IU000135 : Personalisation and Machine Learning - Mini Project

Content-based Movie Recommender with Salient Sentences Extracted by Bert

Xiaolin Deng - 21001319

📒Report

Project Rationale

The emergence of short-form videos significantly impacted human lives in several dimensions. People nowadays live fast-paced lives with the explosion of data. Therefore, the demands of text-summarization have notably arisen. Although human power is the most reliable resource for text-summarization, the needs are way beyond the supply. Computers have long been considered to be unable to perform complex and creative tasks. Text generation is an example of an exciting new wave of deep learning-based methods which threaten to challenge this belief. It's a field that has an easy to interpret output with straightforward applications. With the ongoing rise of streaming services, consumers have a much larger pool of movies to select from. In order to identify which movies are worth watching, it is important to have quality movie trailers. However, trailer creation requires substantial resources.

Project Description

In this project, we propose a method to assist movie trailer creation by automatically recommending the most salient sentences from the original film.We were able to successfully perform text recommendation tasks using a pre-trained BERT model for the task of movie trailer generation. By implementing the DistilBERT on our dataset, we were able to get the rouge-4 score from an average of 0.0062, although a better measure of the movie trailer's quality would be human testers.

Document Instruction

1. Data collection (DC|Marvel) / (movie|trailer)_corpus

```
    a) MarvelMovieCrawl.ipynb Crawling for Marvel movie titles
    b) Download the subtitles for the corresponding movie manually.srt
    c) Search for each movie trailer in turn from Youtube, then copy and
paste the automatically generated Youtube subtitle text into CSV
```

2. Data preprocess

```
    a) CSV2Trailer.ipynb
    b) DataReader.ipynb pre-processes the raw data into formats
```

3. Run Preprocess.ipynb to preprocess the subtitles.

4. Run Testing.ipynb for model prediction and evaluation

Adjusted pretrained model, encoder, etc.

The adujusted pretrained model and generated files was not fully placed in the repository because it was too large,Download those files from links,please.

1. bert-extractive-summarization
2. PreSumm

Movie Data Source

Dataset
Marvel
DC
Youtube
OpenSubtitles

Our datasets comprise a total of 49 movies from Marvel and DC, along with their corresponding trailers. We initially attempted to collect data using regular expressions. However, due to the complexity of variable download resources and anti-crawler protections on the subtitles websites, we ended up manually collecting the movie and trailer subtitles from the internet.

There are 73892 sentences from the movie and a total of 7927 sentences from trailers. At the beginning of each sentence, we add an integer to annotate how many seconds elapsed from the start of the movie. For example, suppose the sentence “This is the Asgardian”(from Avengers: Infinity War) appears right at the movie's beginning. After preprocessing, it would become “0 This is the Asgardian”.

Library used:

- `pandas`
- `numpy`
- `pytorch`
- `scikit learn`
- `nltk`
- `bert-extractive-summarization`

Original Pretrained Bert

1. stanfordnlp
2. bert-extractive-summarizer