**Lab Exercise -1**: Data Extraction -**“Hashtag Campaigns”**

(S Divya Naimisha (17XJ1A0513) and Meda Sahithi(17XJ1A0521))

Our topic in mind for the extraction of data is the use of Hashtag for different Campaigns. The motivation behind choosing this topic is to know the reaction of people to different hashtags like **#BlacksLivesMatter, #metoo** and to know how the brands promote and make use different hashtags like **#WantAnR8 by Audi**.

To achieve the task we built a module that takes a list of hashtags and generates a .txt file containing all the captions of the posts posted using the given hashtags. The data source we used for extraction is Instagram and the tools used are beautiful soup library from python and selenium bot to have a visual of data that is being stored while the data is being scrapped. The reason behind using data source as Instagram is because of its popularity among the teens nearly 72% of teens uses Instagram. The other sources for extraction can be twitter and some blogs by people on the given moments.

**Extraction**:

The web page used for extraction is - “<https://www.instagram.com/explore/tags/BlackLivesMatter/>”.

(https://www.instagram.com/explore/tags/hashtag\_list[count]).

Links for the post on the page are scrapped using **selenium bot** and are store in Links.txt file.

The captions are collected for all the post’s links in the Link.txt file using **Beautiful Soup**, translated to English using **google translate API** and stored in theHashtag.xlsx file.

**Prons:**

* A large amount of posts available on Instagram helps get many captions.

**Cons:**

* The data has to be filtered by a person after collection because not all captions are relevant to the hashtags.

**About the extracted data:**

Captions for the hashtag: #food is scrapped and stored in the file. The scrapped data contains 885 captions with an average of 20 tokens per caption. The tokens are computed using TweetTokenizer from nltk.tokenize. The word types of each word are computed using the pos\_tag of nltk library.

**Problems faced and Learnings:**

* Thecaptions are written by people and people are prone to make grammatical errors. This could affect the performance of the google translate API and other functions used.
* Some captions involve a non-verbal form of communication like emoji.
* Irrelevance of the caption with the used hashtags.
* The provided a good hands-on experience working with unstructured data.