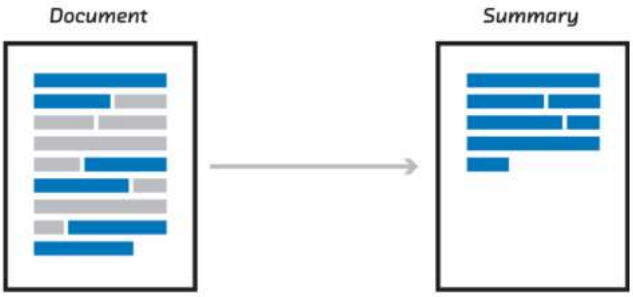
Eric Ondenyi · Follow
Dec 14, 2017 · 3 min read

Extractive Text Summarization Techniques With sumy



Extractive summarizers

Extractive text summarization techniques perform summarization by picking portions of texts and constructing a summary, unlike abstractive techniques which conceptualize a summary and paraphrases it .

Recently while I was researching I come across sumy by miso-belica which abstracts you from implementing this algorithms by yourself. From the legendary Luhn to Edmundson summarizer this library provides an easy way to perform extractive summarization as shown below.

This is was text summarization task on BBC news datasets [http://mlg.ucd.ie/datasets/bbc.html] while comparing the performance of this extractive algorithms.

First, begin by installing sumy

```
sudo pip install sumy
```

We will review all relevant abstractive summarizers and implement them using sumy. There are various techniques under abstractive summarization each technique is implemented differently based on researchers approaches, this techniques include clustering, graph theory, lexical chains, word-net etc, some are statistical in nature others deep rooted in linguistics while others robustly try to combine two or more techniques

```
#Plain text parsers since we are parsing through text
from sumy.parsers.plaintext import PlaintextParser


#for tokenization
from sumy.nlp.tokenizers import Tokenize
```

After importing relevant libs for our per-processing, we load the text file parse then tokenize it

```
#name of the plain-text file ~ bbc news dataset
file = "001.txt"
parser = PlaintextParser.from_file(file, Tokenizer("english"))
```

This a graphical based text summarizer

```
from sumy.summarizers.lex_rank import LexRankSummarizer
summarizer = LexRankSummarizer()

#Summarize the document with 2 sentences

summary = summarizer(parser.document, 2)

for sentence in summary:
 print(sentence)
```

Results from my notebook

```
In an effort to live up to its reputation in the 1990s as “an island of democracy”, the Kyrgyz President,
Askar Akaev, pushed through the law requiring the use of ink during the upcoming Parliamentary and
Presidential elections.
It dries and is not visible under normal light.
```

2. Luhn

It is one of the earliest suggested algorithm by the famous IBM researcher it was named after. It scores sentences based on frequency of the most important words.

```
from sumy.summarizers.luhn import LuhnSummarizer
summarizer_1 = LuhnSummarizer()
summary_1 =summarizer_1(parser.document,2)

for sentence in summary_1:
 print(sentence)
```

Results for Luhn Summarizer

```
The Kyrgyz Republic, a small, mountainous state of the former Soviet republic, is using invisible ink and
ultraviolet readers in the country's elections as part of a drive to prevent multiple voting.
In an effort to live up to its reputation in the 1990s as "an island of democracy", the Kyrgyz President,
Askar Akaev, pushed through the law requiring the use of ink during the upcoming Parliamentary and
Presidential elections.
```

3. LSA

Latent semantic analysis is an unsupervised method of summarization it combines term frequency techniques with singular value decomposition to summarize texts. It is one of the most recent suggested technique for summerization

```
from sumy.summarizers.lsa import LsaSummarizer
summarizer_2 = LsaSummarizer()
summary_2 =summarizer_2(parser.document,2)

for sentence in summary_2:
 print(sentence)
```

Results for LSA summarizer

```
This new technology is causing both worries and guarded optimism among different sectors of the
population.
```




Text rank is a graph-based summarization technique with keyword extractions in from document.

```
from sumy.summarizers.text_rank import TextRankSummarizer
summarizer_3 = TextRankSummarizer()
summary_3 =summarizer_3(parser.document,2)

for sentence in summary_3:
 print(sentence)
```

Results for text rank

```
In an effort to live up to its reputation in the 1990s as "an island of democracy", the Kyrgyz President,
Askar Akaev, pushed through the law requiring the use of ink during the upcoming Parliamentary and
Presidential elections.
The use of ink is only one part of a general effort to show commitment towards more open elections - the
German Embassy, the Soros Foundation and the Kyrgyz government have all contributed to purchase
transparent ballot boxes.
```

Conclusion

Sampling just a few. The results are reasonable and can be used by humans to generally understand long texts and their contents. The choice of the algorithm is now your choice

This has made it quite easy to summarize document but its also important for the engineer to understand the underlying statistics and mathematical implementation of each algorithm to see which one suites your task well.