**Implementarea unui tool pentru colectarea meta-datelor video-urilor de pe Youtube**

**1 Introducere**

Majoritatea site-urilor moderne utilizeaza diverse metode pentru a incarca toate resursele necesare randarii unei pagini. Complexitatea acestei randari face imposibila obtinerea continutului printr-o simpla cerere HTTP. Din acest motiv, implementarea unui web-scraper pentru un astfel de site are nevoie de o modalitate de a incarca toate resursele site-ului. O solutie utilizata des este Selenium.

In timp ce Selenium nu este o biblioteca al carei scop principal este web-scraping-ul, faptul ca utilizeaza un browser real rezolva problema mentionata mai sus. In fapt, Selenium este o bilbioteca menita pentru testarea aplicatiilor web.

In tool-ul creat pentru aceasta lucrare folosim Selenium pentru a prelua, mai intai, o lista de link-uri catre video-urile unui canal ales, iar apoi, parcurgand aceasta lista, pentru a prelua meta-datele disponibile fiecarui video.

Principalul scop al acestui corpus este sa ajute in lucrari viitoare legate de procesarea limbajului natural, mai ales analiza trendurilor [[1](https://aclanthology.org/W17-5213.pdf)] si a sentimentelor [[2](https://aclanthology.org/D18-1394/)] (cel mai relevant aspect fiind limbajul utilizat in descriere si titlu relativ la numarul de view-uri si de like-uri).

In aceasta lucrare am realizat urmatoarele puncte:

* Implementarea unui tool de web-scraping pentru preluarea de meta-date a video-urilor de pe Youtube.
* Colectarea a 7 seturi de date pe canale cu o tematica diversa (canale de stiri / divertisment, canale de radio, canale de vlog)
* Colectarea de date despre video-uri in limba romana (format rar tratat in crearea de corpus-uri pentru lucrari de tip nlp)

## 2 Video meta-data

## In cautarea a diverse resurse pentru aceasta lucrare, am observat o lipsa de seturi de date in limba romana care sa aiba ca origine site-ul Youtube. [[3](https://aclanthology.org/)]

## De asemenea, majoritatea seturilor de date pe care le-am gasit constau in transcrieri [[4](https://aclanthology.org/D18-1394/)] ale video-urilor, nu date despre video-uri in sine (titlu, descriere etc.).

## Am ales acest format deoarece credem ca ar fi util pentru diferite tipuri de analiza. Consideram ca titlurile video-urilor, in mod special, au un format unic de limbaj formal (asemanatoare cu titlurile de stiri), sau informal.

## 3 Setul de date

## Am ales urmatoarele canale pentru formarea setului de date:

## Tvr (formatul este asemanator cu stirile televizate)

## Stirile ProTv (formatul este asemanator cu stirile televizate)

## Digi24HD (formatul este asemanator cu stirile televizate)

## Dorian Popa (Vlogger, stil informal)

## Selly (Vlogger, stil informal)

## Radio Zu (stil semi-formal)

## Kiss Fm (stil semi-formal)

## 4 Cod

## Lista a elemetelor relevante:

## LINKS – lista a argumentelor date script-ului

## CHANNEL\_NAMES – lista a numelor fiecarui canal (dedusa din LINKS)

## RESULT\_FILE\_NAMES – lista numelor fisierelor in care vor fi scrise rezultatele

## parametrul hard\_cap al functiei scroll\_to\_bottom – numarul de video-uri preluate de la un canal

## functia write\_video\_100 – functie care scrie rezultatele la fiecare 100 de video-uri preluate de script (admite modificari pentru a mari sau micsora intervalul la care se salveaza aceste date)

## functia get\_video\_info – functia principala a script-ului, aici este parcursa lista de link-uri si sunt preluate informatiile pentru fiecare video din aceasta.

## Implementarea noastra, fiind modulara (fiecare pas este izolat intr-o functie), este usor extensibila (optimizare cu multithreading, preluarea comentariilor pentru fiecare video, etc.). Procesul utilizat pentru a randa continutul unei pagini adauga un overhead considerabil, iar o integrare cu API-ul expus de Youtube ar reduce mult din acesta.

## 5 Referinte

* [1] - <https://aclanthology.org/W17-5213.pdf>
* [2] - <https://aclanthology.org/D18-1394/>
* [3] - <https://aclanthology.org/>
* [4] - <https://aclanthology.org/D18-1394/>