Стандардна семинарска по предметот
Вовед во науката на податоците

Тема: YouTube Analyzer: Scraping YouTube data from channels. Analysis of popular channels, type of content posted, frequency of posting, user engagement, etc.

Студент:
Александар Стојанов 211067

Линк до кодот:
https://colab.research.google.com/drive/1AJruv_DmonYqHihUZgsl5HrzrReblnWP?usp=sharing

Линк до видеото:
https://www.youtube.com/watch?v=WB3sEJF4OMk

Содржина

1. Вовед

1.1Цел на проектот

Целта на овој проект е да се разгледаат податоците од YouTube за одреден канал и да се анализира популарен канал на оваа платформа. Исто во овој проект се анализира и ангажманот на корисниците. Со брзиот раст на YouTube како платформа за споделување содржини, важно е да се разберат трендовите кои можат да им помогнат на креаторите да ги оптимизираат своите стратегии за објавување и интеракција со публиката. Исто така можеме да направиме анализа и за расположението на корисниците на платформата Youtube во одредени периоди од денот во зависност од коментарите кои ги споделуваат на популарен канал.

1.2Опсег на проектот

Проектот е насочен кон три главни анализи:

Фреквенција на објавување на видеа по месец: Анализа на тоа во кој месец се објавуваат најмногу видеа.

Фреквенција на објавување на видеа по ден од неделата: Анализа на тоа во кој ден од неделата се објавуваат најмногу видеа.

Периоди од денот со најдобри и најлоши коментари: Идентификација на периодите од денот кога се објавуваат најпозитивните и најнегативните коментари на видеата.

Овие анализи ќе обезбедат увид во ангажираноста на креаторот на видеа на еден популарен канал, и кога е најдобро време за интеракција со публиката за да се добие што повеќе позитивни коментари.

2. Алатки кои се користат

- За реализација на проектот се користат следниве алатки:

Google Colab: Бесплатна онлајн платформа за пишување и извршување на Python код. Тоа е всушност хостирана верзија на Jupyter Notebook. Обезбедува бесплатен пристап до компјутерски ресурси, вклучително и графички процесор. Colab е особено погоден за машинско учење, наука за податоци и образование.

YouTube API: Интерфејс кој овозможува пристап до податоците од YouTube, како што се статистики за канали, информации за видеа и коментари. YouTube API овозможува автоматско преземање и обработка на голем обем на податоци директно од YouTube.

- Python библиотеки кои се користат:

pandas: Библиотека за манипулација и анализа на податоци. Обезбедува структури за податоци и алатки за работа со табеларни податоци.

matplotlib и seaborn: Библиотеки за визуелизација на податоци. Се користат за креирање графикони и визуелизации кои помагаат во интерпретацијата на резултатите од анализата.

transformers: Библиотека за обработка на природен јазик (NLP). Се користи за анализа на сентиментот на коментарите, односно за одредување дали коментарите се позитивни или негативни.

Овие алатки и библиотеки овозможуваат ефективно собирање, обработка и анализа на голем број податоци од YouTube, како и визуелизација на резултатите за полесно разбирање и донесување одлуки во зависност од резултатите.

3. Објаснување на кодот

3.1Собирање на податоци

Во овој проект се користи YouTube API за собирање на податоци од YouTube. YouTube API овозможува пристап до различни информации за каналите, видеата и коментарите на платформата. Тоа го постигнуваме со следните постапки од кодот:

- Со користење на YouTube API, се прибираат основните статистики за каналот, како што се бројот на претплатници, вкупниот број на прегледи и бројот на видеа.
- Се прибираат идентификаторите на сите видеа објавени на даден канал.
- Со користење на видео ID-овите, се добиваат детали за секое видео, како и коментари објавени на видеата.

Ова собирање на податоците се прави со помош на неколку функции:

- get_channel_statistics(channel_id): Преземање на статистики за даден канал.
- get_all_video_ids(channel_id): Добивање на листа на сите видео ID-а за даден канал.
- get_video_info(video_ids): Преземање на детални информации за секое видео.
- get_video_comments(video_id): Преземање на коментари за дадено видео (се земаат само по 100 коментари за секое видео).

3.2Обработка на податоците:

Со овој кој е испишан се добиваат следните податоци кои се складирани во табелата за видеа (video_info_df):

videoId: идентификатор на секое видео,
title: наслов на видеото,
description: опис на видеото напишан од креаторот на видеото,
publishedAt: датум и време на објавување на видеото во формат на datetime типот,
dayPublished: изведен податок од податокот publishedAt, во овој податок е прикажан денот на објавување на видеото (Monday, Tuesday…),
tags: тагови на видеото,
viewCount: колку вкупно прегледи има на видеото,

likeCount: на колку корисници вкупно им се допаѓа видеото,
comments: број на коментари на видеото,
month: изведен податок од податокот publishedAt, во овој податок е прикажан месецот во кој било објавено видеото.

Во табелата во која се складирани коментарите (comments_df), се следните податоци:

commentId: Идентификатор на секој коментар,
videoId: идентификатор на видеото на кое припаѓа одреден коментар,
author: автор на коментарот,
text: содржина на коментарот,
likes: на колку луѓе им се допаѓа коментарот кој е објавен,
publishedAt: датум и време на објавување на коментарот во формат на datetime типот,
periodOfDay: изведен податок од податокот publishedAt со помош на функцијата get_period(hour), во овој податок е прикажан периодот од денот во кој бил објавен коментарот (Morning, Afternoon, Evening, Night).

3.3Анализа на податоците

Со цел да направиме анализа на конкретен Youtube канал, во овој проект ќе го разгледаме каналот „Kurzgesagt – In a Nutshell“

Овој канал ги има следните општи статистики:

"viewCount": "2815035544", вкупно прегледи на сите видеа,

"subscriberCount": "22500000", вкупно претплатници на каналот,

"hiddenSubscriberCount": false, податок кој кажува дека бројот на претплатници на каналот не е скриен

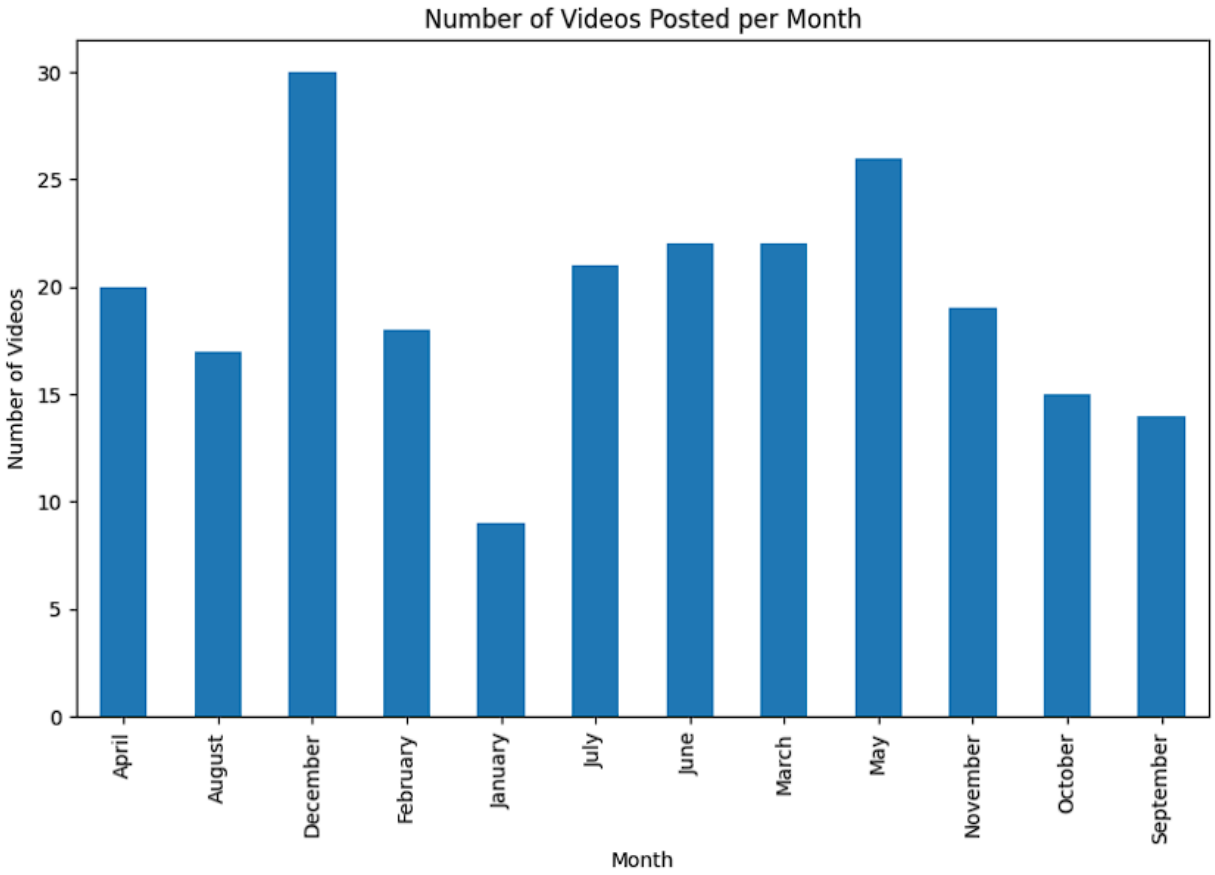
"videoCount": "235", вкупен број на видеа на каналот

3.3.1 Фреквенција на објавување на видеа

- Фреквенција на објавување според месец:

Податоците за објавувањето на видеата се групираат по месец за да се идентификува фреквенцијата на објавување. Овие податоци се визуелизираат со помош на библиотеките matplotlib и seaborn со бар графикони.

Според оваа анализа најголем број на видеа биле објавени во месец декември (30 видеа), а најмалку во месец јануари (9 видеа).

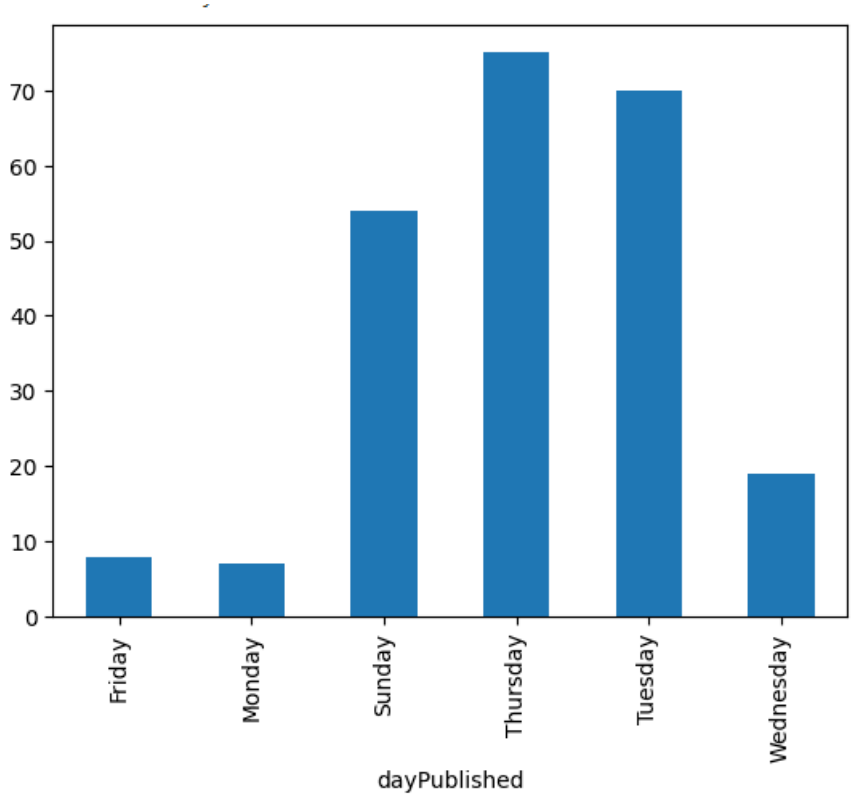


слика 1: Бар графикон за анализа на објавување на видеа по месец

- Фреквенција на објавување според ден во неделата:

Слично како и за групирањето на видеата по месец, така тие се групираат и по ден од неделата. Се користат истите библиотеки за визуелизација на податоците.

Според оваа анализа најголем број на видеа биле објавени во четврток (75 видеа), а најмалку во понеделник (7 видеа).



слика 2: Бар графикон за анализа на објавување на видеа по ден од неделата

.

3.3.2 Анализа на коментарите

- Категоризација на коментарите по сентимент (позитивни или негативни) користејќи ја библиотеката transformers:

Коментарите се анализираат за да се одреди нивниот сентимент користејќи напредни модели за обработка на природен јазик. Поради ограниченоста на моделот за обработка на природен јазик во врска со големината на текст на која моделот може да предвиде сентимент, содржината на коментарот ја делиме на токени и земаме само одреден број на токени (толку колку што може моделот да предвиде одеднаш).

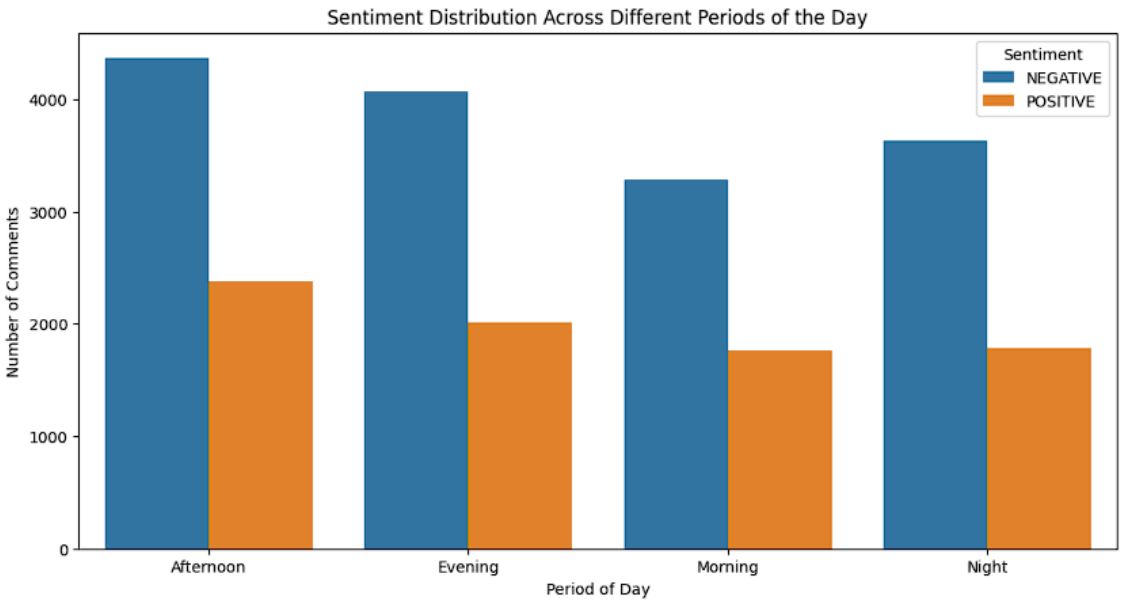
Доколку ги земеме сите сегменти и за сите сегменти предвидуваме сентимент, тоа би ни одзело премногу време. Со таа цел, во оваа анализа, сентиментот го предвидуваме само по првата поделба на содржината на коментарот по токени (првиот сегмент од токени). Тоа го правиме со функцијата def split_into_chunks(text, tokenizer, max_length) каде text е содржината на коментарот, tokenizer е моделот кој го претвора текстот во токени и max_length е големината на еден сегмент. Оваа функција како резултат го враќа сентиментот на овој дел од коментарот.

- Анализа на времето кога се објавуваат коментарите и идентификација на периодите со најмногу позитивни и негативни коментари:

Се анализира во кој период од денот (Morning/Наутро, Afternoon/Попладне, Evening/Навечер, Night/Во текот на ноќта) колку позитивни и негативни коментари биле оставени и кој е периодот со најпозитивните и најнегативните коментари. Ова се прави за да се идентификуваат оптималните времиња за интеракција со публиката.

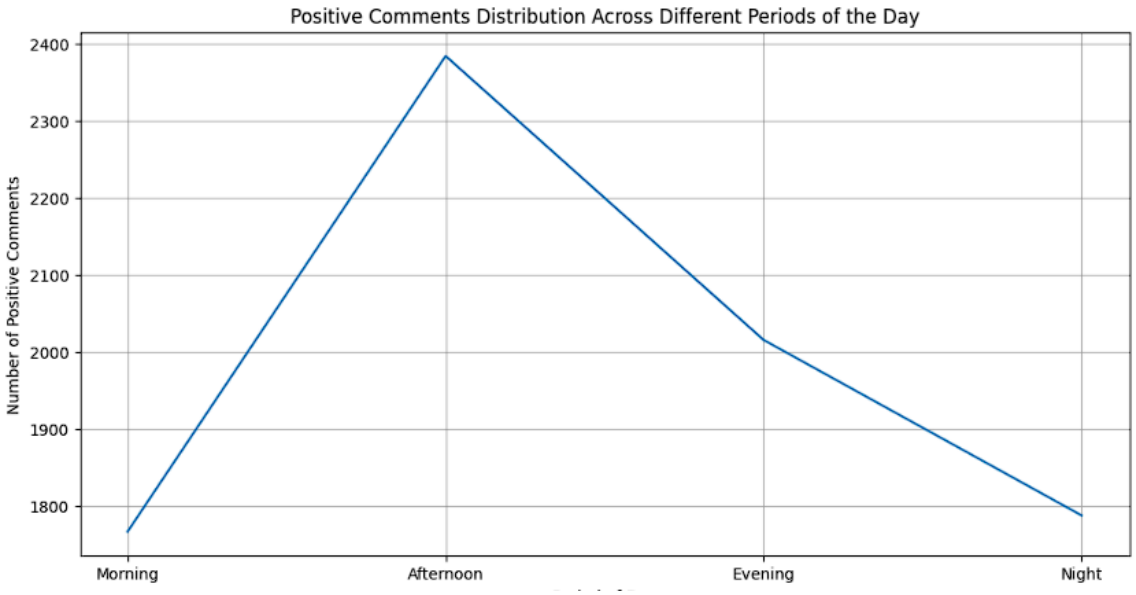
Се добива следната анализа:

	Позитивни	Негативни
Наутро	1766	3282
Попладне	2384	4369
Навечер	2015	4068
Во текот на ноќта	1787	3629



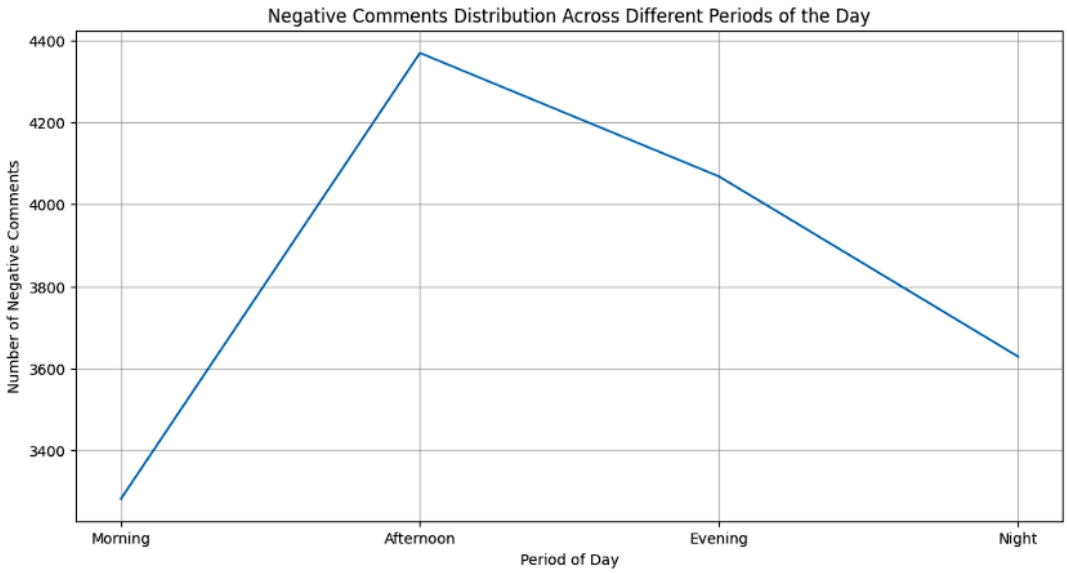
слика 3: Бар графикон за анализа на објавување на коментари во зависност од период во денот.

Според оваа анализа, луѓето оставаат најмногу позитивни коментари како и најмногу негативни коментари попладне.



слика 4: Line chart графикон за анализа на објавување на позитивни коментари во зависност од период во денот.

Според слика 4, можеме да забележиме дека имаме нагорна линија од наутро кон попладне за оставање на позитивни коментари, а потоа линијата се симнува надолу кога станува збор за периодот „Навечер“ подрастично надолу за периодот „Во текот на ноќта“.



слика 5: Line chart графикон за анализа на објавување на негативни коментари во зависност од период во денот.

Според слика 5, можеме да забележиме дека имаме нагорна линија од наутро кон попладне за оставање на негативни коментари, а потоа линијата благо се симнува надолу кога станува збор за периодот „Навечер“ и „Во текот на ноќта“.

4. Заклучок

Анализата на податоците од YouTube ни открива важни трендови во фреквенцијата на објавување на видеа и корисничкиот ангажман. Според добиените резултати од каналот „Kurzgesagt – In a Nutshell “, најголем број видеа се објавуваат во декември, а најмалку во јануари. Од аспект на денови во неделата, четврток е најактивен ден за објавување видеа, додека понеделник е најмалку активен. Ова укажува на тоа дека креаторите на содржини можат да ги планираат своите објави во тие периоди за поголема видливост и ангажман како избраниот канал.

Анализата на коментарите покажува дека позитивните коментари најчесто се објавуваат попладне исто како и негативните коментари. Ова е важен податок за креаторите на содржини бидејќи им помага да ги идентификуваат оптималните времиња за интеракција со својата публика.

Покрај тоа што е направено анализа на еден популарен канал, оваа анализа може да се примени и на останати канали и да се разгледаат добиените резултати за различни типови на канали. Исто така оваа анализа дава корисни увиди кои можат да им помогнат на креаторите на YouTube да ја подобрат својата стратегија за објавување и да го зголемат позитивниот ангажман со својата публика.