NLP For Economists

Text Summarization

Sowmya Vajjala

Munich Graduate School of Economics - LMU Munich

Guest Course, October 2020

Goals for this session

- What is text summarization?
- How is it potentially relevant for you?
- What are some existing solutions?

source material: small subtopic in Chapter 7 from `https://practicalnlp.ai`

What is text summarization?

- Text summarization refers to the task of creating a summary of a longer piece of text.
- The goal of this task is to create a coherent summary that captures the key ideas in the text.
- It's useful to do a quick read of large documents, to extract key phrases/information etc.

Extractive vs Abstractive Summarization

- Extractive summarization refers to selecting important sentences from a piece of text and showing them together as a summary.
- Abstractive summarization refers to the task of generating an abstract of the text; i.e., instead of picking sentences from within the text, a new summary is generated.

- Extractive is relatively simpler to implement/use with given SOTA.

Query-focused versus query-independent summarization

- query-focused summarization refers to creating the summary of the text depending on the user query
- query-independent summarization creates a general summary.

- query independent is relatively easily usable with existing python libraries.

Single versus multi-document summarization

- single-document summarization is the task of creating a summary from a single document
- multi-document summarization creates a summary from a collection of documents.

- again, single document is a more commonly usable solution

Using a summarizer

- Research in this area has explored rule-based, supervised, and unsupervised approaches and, more recently, DL-based architectures.
- However, popular extractive summarization algorithms used in real-world scenarios use a graph-based sentence-ranking approach.
- Each sentence in a document is given a score based on its relation to other sentences in the text, and this is captured differently in different algorithms.
- The Top N sentences are then returned as a summary.
- TextRank is a well known algorithm.

Textrank to summarize a webpage

using the library sumy: `https://pypi.org/project/sumy/`

```
from sumy.parsers.html import HtmlParser
from sumy.nlp.tokenizers import Tokenizer
from sumy.summarizers.text_rank import TextRankSummarizer

url = "https://en.wikipedia.org/wiki/Automatic_summarization"
parser = HtmlParser.from_url(url, Tokenizer("english"))
summarizer = TextRankSummarizer()
for sentence in summarizer(parser.document, 5):
    print(sentence)
```

You can use plain text as well. (important point to note: textrank summarizer take longer time for longer texts!!)

Conclusion

- Summarization can potentially be useful to you as economists
- It is a very active area of research, I took a particular snapshot that is relatively readily usable for non-NLP researchers.
- Explore the library further if you want, and be aware that the standard disclaimer that these tools are not perfect holds!

What Next?

- Last item in this topic "NLP and Machine Learning Methods"!
- Check out the Assignment 3 description
- I will sit back and follow your discussions on NLP use in Economics research!