Memelysis

Analytics of memes

Michał Dyczko
Jakub Kała
Maciej Pawłowski
Krzysztof Spaliński
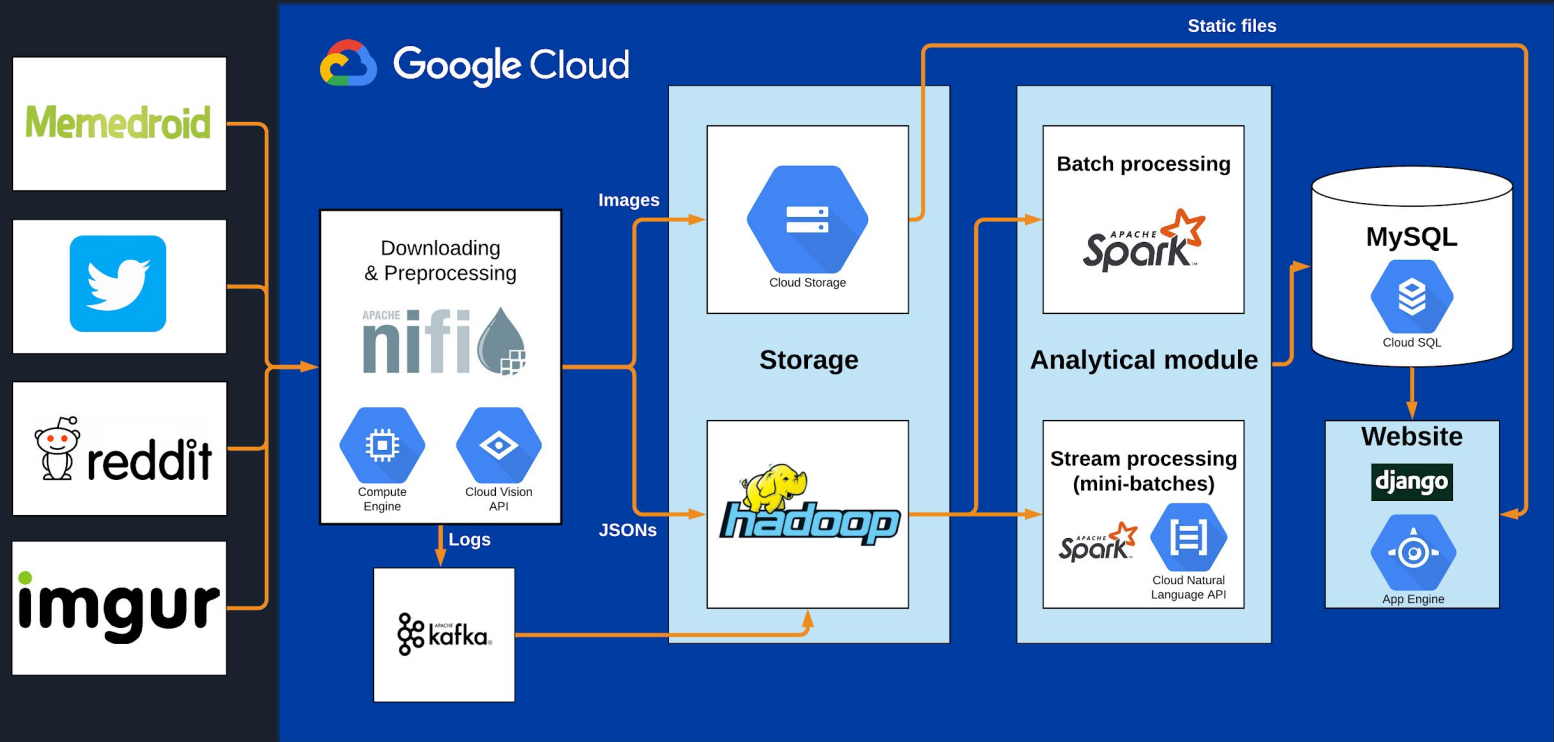
Project description

The goal of this project is to prepare web-based dashboard consisting of:

1. Live stream of memes from various sources:
   a. Memes are automatically classified by category;
   b. Each meme contains information about it's score against other memes from that source.
2. Analytics of meme categories and formats popularity.

# Data architecture

Data processing

Batch processing:

1. Cluster analysis model re-estimation (every 24 hours)

2. Re-estimation of virality factor fitting model

3. Trends analysis

Stream processing:

1. Extracting text from images (OCR)

2. Filtering explicit content

3. Category assignment for each incoming meme

4. **Virality factor** computation

APACHE **nifi**

Data collecting statistics

Aggregated meme volume over time
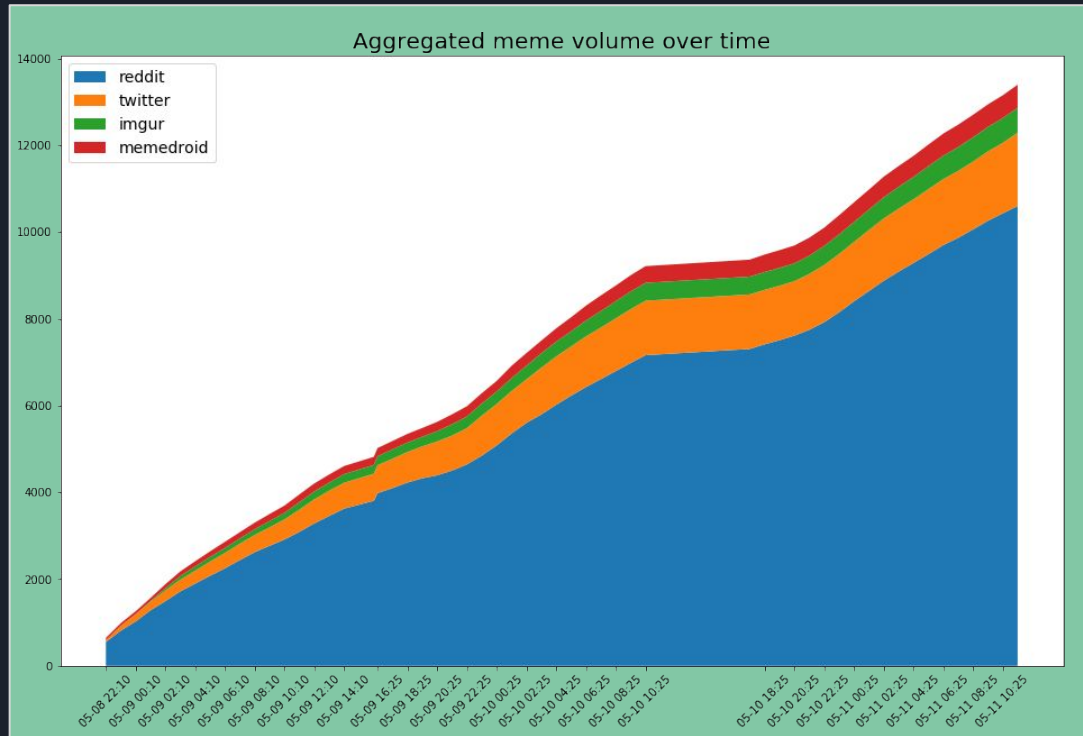
Analytical module
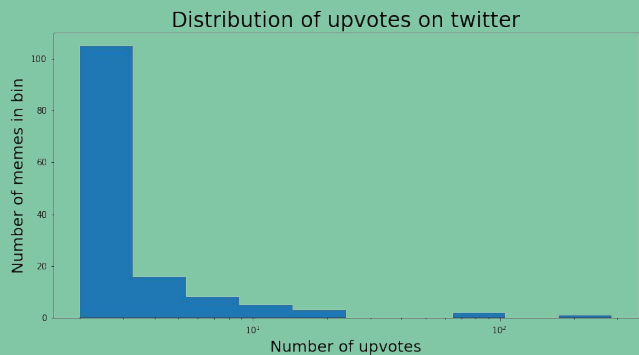
Optical Character Recognition

Clustering

Virality factor analysis

Upvote distribution on sources

Sample clustering

k=7, wn=50, Silhouette=0.50983127194143
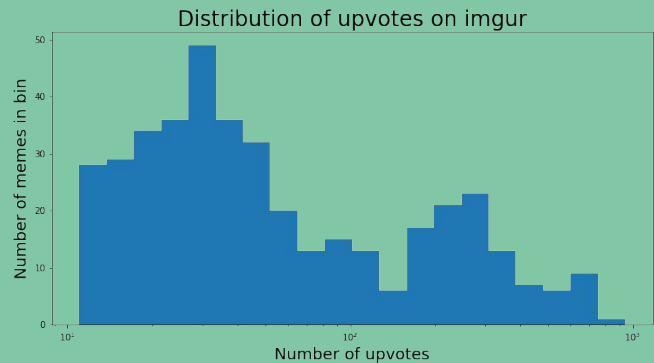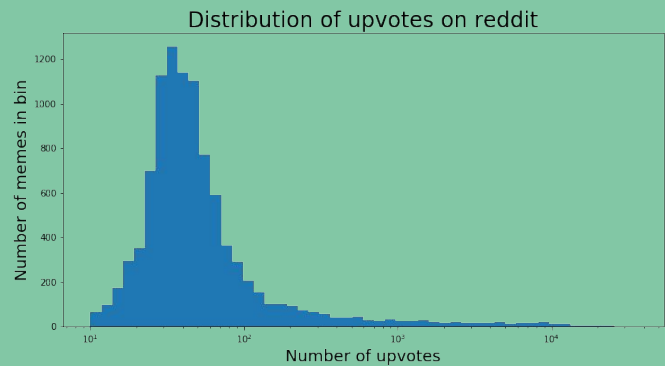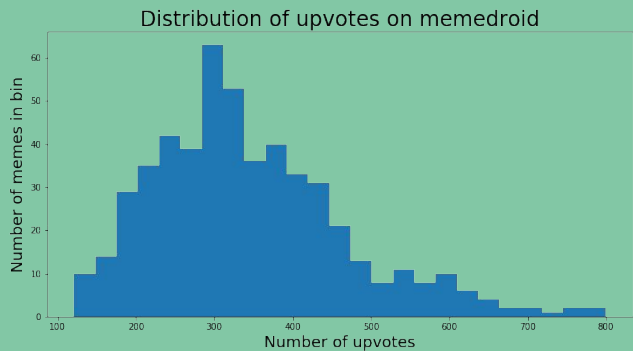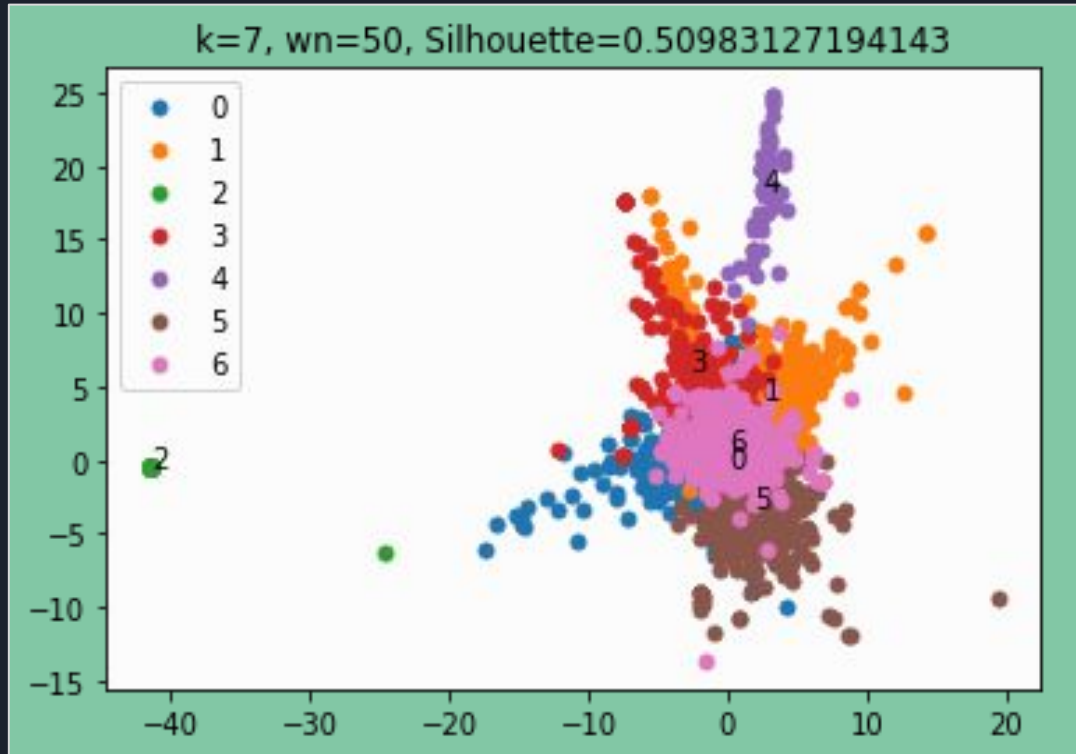
id	name
0	funny
1	rpg
2	-
3	-
4	minimumrequirements
5	cyberpunk2077
6	minecraft

Tools

Scraping:

- ❖ Python libs: BeautifulSoup 4, requests, Django 3;
- ❖ API's: Twitter API, Reddit API.

Cloud - GCP:

- ❖ Compute Engine running Nifi for scraping and image downloading;
- ❖ Dataproc: Hadoop and Spark;
- ❖ Google Storage for image hosting;
- ❖ App Engine running Django app.

Analytics:

- ❖ OCR - Google Cloud Vision API.

Product

Sample object

Click for more