**Проект по Вовед во науката на податоците**

Сумаризација на твитови

1. Вовед

Во рамките на овој проект е разработена темата за сумаризација на твитови. Како првична фаза од проектот е собирање на твитови од различни теми, политика, астрономија, технологија и слично. Сите овие твитови понатаму се средуваат со цел моделот да може по добро да ги обработи, всушност се прави прочистување на податоци, бришење на непотребни карактери и слично. Во следната фаза твитовите се кластерираат со цел да се добие одредена претстава за тоа како твитовите се групирани. Со предходни истражувања од проекти од ваков тип и разгледување на различни можности за истото е реализирана последната фаза од проектот т.е. сумаризација на твитовите и споредување сличност помеѓу твитовите за евалуација на резултатите од моделите за сумаризација.

2. Подготовка на податоците

Подтотовка на податоците претставува прва фаза од целиот процес во рамките на проектот. Првично беа собрани 40-50 твитови од различни теми. Во рамките на твитовите постојат карактери и информации кои се речиси не потребни и додаваат дополнителна сложеност во целиот процес па потребно е истите да се исчистат од истите. Со користење на техниките за прочистување на податоците се заменуваат овие карактери со цел да се избришат истите и се пополнуваат сите празни делови во истите. Исто така во рамките на овој дел спаѓа и делот за токенизација на твитовите со цел да бидат подготвени за користење во моделите на сумаризација. Со користење на методите за токенизација во зависност од моделот (T5 Tokenizer, GPT Tokenizer и слично) се прави токенизација на карактерите така што карактерите од еден твит се претвараат во токени за да може моделот да ги обработи истите.

3. Кластерирање

Во рамките на оваа фаза од проектот се прави кластерирање на податоците со цел да се добие претстава за тоа како тие се организирани по категории. Со користење на KMeans методот за кластерирање се кластерираат твитовите. По добиениот резултат од овој процес се добива дека податоците се организирани во 3 категории што се поклопува со тоа како тие се организирани при собирање на истите.

4. Барање сличност

Сличноста помеѓу документите ја оценуваме со TextRank, со цел да видиме колку се разликуваат сумаризациите. Сличноста се движи од [0,1]. Ги броиме зборовите кои се слични(еднакви) во документите. Ја пресметуваме нивната норма, односно збир од нивни логаритми од должините на реченициите. На крај враќаме количник од бројот на исти зборови и нормата.

5. Споредба на резултатите за сличност

Два документи се споредуваат со нивните сумаризации, со секој од моделите. T5, GPT2, roBERTa, Xlnet. Се покажува дека T5 е добар сумаризатор, како и roBERTa.

6. Референци

https://anubhav20057.medium.com/step-by-step-guide-abstractive-text-summarization-using-roberta-e93978234a90
https://huggingface.co/transformers/main_classes/output.html
https://huggingface.co/transformers/model_doc/roberta.html
https://iq.opengenus.org/textrank-for-text-summarization/