Big Data report

team: Magiczny Krzysztof

Spaliński Krzysztof,
Kała Jakub,
Dyczko Michał,
Pawłowski Maciej

March 2020

1 Project description

Memelysis - Website for memes analysis, with respect to themes and formats and a place to find your favourite meme. Data sources will consist of: 9GAG, Reddit, Twitter, Memedroid.

1.1 Delivered product

User will be provided with a web application consisting of three main components - a dashboard, search engine and statistical analysis.

The first one will present the latest data obtained from the most popular portals providing user-generated content. Each presented object will be extended with information such as source, matched category and virality value. Instead of looking through dozens of websites one will be able to enjoy all the memes in one place. Moreover each user will have its own personalised content thanks to minimum virality level that can be set in dashboard.

The latter one will enable the client to extract memes on dedicated topic. In such case the user will have the access to the most well-received content from the most trendy websites. Furthermore they will be provided with memes sources, so they will have a possibility to contact with other viewers, the original poster and express their opinion.

Finally, with the statistical analysis provided, users might examine current trends. They can track their favourite content and observe others memes preferences. Such tool might broaden their taste in memes and invite them to a meme world they have not experienced before.

2 Data

2.1 Data Sources

2.1.1 9GAG

"9GAG is a Hong Kong–based online platform and social media website, which allows its users to upload and share "user-generated content" or other content from external social media websites. It was launched on July 1, 2008." [1]. Currently it is used by more than 200 million users a month.

[1] https://en.wikipedia.org/wiki/9GAG

2.1.2 Twitter

Twitter is an American microblogging and social networking service on which users post and interact with messages known as "tweets". Twitter was launched in July 2006. The service rapidly gained worldwide popularity. In 2012, more than 100 million users posted 340 million tweets a day. Users can group posts together by topic or type by use of hashtags – words or phrases prefixed with a "#" sign. Tweets include also a functionality of attaching an image.

2.1.3 Reddit

"Reddit is an American social news aggregation, web content rating, and discussion website. Registered members submit content to the site such as links, text posts, and images, which are then voted up or down by other members. Posts are organized by subject into user-created boards called "subreddits", which cover a variety of topics like news, science, movies, video games, music, books, fitness, food, and image-sharing. As of July 2019, Reddit ranks as the 5th most visited website in the U.S. and 13th in the world". [2]

2.1.4 Memedroid

Memedroid is also a website for user-generated content. It might be not as popular nowadays as it used to be in the past, however it still possesses its main feature, which is the ability to easily create your own memes and share them among others. Therefore Memedroid contains dozens of themes, formats which will be highly demanded while clustering, where variety of content improves the algorithm accuracy.

2.2 Data collecting

Besides memes themselves: tags, titles, urls and the popularity indicators will be extracted. Data will be extracted using Python libraries dedicated to scrapping (requests, BeautifulSoup) with the help of Reddit's and Twitter's API. The data, hoarded by web-crawlers and API's, will be in the form of JSON. Each sample will contain a few keys containing data in size varying up to a few kB.

[2] https://en.wikipedia.org/wiki/Reddit

	Title	Image	Url	Rating	Tags
Reddit	✓	✓	✓	Vote count and upvote rate	✗
9gag	✓	✓	✓	Upvote count	✓
Memedroid	✓	✓	✓	Likeability percentage	✓
Twitter	✓	✓	✓	Likes	✓

Table 1: Sources content

3 Architecture and tech stack

3.1 Architecture

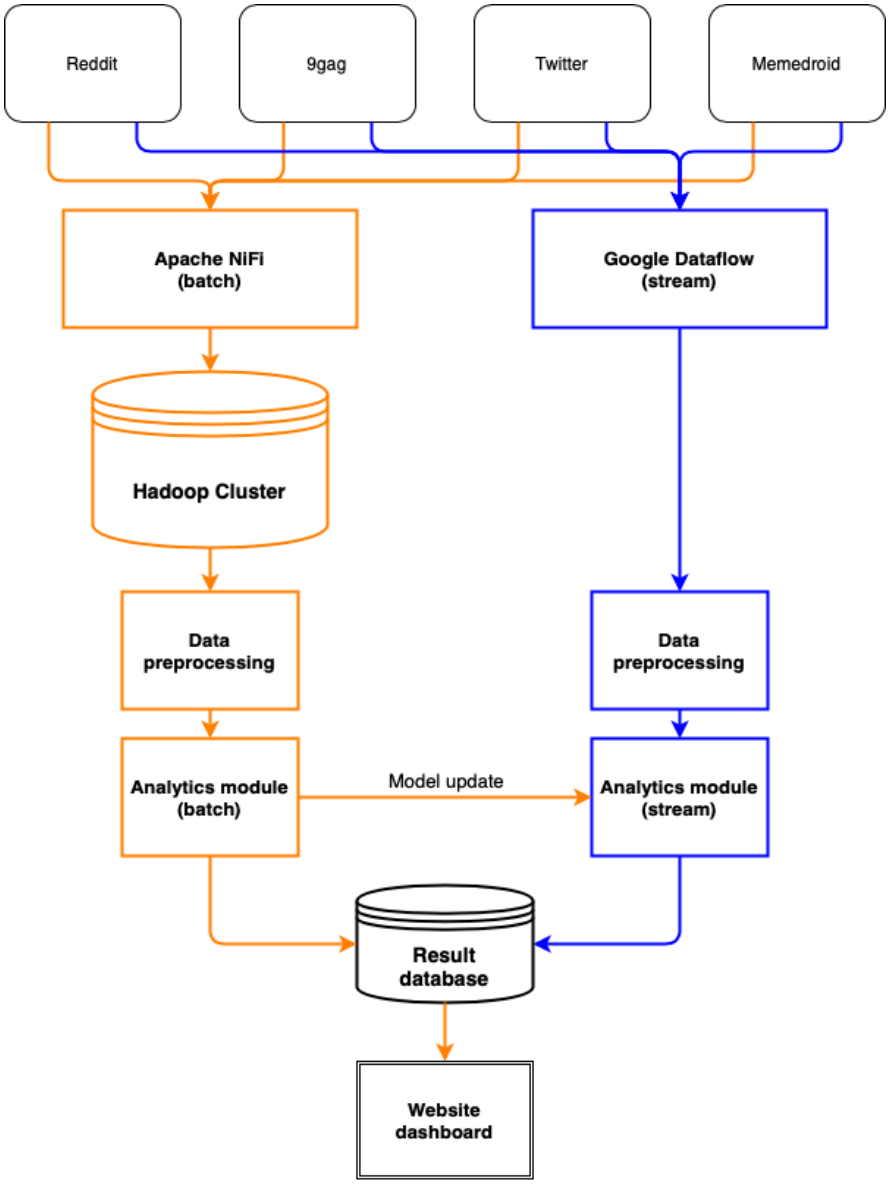

Figure 1: Project architecture

3.2 Tools

- Google Cloud Platform
- Apache NiFi
- Google Dataflow
- Hadoop Cluster
- Spark
- Python
- Django
- SQL (MariaDB)
- Google Cloud Vision

4 Data processing

4.1 Feature extraction

The main task is to extract text from each meme. To achieve such result Google Vision will be used. Furthermore, data will be cleaned in order to avoid inappropriate content which might badly affect the later analysis. Moreover all duplicates will be removed.

The whole process might take such form:

1. extract each desired feature from JSON items,
2. download pictures in according to given urls,
3. remove pictures not considered as memes
4. extract texts from pictures
5. remove potential duplicates
6. forward data to analytical segment

4.2 Data Analysis

The main task to be covered is to label each meme based on its content and format. To achieve such results OCRed content will be clustered and the most commonly used tags will be used as categories names. On the other hand Google Vision will perform clusterization of images to find the most popular formats for memes.

5 Tasks

The task to be performed within this project are:

Task	Michał	Jakub	Maciej	Krzysztof
Project Management				✓
Human Resources			✓	
Data collecting	9GAG	Twitter	Memedroid, 9GAG	Reddit
Cloud initialization	✓	✓	✓	✓
Apache NiFi, HDFS	✓		✓	
Google Dataflow		✓		✓
Data Preprocessing			✓	
Unit tests	✓	✓	✓	✓
Advanced Analytics	✓	✓	✓	✓
Final Webapp	✓			
Data Visualization				✓

Table 2: Tasks distribution