and requests the JSON from it
    request_url = f"https://www.googleapis.com/youtube/v3/videos?part=id,statistics,snippet{page_token}chart=mostPopular"
    request = requests.get(request_url)
    if request.status_code == 429:
        print("Temp-Banned due to excess requests, please wait and continue later")
```

```

if request.status_code == 429:
    print("Temp-Banned due to excess requests, please wait and continue later")
    sys.exit()
return request.json()

def get_tags(tags_list):
    # Takes a list of tags, prepares each tag and joins them into a string by the pipe character
    return prepare_feature("|".join(tags_list))

def get_videos(items):
    lines = []
    for video in items:
        comments_disabled = False
        ratings_disabled = False

        # We can assume something is wrong with the video if it has no statistics, often this means it has been c
        # so we can just skip it
        if "statistics" not in video:
            continue

        # A full explanation of all of these features can be found on the GitHub page for this project
        video_id = prepare_feature(video['id'])

        # Snippet and statistics are sub-dicts of video, containing the most useful info
        snippet = video['snippet']
        statistics = video['statistics']

        # This list contains all of the features in snippet that are 1 deep and require no special processing
        features = [prepare_feature(snippet.get(feature, "")) for feature in snippet_features]

        # The following are special case features which require unique processing, or are not within the snippet
        description = snippet.get("description", "")
        thumbnail_link = snippet.get("thumbnails", dict()).get("default", dict()).get("url", "")
        trending_date = time.strftime("%y.%d.%m")
        tags = get_tags(snippet.get("tags", ["none"]))
        view_count = statistics.get("viewCount", 0)

        # This may be unclear, essentially the way the API works is that if a video has comments or ratings disab
        # then it has no feature for it, thus if they don't exist in the statistics dict we know they are disable

```

```

# then it has no feature for it, thus if they don't exist in the statistics dict we know they are disabled
if 'likeCount' in statistics and 'dislikeCount' in statistics:
    likes = statistics['likeCount']
    dislikes = statistics['dislikeCount']
else:
    ratings_disabled = True
    likes = 0
    dislikes = 0

if 'commentCount' in statistics:
    comment_count = statistics['commentCount']
else:
    comments_disabled = True
    comment_count = 0

# Compiles all of the various bits of info into one consistently formatted line
line = [video_id] + features + [prepare_feature(x) for x in [trending_date, tags, view_count, likes, dislikes,
comment_count, thumbnail_link, comments_disabled, ratings_disabled, description]]

lines.append(",".join(line))
return lines

def get_pages(country_code, next_page_token="&"):
    country_data = []

    # Because the API uses page tokens (which are literally just the same function of numbers everywhere) it is much
    # more inconvenient to iterate over pages, but that is what is done here.
    while next_page_token is not None:
        # A page of data i.e. a list of videos and all needed data
        video_data_page = api_request(next_page_token, country_code)

        # Get the next page token and build a string which can be injected into the request with it, unless it's None
        # then let the whole thing be None so that the loop ends after this cycle
        next_page_token = video_data_page.get("nextPageToken", None)
        next_page_token = f"&pageToken={next_page_token}&" if next_page_token is not None else next_page_token

        # Get all of the items as a list and let get_videos return the needed features
        items = video_data_page.get('items', [])
        country_data += get_videos(items)

```

```

next_page_token = video_data_page.get("nextPageToken", None)
next_page_token = f"&pageToken={next_page_token}&" if next_page_token is not None else next_page_token

# Get all of the items as a list and let get_videos return the needed features
items = video_data_page.get('items', [])
country_data += get_videos(items)

return country_data

def write_to_file(country_code, country_data):

    print(f"Writing {country_code} data to file...")

    if not os.path.exists(output_dir):
        os.makedirs(output_dir)

    with open(f"{output_dir}/{time.strftime('%y.%d.%m')}_{country_code}_videos.csv", "w+", encoding='utf-8') as file:
        for row in country_data:
            file.write(f"{row}\n")

def get_data():
    for country_code in country_codes:
        country_data = [", ".join(header)] + get_pages(country_code)
        write_to_file(country_code, country_data)

if __name__ == "__main__":

    parser = argparse.ArgumentParser()
    parser.add_argument('--key_path', help='Path to the file containing the api key, by default will use api_key.txt')
    parser.add_argument('--country_code_path', help='Path to the file containing the list of country codes to scrape')
    parser.add_argument('--output_dir', help='Path to save the outputted files in', default='output/')

    args = parser.parse_args()

    output_dir = args.output_dir
    api_key, country_codes = setup(args.key_path, args.country_code_path)

    get_data()

```

PRIVACY/ETHICS CONSIDERATION

This Trending YouTube Video Statistics dataset is a public dataset from Kaggle built using YouTube's API Services. Developers have permission to use data collected from YouTube's API Services as long as they follow the policies listed in the YouTube API Services Policies document. Overall, online video research does raise valid ethical and data privacy concerns regarding informed consent. However, the unique potential of such a dataset outweighs the cost since this report is not for a commercial purpose and in our analysis, we have minimized the use of unique identifiers of the YouTube Video by using a bag-of-word model to analyse common terms in the title, tags, and description.

CONCLUSIONS AND DISCUSSION

Do find a positive correlation between controversial of a video and viewer engagement in categories concerning politics and social issues. However, there are limitations of this project due to the incompleteness of dataset, confounding variables in real world scenario, and assumptions to simplify a complicated problem. We will handle this potential issue by recognizing the impact of the bias in the conclusion and discussion of our results.

As for keywords in video titles. Although we found that there are some correlations between keywords and the number of views, the impact is not completely clear. We observed that in some category some words dominate the context of video title, but videos whose title contains some of those top words does not mercenarily become more popular than those who do not.

Since the Trending YouTube Video dataset includes only top trending videos, there is a potential for survivor bias in the results. Videos with similar characteristics, that did not trend, are not included in the dataset and thus will not be studied.

We assume that video statistics such as views are independent of category, title, and many other variables presented in this dataset. Each view count is assumed to have equal weights. However, we do not consider the replay ability of a video because it cannot be distinguished from the dataset. In addition, we assume that there

is no underlying association between individual ratings of a video. However, ratings in this dataset can inaccurately reflect the hidden distribution as comment and rating bots might be involved.

For convenience, we assume that frequent words in each category has higher impact on the number of views. However, some less frequent words that appear in only a few videos might also have great impact (by the assumption of tf-idf). In our study, we only focus on more general cases that covers the majority videos of the category.

Regarding future work, our group will try more advanced techniques on video statistics analysis, such as regression models and learning algorithms. We will also employ more language models in the realm of Natural Language Processing, such as word2vec, topic models and tied. Besides working on numerical and string-based data, we will extend this analysis to image data by encoding thumbnails. As for content-based analysis, we could obtain newspaper headlines from other dataset or using topic model to study the effect of trending news on trending videos.

REFERENCE

- [1] YouTube Statistics,
<http://www.youtube.com/yt/press/statistics.html>
- [2] YouTube data API documentation
<https://developers.google.com/youtube/2.0/reference>
- [3] Yessenov K, Misailovic S (2009) Sentiment analysis of movie review comments, pp 1–17. <http://people.csail.mit.edu/kuat/courses/6.863/report.pdf>.
- [4] University of Edinburgh, pp 1–74.
<http://homepages.inf.ed.ac.uk/miles/mscprojects/wolfram.pdf>.
- [5] Wikimedia Foundation (2014) Wikipedia:Database download.
http://en.wikipedia.org/wiki/Wikipedia:Database_download.
- [6] ["What is structured data?-Definition from WhatIs.com"](#).
WhatIs.com. Retrieved 2016-12-06.
- [7] [https://github.com/meyhaa/trendingyoutubevideos/blob/master/full trending video analysis.ipynb](https://github.com/meyhaa/trendingyoutubevideos/blob/master/full_trending_video_analysis.ipynb)
- [8] <https://www.kaggle.com/ammarr111/youtube-trending-videos-analysis>
- [9] Trending-YouTube-Scraper
<https://github.com/mitchelljy/Trending-YouTube-Scraper>
- [10] <https://github.com/rahulanand1103/Web-Scraping-Youtube>

