COP5725 DATABASE MANAGEMENT SYSTEMS

FALL 2019

Instructor: Dr. Markus Schneider

TA: Kyuseo Park

DATABASE SOFTWARE PROJECT ON ANALYSIS OF YOUTUBE

TRENDING VIDEOS

PROJECT DELIVERABLE-1

**GROUP NUMBER: 14**

1. 1914 5414 - VENKATA SINDHU KANDULA
2. 7179 9872 - SAI PUJITHA SELVAKUMAR
3. 9693-6544 - VEERA VENKATA SHIVA TEJA NEMALA
4. 1733-6715 - SPANDANA EDUPUGANTI

# CONTENTS

1. Introduction
2. Data Requirements
3. Requirement Analysis

3.1 Functions of web-based interface

3.2 Details of real-world data used

3.3 Source for the real-world data

3.4 Colloquial Queries important for the application

3.5 Technical Aspects

1. INTRODUCTION

YouTube is the most used and most popular video platform in the world today. A list of trending videos are updated in YouTube constantly. For this project, a dataset with data about trending videos collected for about 205days over five different countries USA, Great Britain, Germany, Canada and France is taken for analysis.

This project deals with designing and implementing a database software system to analyze the data of more than 40,000 trending videos over different countries.

This project is aimed to provide an understanding to people or channels to increase popularity of their videos or channels and to identify the contribution and pattern by statistically analyzing different demographic factors like region, time of upload, trends, category of the video etc.,

1. Data Requirements

The database for this management system will include the following data:

* Titles information – This includes titles for various trending videos across various countries.
* Channels Information – This includes Channels which uploaded the collected trending videos.
* Category Information – This includes data about the various Categories of the collected trending videos.
* Titles information – This includes titles for various trending videos across various countries.
* Views Information – This includes total numbers of views for the trending videos at the time of data collection.
* Likes information – This includes total numbers of likes at the time of data collection.
* Dislikes information – This includes total numbers of dislikes at the time of data collection.
* Comments count Information – This includes total numbers of comments for the trending videos at the point of data collection.
* Ratings disabled information – This includes whether the ratings for the video is disabled or not.
* Removed Videos information – This indicates whether the trending video is removed either by the user or by the YouTube admin.

1. Requirement Analysis
   1. What are the main functions that the web-based user interface should provide?
2. Preference of country (USA, Germany etc.,) – the user can choose one among the five countries for the statistical analysis of the trending videos.
3. Preference of channel (ESPN, Netflix etc.,) – The data base web application evaluates the factors that effect the trending pattern of videos and displays the detailed view with respect to the channel desired.
4. Interested categories (Music, News, Sports, etc.,) – The aim is to provide what kind of category of videos to put in order to attract the subscribers or viewers.
5. Graphical representation of the trend analysis based on the data deducted from the different factors that affect users’ interest in making the videos trending.
   1. Details of real-world data used

Real-world dataset used for this project includes several months of data on daily trending YouTube videos. Data is included for the USA, Great Britain, Germany, Canada, and France regions with up to 200 listed trending videos per day.

This dataset comprises of statistical information on views, comments, likes, dislikes, channel title, video title, published time, errors, category of the trending videos in the YouTube for the period (2017 - 2018).

* 1. Source for the real-world data

Dataset for the project is procured from an opensource data provider platform in the internet named Kaggle, which shares datasets to enable developers to conduct analysis and predictions based on this information.

Web Link: <https://www.kaggle.com/datasnaek/youtube-new>

* 1. Colloquial Queries important for the application

1. Which country has the most active users?
2. Co-relation of trending videos in different countries vs the number of days the video has been trending
3. Which channel has largest trending videos with respect to different countries? Also, what kind of category of videos to put in order to attract the subscribers or viewers. (based on category ID)
4. Which category (music, education, sports, news etc.,) is the most trending? Country-wise as well as world-wide?
5. Analysis on the videos that are the most listed on the trending list based on categories.
6. How does the trending of videos corelate with likes and comments ratio?
7. How many videos have been rendered controversial and removed after they became trending?
8. How is trending of videos related to views and comments? (analysis of controversial videos)
   1. Technical Aspects:

**3.5.1 Oracle DB**

Oracle database (Oracle DB) is a relational database management system (RDBMS) from the Oracle Corporation and is one of the widely used relational database systems. The database system we will be using is the CISE Oracle.

**3.5.2 JAVA**

JAVA is a widely used general-purpose programming language that is especially suited for Web development and can be embedded into HTML. All server-side code will be written in Java. The code will be hosted on Apache Tomcat in Windows Environment.

**3.5.3 APACHE Tomcat**

Apache Tomcat (also referred to as Tomcat Server) is an open source software which can provide a "pure Java" HTTP web server environment in which Java code can run.

**3.5.4 HTML**

Hypertext Markup Language (HTML) is the standard markup language for documents designed to be displayed in a web browser.

**3.5.5 JavaScript**

JavaScript (JS) is a lightweight, interpreted, or just-in-time compiled programming language with first-class functions.