Wikipedia Discovery

-using different text summarization-

Extractive Summarization

Iliescu Mihail Dorin

Problem statement

This project is aiming to create summary text based on a wikipedia article.

Motivations

Our motivations for this project were to implement and compare text summarization through different algorithms and state what could be the best ways to achieve text summarization.

Pipeline

Our program is defined and is excuted in this order:

1. Web crawler/scraper
2. Pre-processsing, tokenization
3. Algorithms computation
4. Results formationg
5. Results evaluation
6. Display output

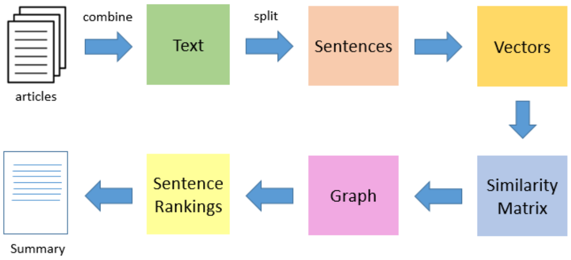
Proposed solution & Implementation

This section discuss about what methods we decided to use for our project and why.

- Cosine similarity

Cosine similarity through vector space model can be a way to find similarity in dataset. It means that if we can get the most similar sentences to the wole text, we can simply select the n most relevant sentences in the result set.

- TextRank



Derived from Google Page Rank algorithm [1], TextRank algorithm ranks parts of text. This ranking is defined by the number of relation between sentences. We base our evaluation on word frequency such as TF-IDF.

- K-Mean clustering

Wikipedia Discovery

-using different text summarization-

Extractive Summarization

K-Mean clustering algorithm [2] is an unsupervised classification algorithm frequently used in the world of Machine Learning and Data Science. Its main purpose is to, given a n dimension data set, be able to classify this data in categories (called clusters) according to them features. So using this algorithm, we can classify any type of data (images, text...).

Even K-Mean clustering has a data classification purpose, we decided to try to adapt its feature to text summarization. In our case, if we give as input our sentences from original text to K-Mean, it will classify our sentences by topics. Then we can pick the n most relevant sentences from each cluster/topic to form our summary.

I based my implementation on these two papers [3] and [4].

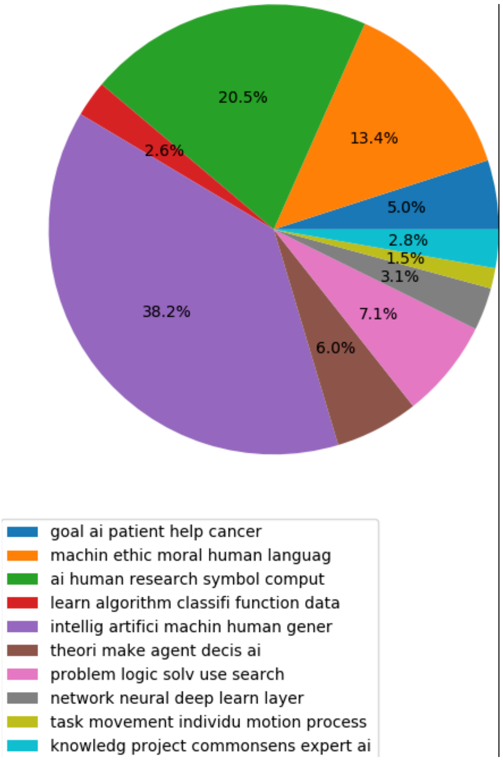


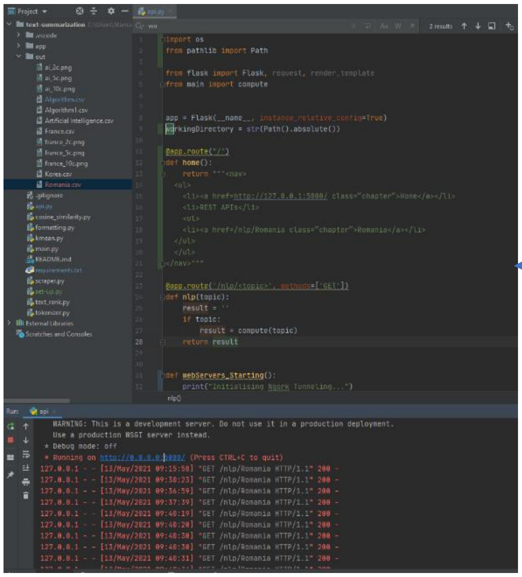
Chart showing data repartition of topic "Artificial Intelligence" over 10 clusters. Each category shows the top words in the cluster (note that words are stemmed).

To know what are the most important sentences of each cluster, we proceed as follow. After classifying input text in n clusters, we then pick one sentence from each cluster based on minimum distance from sentences of cluster. We apply this until we reach the number of sentences present in the reference summary

Wikipedia Discovery
-using different text summarization-

Extractive Summarization

Application

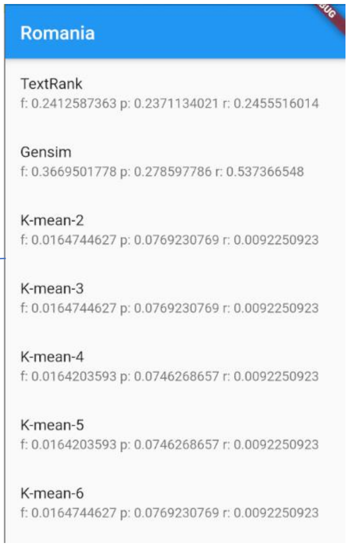


```python
import os
from pathlib import Path

from flask import Flask, request, render_template
from main import compute


app = Flask(__name__, instance_relative_config=True)
workingDirectory = str(Path().absolute())

@app.route("/")
def home():
    return """<nav>
  <ul>
    <li><a href=http://127.0.0.1:5000/ class="chapter">Home</a></li>
    <li>REST APIs</li>
    <ul>
    <li><a href=/nlp/Romania class="chapter">Romania</a></li>
  </ul>
  </ul>
</nav>"""

@app.route('/nlp/<topic>', methods=['GET'])
def nlp(topic):
    result = ''
    if topic:
        result = compute(topic)
    return result


def webServers_Starting():
    print("Initialising Ngork Tunneling...")
```

```
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: off
 * Running on http://0.0.0.0:5000/ (Press CTRL+C to quit)
127.0.0.1 - - [13/May/2021 09:15:58] "GET /nlp/Romania HTTP/1.1" 200 -
127.0.0.1 - - [13/May/2021 09:30:23] "GET /nlp/Romania HTTP/1.1" 200 -
127.0.0.1 - - [13/May/2021 09:36:59] "GET /nlp/Romania HTTP/1.1" 200 -
127.0.0.1 - - [13/May/2021 09:37:39] "GET /nlp/Romania HTTP/1.1" 200 -
127.0.0.1 - - [13/May/2021 09:40:19] "GET /nlp/Romania HTTP/1.1" 200 -
127.0.0.1 - - [13/May/2021 09:40:20] "GET /nlp/Romania HTTP/1.1" 200 -
127.0.0.1 - - [13/May/2021 09:40:30] "GET /nlp/Romania HTTP/1.1" 200 -
127.0.0.1 - - [13/May/2021 09:40:30] "GET /nlp/Romania HTTP/1.1" 200 -
127.0.0.1 - - [13/May/2021 09:40:31] "GET /nlp/Romania HTTP/1.1" 200 -
```

NGROK

Tunneling



Romania

TextRank
f: 0.2412587363 p: 0.2371134021 r: 0.2455516014

Gensim
f: 0.3669501778 p: 0.278597786 r: 0.537366548

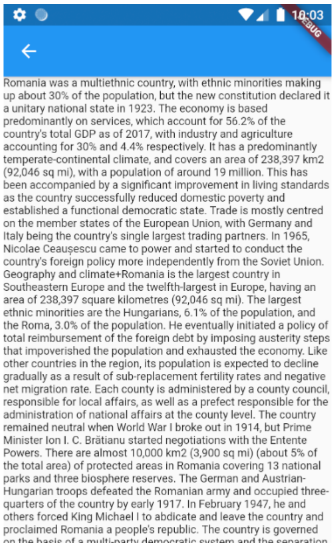
K-mean-2
f: 0.0164744627 p: 0.0769230769 r: 0.0092250923

K-mean-3
f: 0.0164744627 p: 0.0769230769 r: 0.0092250923

K-mean-4
f: 0.0164203593 p: 0.0746268657 r: 0.0092250923

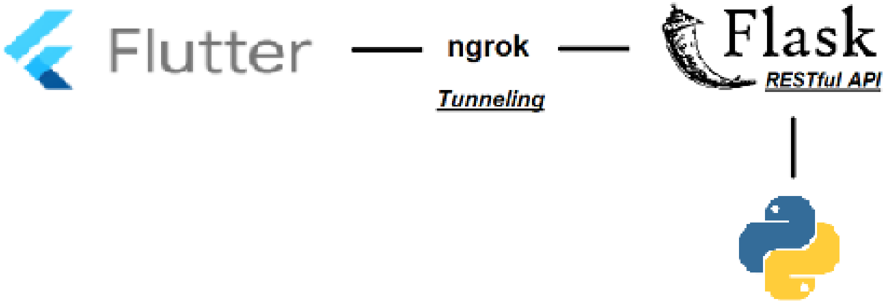
K-mean-5
f: 0.0164203593 p: 0.0746268657 r: 0.0092250923

K-mean-6
f: 0.0164744627 p: 0.0769230769 r: 0.0092250923



Romania was a multiethnic country, with ethnic minorities making up about 30% of the population, but the new constitution declared it a unitary national state in 1923. The economy is based predominantly on services, which account for 56.2% of the country's total GDP as of 2017, with industry and agriculture accounting for 30% and 4.4% respectively. It has a predominantly temperate-continental climate, and covers an area of 238,397 km2 (92,046 sq mi), with a population of around 19 million. This has been accompanied by a significant improvement in living standards as the country successfully reduced domestic poverty and established a functional democratic state. Trade is mostly centred on the member states of the European Union, with Germany and Italy being the country's single largest trading partners. In 1965, Nicolae Ceaușescu came to power and started to conduct the country's foreign policy more independently from the Soviet Union. Geography and climate+Romania is the largest country in Southeastern Europe and the twelfth-largest in Europe, having an area of 238,397 square kilometres (92,046 sq mi). The largest ethnic minorities are the Hungarians, 6.1% of the population, and the Roma, 3.0% of the population. He eventually initiated a policy of total reimbursement of the foreign debt by imposing austerity steps that impoverished the population and exhausted the economy. Like other countries in the region, its population is expected to decline gradually as a result of sub-replacement fertility rates and negative net migration rate. Each county is administered by a county council, responsible for local affairs, as well as a prefect responsible for the administration of national affairs at the county level. The country remained neutral when World War I broke out in 1914, but Prime Minister Ion I. C. Brătianu started negotiations with the Entente Powers. There are almost 10,000 km2 (3,900 sq mi) (about 5% of the total area) of protected areas in Romania covering 13 national parks and three biosphere reserves. The German and Austrian-Hungarian troops defeated the Romanian army and occupied three-quarters of the country by early 1917. In February 1947, he and others forced King Michael I to abdicate and leave the country and proclaimed Romania a people's republic. The country is governed on the basis of a multi-party democratic system and the separation

Backend (Python) **Frontend (Flutter)**



Frameworks used:

- numpy~=1.20.1
- pandas~=1.2.4
- scikit-learn~=0.24.1
- scipy~=1.6.2
- **gensim**~=3.8.3 - an open-source library for unsupervised topic modeling and natural language processing
- **nltk**~=3.6.1 -a leading platform for building Python programs to work with human language data
- **textblob**~=0.15.3 -part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.
- flask~=1.1.2
- **rouge**~=1.0.0
- Flutter (frontend)
- Python 3.9
- Ngrok.exe (tunneling on the web)

Wikipedia Discovery

-using different text summarization-

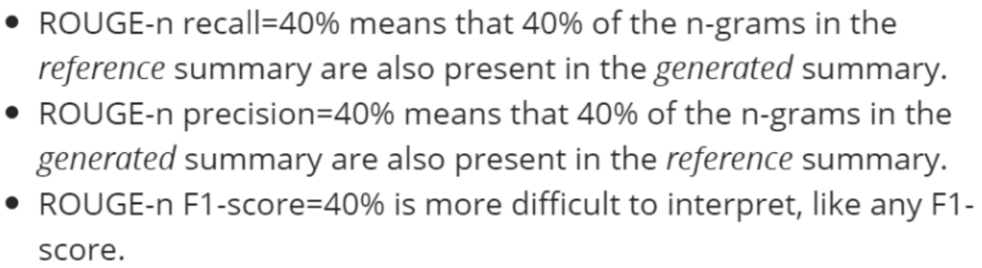
Extractive Summarization

Experiments and results

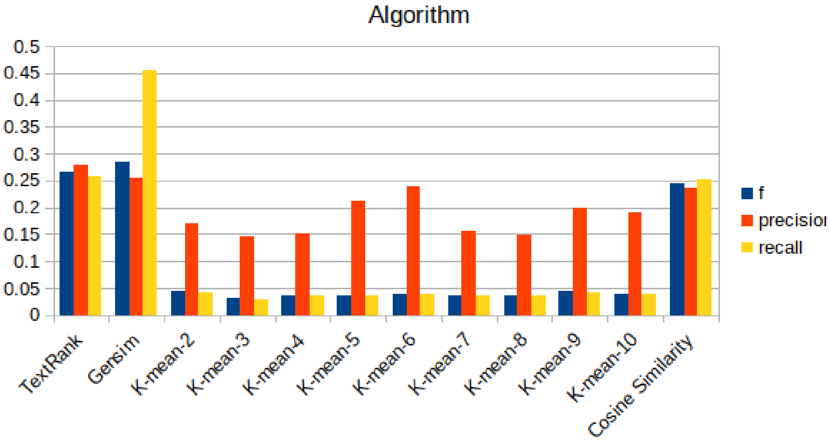
In order to evalute generated summaries with reference summary, we need a relavant evaluation tool. How can we state in term of number how close is our summary to the reference one.

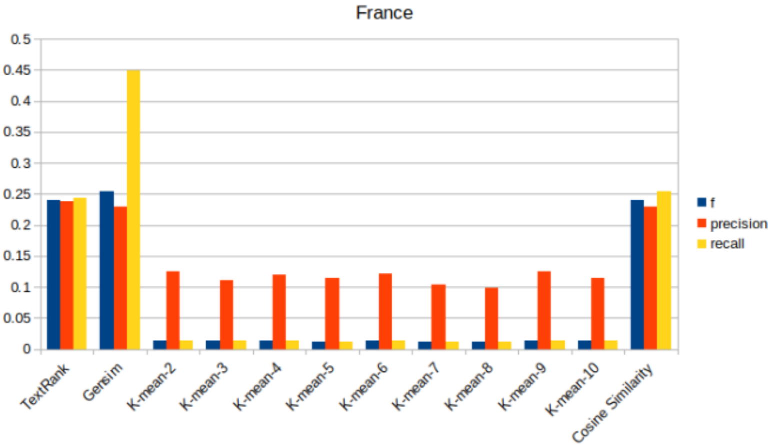
We decided to use ROUGE evaluation system [5]. ROUGE evaluation is a method to calculate the percentage of generate summary in reference summary and vice versa

$$\frac{\sum_r \sum_s \mathrm{match}(\mathrm{gram}_{s,c})}{\sum_r \sum_s \mathrm{count}(\mathrm{gram}_s)}$$



- ROUGE-n recall=40% means that 40% of the n-grams in the *reference* summary are also present in the *generated* summary.
- ROUGE-n precision=40% means that 40% of the n-grams in the *generated* summary are also present in the *reference* summary.
- ROUGE-n F1-score=40% is more difficult to interpret, like any F1-score.





Overall precision for all algorithms varies from 15% to 30%. It means that 15-30% of the n-grams in the generated summary are also present in the reference summary.

RECALL $$\frac{number\_of\_overlapping\_words}{total\_words\_in\_reference\_summary}$$

Overall recall gives us quite same values (about 30%) except Gensim which is way higher. It means that 30% of the n-grams in the reference summary are also present in the generated summary. However, K-mean implementation leads to a very low recall no matter how many clusters we choose.

PRECISION $$\frac{number\_of\_overlapping\_words}{total\_words\_in\_system\_summary}$$

Considering f score, it's way tougher to evaluate but its pretty well in agreement with precision and recall (except for K-Mean).

So why K-mean results are not relevant ?

We tried to figure out and tweak the algorithm but we guess that the approach we took was not optimal. It would have been way more relevant to combine K-Mean classification with other algorithm. We could have apply Text Rank Algorithm inside of each cluster.

Conclusion

Even K-mean can be in some way an approach for summarize text, our implementation didnt provide expected results. Although this basic implementation didnt do the trick, be believe that combined with other evaluation algorithms, it can be used in text summarization domain.

References

[1] Google Page Rank algorithm. https://en.wikipedia.org/wiki/PageRank

[2] K-means clustering, Wikipedia. https://en.wikipedia.org/wiki/K-means_clustering

[3] Automatic document summarization by sentence extraction. https://pdfs.semanticscholar.org/e7a4/8350000cec2025a212e7e3ca533b76351027.pdf

[4] Automatic extractive single document summarization, An unsupervised approach https://pdfs.semanticscholar.org/e7a4/8350000cec2025a212e7e3ca533b76351027.pdf

[5] ROUGE (Metric), Wikipedia. https://en.wikipedia.org/wiki/ROUGE_(metric)

[6] n-gram https://en.wikipedia.org/wiki/N-gram

[7] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." http://anthology.aclweb.org/W/W04/W04-1013.pdf