COMP 5313- ARTIFICIAL INTELLIGENCE

PROJECT 2

Text Summarization Techniques in Python

Methodology:

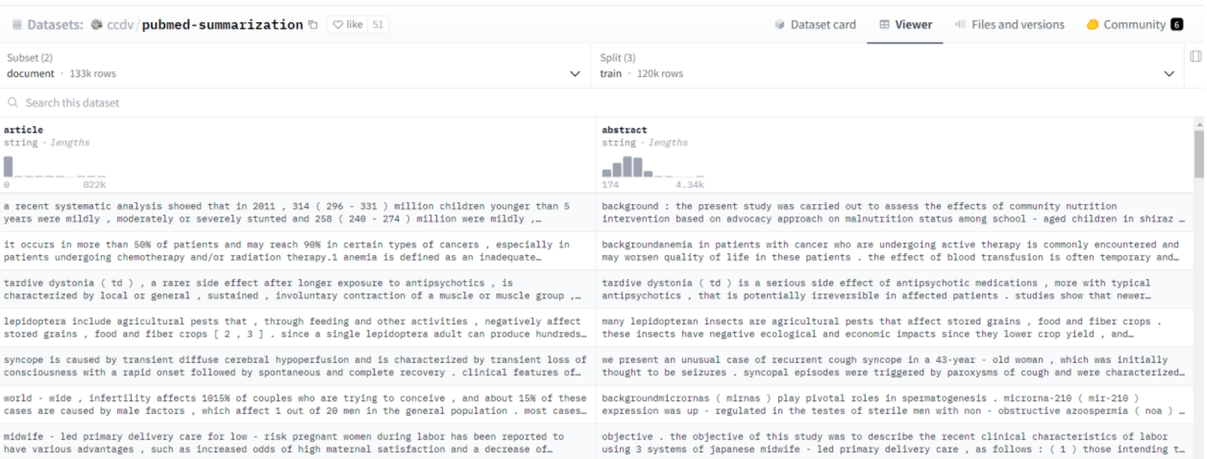
I created one .py file in order to do text summarization techniques in python.

I used four techniques that are BART, T5, BERT and Pegasus to do text summarization.

Dataset:

Here I used ccdv/pubmed-summarization dataset from hugging face. This dataset has two columns article and abstract. Abstract is the summarization of the article which will be the baseline of our project.

This is the dataset.


Datasets: ccdv/pubmed-summarization like 51 Dataset card Viewer Files and versions Community 6

Subset (2) document · 133k rows

Split (3) train · 120k rows

Search this dataset

article string · lengths 0 822k	abstract string · lengths 174 4.34k
a recent systematic analysis showed that in 2011 , 314 (296 - 331) million children younger than 5 years were mildly , moderately or severely stunted and 258 (240 - 274) million were mildly ,…	background : the present study was carried out to assess the effects of community nutrition intervention based on advocacy approach on malnutrition status among school - aged children in shiraz …
it occurs in more than 50% of patients and may reach 90% in certain types of cancers , especially in patients undergoing chemotherapy and/or radiation therapy.1 anemia is defined as an inadequate…	backgroundanemia in patients with cancer who are undergoing active therapy is commonly encountered and may worsen quality of life in these patients . the effect of blood transfusion is often temporary and…
tardive dystonia (td) , a rarer side effect after longer exposure to antipsychotics , is characterized by local or general , sustained , involuntary contraction of a muscle or muscle group ,…	tardive dystonia (td) is a serious side effect of antipsychotic medications , more with typical antipsychotics , that is potentially irreversible in affected patients . studies show that newer…
lepidoptera include agricultural pests that , through feeding and other activities , negatively affect stored grains , food and fiber crops [2 , 3] . since a single lepidoptera adult can produce hundreds…	many lepidopteran insects are agricultural pests that affect stored grains , food and fiber crops . these insects have negative ecological and economic impacts since they lower crop yield , and…
syncope is caused by transient diffuse cerebral hypoperfusion and is characterized by transient loss of consciousness with a rapid onset followed by spontaneous and complete recovery . clinical features of…	we present an unusual case of recurrent cough syncope in a 43-year - old woman , which was initially thought to be seizures . syncopal episodes were triggered by paroxysms of cough and were characterized…
world - wide , infertility affects 1015% of couples who are trying to conceive , and about 15% of these cases are caused by male factors , which affect 1 out of 20 men in the general population . most cases…	backgroundmicrornas (mirnas) play pivotal roles in spermatogenesis . microrna-210 (mir-210) expression was up - regulated in the testes of sterile men with non - obstructive azoospermia (noa) …
midwife - led primary delivery care for low - risk pregnant women during labor has been reported to have various advantages , such as increased odds of high maternal satisfaction and a decrease of…	objective . the objective of this study was to describe the recent clinical characteristics of labor using 3 systems of japanese midwife - led primary delivery care , as follows : (1) those intending t…

```python
# Load the dataset
dataset = load_dataset("ccdv/pubmed-summarization", ignore_verifications=True)
```

Abstractive text summarization

Abstractive text summarization generates legible sentences from the entirety of the text provided. It rewrites large amounts of text by creating acceptable representations, which is further processed and summarized by natural language processing.

Then I created a sample text with first 1000 characters of the article

```
sample_text = dataset["train"][1]["article"][:1000]
sample_text
```

```
'it occurs in more than 50% of patients and may reach 90% in certain types of cancers , especially in patients
undergoing chemotherapy and/or radiation therapy.1 anemia is defined as an inadequate circulating level of hemo
globin ( hb ) ( hb < 12 g / dl ) and may arise as a result of the underlying disease , bleeding , poor nutritio
n , chemotherapy , or radiation therapy . \n preliminary studies suggest that survival and loco - regional cont
rol after radiation therapy , especially in head and neck cancers , may be compromised by anemia.24 anemia ofte
n worsens symptoms such as fatigue , weakness , and dyspnea , and thus may have a negative effect on quality of
life ( qol ) and performance status in patients with cancer . \n thus , to improve physical functioning , qol ,
and prognosis in patients with cancer , it would be reasonable to take a proactive approach in identifying popu
lations who need treatment for cancer - associated anemia ( caa ) and provide timely management . \n blood tra
n...'
```

T5:

T5, or **Text-to-Text Transfer Transformer**, is a Transformer based architecture that uses a text-to-text approach. Every task – including translation, question answering, and classification – is cast as feeding the model text as input and training it to generate some target text.

```
# Initializing T5 pipeline
t5_pipeline = pipeline('summarization', model='t5-small')
t5_output = t5_pipeline(sample_text)
summaries['t5'] = '\n'.join(sent_tokenize(t5_output[0]['summary_text']))
```

I created a pipeline for the T5 model and passed sample_text through the model to generate summarization.

BART:

BART, or **Bidirectional and Auto-Regressive Transformers** is a denoising autoencoder for pretraining sequence-to-sequence models. It is trained by corrupting text with an arbitrary noising function, and learning a model to reconstruct the original text. It uses a standard Transformer-based neural machine translation architecture.

```
# Initialize BART pipeline
bart_pipeline = pipeline("summarization", model="facebook/bart-large-cnn")
bart_output = bart_pipeline(sample_text)
summaries['bart'] = '\n'.join(sent_tokenize(bart_output[0]['summary_text']))
```

Here, I created a pipeline for the BART and passed sample_text through the model to generate summarization.

PEGASUS:

PEGASUS proposes a transformer-based model for abstractive summarization. It uses a special self-supervised pre-training objective called gap-sentences generation (GSG) that's designed to perform well on summarization-related downstream tasks.

```
# Initialize PEGASUS pipeline
pegasus_tokenizer = AutoTokenizer.from_pretrained("google/pegasus-large")
pegasus_model = AutoModelForSeq2SeqLM.from_pretrained("google/pegasus-large")
pegasus_pipeline = pipeline("summarization", model=pegasus_model, tokenizer=pegasus_tokenizer)
pegasus_output = pegasus_pipeline(sample_text)
summaries['pegasus'] = '\n'.join(sent_tokenize(pegasus_output[0]['summary_text']))
```

First, I initialized the PEGASUS tokenizer using 'AutoTokenizer' from the Hugging Face Transformers library. Then, I created a pipeline for the PEGASUS and passed sample_text through the model to generate summarization.

BERT:

BERT, or **Bidirectional Encoder Representations from Transformers**, improves upon standard Transformers by removing the unidirectionality constraint by using a masked language model (MLM) pre-training objective.

```
# Initialize BERT model
bert_tokenizer = BertTokenizer.from_pretrained("bert-base-uncased")
bert_model = BertForNextSentencePrediction.from_pretrained("bert-base-uncased")
bert_pipeline = pipeline("feature-extraction", model=bert_model, tokenizer=bert_tokenizer)
bert_output = bert_pipeline(sample_text)
top_sentences = sorted(list(enumerate(bert_output[0])), key=lambda x: x[1], reverse=True)[:3]
summary_sentences = [sent_tokenize(sample_text)[index] for index, _ in top_sentences]
summaries['bert'] = '\n'.join(summary_sentences)
```

Here I initialized the BERT tokenizer with the pre-trained weights of the "bert-base-uncased" model. Initialized the BERT model for next sentence prediction task using the pre-trained weights. Then, I created a pipeline for the BERT and passed sample_text through the model to generate summarization.

This is what the summarized text looks like.

```
summaries
```

```
{'t5': 'anemia is defined as an inadequate circulating level of hemoglobin ( hb  12 g / dl ) and may arise as a result of the underlying disease .\npreliminary studies suggest survival and loco - regional control after radiation therapy may be compromised by anemia .',
 'bart': 'Anemia is defined as an inadequate circulating level of hemoglobin ( hb) It occurs in more than 50% of patients and may reach 90% in certain types of cancers.\nAnemia often worsens symptoms such as fatigue and dyspnea.\nIt can have a negative effect on quality of life ( qol) and performance status in patients with cancer.',
 'pegasus': 'preliminary studies suggest that survival and loco - regional control after radiation therapy , especially in head and neck cancers , may be compromised by anemia.24 anemia often worsens symptoms such as fatigue , weakness , and dyspnea , and thus may have a negative effect on quality of life ( qol ) and performance status in patients with cancer .',
 'bert': 'it occurs in more than 50% of patients and may reach 90% in certain types of cancers , especially in patients undergoing chemotherapy and/or radiation therapy.1 anemia is defined as an inadequate circulating level of hemoglobin ( hb ) ( hb < 12 g / dl ) and may arise as a result of the underlying disease , bleeding , poor nutrition , chemotherapy , or radiation therapy .\npreliminary studies suggest that survival and loco - regional control after radiation therapy , especially in head and neck cancers , may be compromised by anemia.24 anemia often worsens symptoms such as fatigue , weakness , and dyspnea , and thus may have a negative effect on quality of life ( qol ) and performance status in patients with cancer .'}
```

```
T5 Summary:
anemia is defined as an inadequate circulating level of hemoglobin ( hb  12 g / dl ) and may arise as a result of the underlying disease .
preliminary studies suggest survival and loco - regional control after radiation therapy may be compromised by anemia .

Bart Summary:
Anemia is defined as an inadequate circulating level of hemoglobin ( hb) It occurs in more than 50% of patients and may reach 90% in certain types of cancers.
Anemia often worsens symptoms such as fatigue and dyspnea.
It can have a negative effect on quality of life ( qol) and performance status in patients with cancer.

Pegasus Summary:
preliminary studies suggest that survival and loco - regional control after radiation therapy , especially in head and neck cancers , may be compromised by anemia.24 anemia often worse

Bert Summary:
it occurs in more than 50% of patients and may reach 90% in certain types of cancers , especially in patients undergoing chemotherapy and/or radiation therapy.1 anemia is defined as an
preliminary studies suggest that survival and loco - regional control after radiation therapy , especially in head and neck cancers , may be compromised by anemia.24 anemia often worse
```

Rouge Score:

ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, is a set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing. The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation. ROUGE metrics range between 0 and 1, with higher scores indicating higher similarity between the automatically produced summary and the reference.

```
T5 ROUGE Scores: {'rouge1': 0.11956521739130434, 'rouge2': 0.027322404371584695, 'rougeL': 0.05978260869565217, 'rougeLsum': 0.10869565217391303}
Bart ROUGE Scores: {'rouge1': 0.1671018276762402, 'rouge2': 0.07349081364829396, 'rougeL': 0.10443864229765014, 'rougeLsum': 0.1566579634464752}
Pegasus ROUGE Scores: {'rouge1': 0.15748031496062992, 'rouge2': 0.058047493403693924, 'rougeL': 0.09448818897637797, 'rougeLsum': 0.12073490813648294}
Bert ROUGE Scores: {'rouge1': 0.2681818181818182, 'rouge2': 0.09132420091324202, 'rougeL': 0.1318181818181818, 'rougeLsum': 0.2227272727272727}
          rouge1    rouge2    rougeL  rougeLsum
t5       0.119565  0.027322  0.059783   0.108696
bart     0.167102  0.073491  0.104439   0.156658
pegasus  0.157480  0.058047  0.094488   0.120735
bert     0.268182  0.091324  0.131818   0.222727
```

These are the rouge score for all the above models.

Conclusion:

As you can see above, after using all the four deep learning models for text summarization. B ERT model got the higher rouge score in all rouge1 has 0.268, rouge2 has 0.091, rougeL has 0 .132 and rougeLsum has 0.223 which is the highest among all the other. Next comes BART m odel with rouge1 has 0.167, rouge2 has 0.073, rougeL has 0.104 and rougeLsum has 0.156. T hen comes PEGASUS and then T5. According to my model BERT is the best.

References:

[1] https://paperswithcode.com/method/t5
[2] https://paperswithcode.com/method/bart
[3] https://paperswithcode.com/method/pegasus
[4] https://paperswithcode.com/method/bert
[5] https://en.wikipedia.org/wiki/ROUGE_(metric)
[6] https://www.turing.com/kb/5-powerful-text-summarization-techniques-in-python
[7] https://www.projectpro.io/article/text-summarization-python-nlp/546