뉴스기사를 통한 뉴스 요약 생성

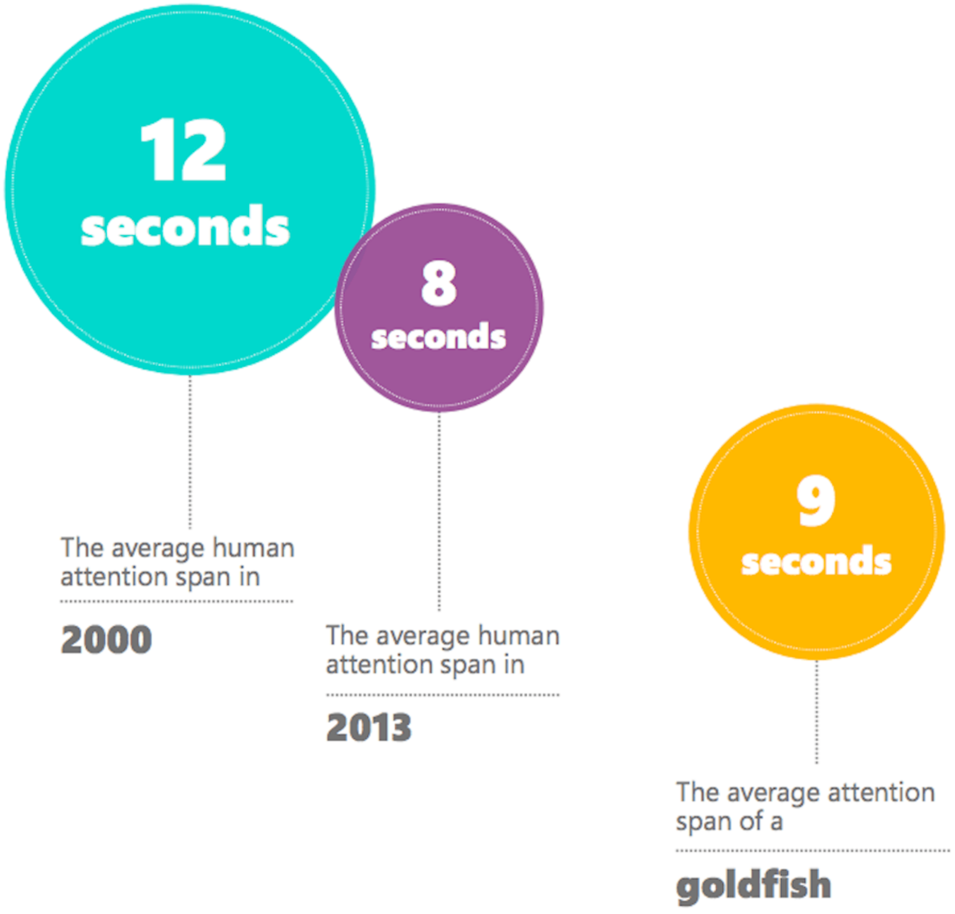
빅데이터경영통계전공

20162522 송민규

INDEX

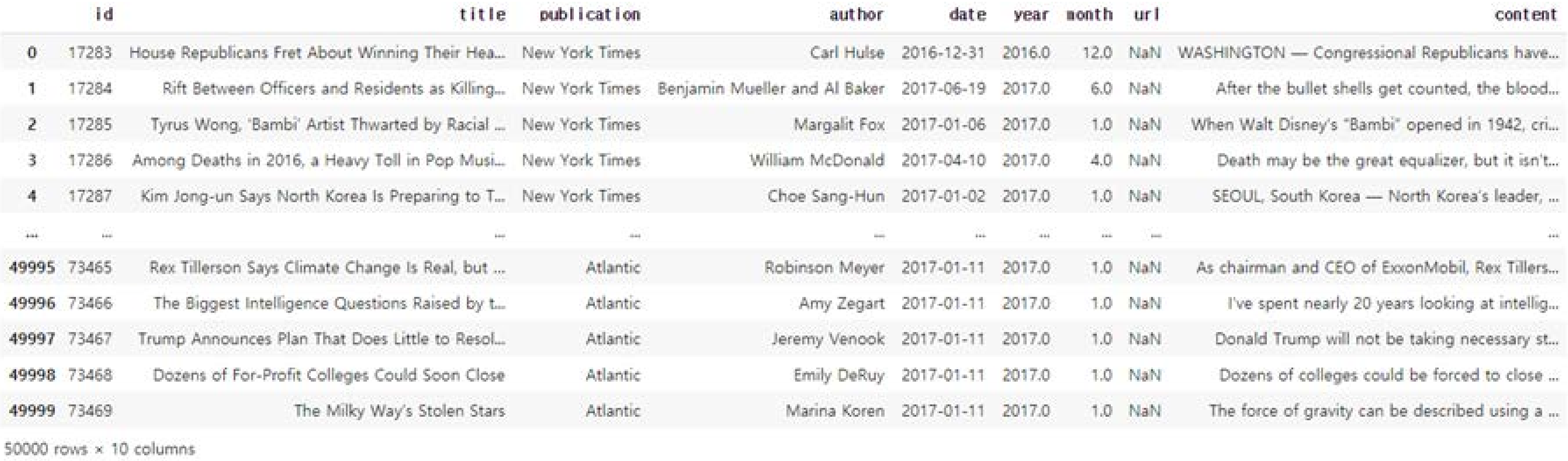
Introduction

최근 사회에서 사람들의 집중력은 날이 갈수록 짧아지고 있다. 옆의 그림을 보면, 2013년 기준으로 사람들의 작업에 집중하는 시간은 8초로, 금붕어의 9초 보다 짧다. 이에 따라 긴 글을 읽는 데에도 어려움을 겪고 있다. 그렇기에 인터넷에서도 내용을 3줄 요약해서 글을 써주는 경우도 많다. 뉴스 기사 또한 핵심만 요약해서 한눈에 보여준다면 뉴스기사를 선택하는 데 도움을 줄 수 있을 것이다.



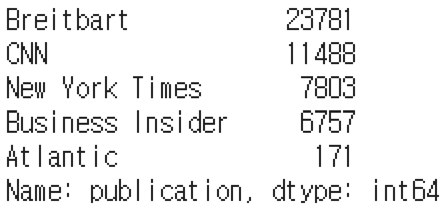
Data

- 데이터 이름 : All the news 중 article1
- 데이터 출처 : kaggle (https://www.kaggle.com/snapcrack/all-the-news)
- 데이터 구성 : 기사고유id, 제목, 발행사, 기자, 날짜, 기사내용
- 데이터 설명 : 15개 발행사에서 발행한 143000개의 기사에 관한 데이터이다. 데이터가 3개 있는데 그 중 Breitbart, CNN, New York Times, Business Insider, Atlantic 의 5개 발행사에서 2011~2017년까지 발행한 기사 50000개에 관한 데이터를 가져왔다.

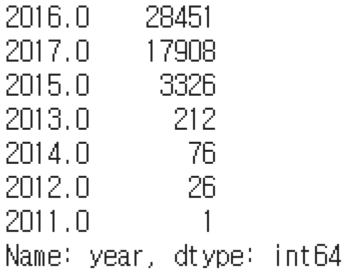


	id	title	publication	author	date	year	month	url	content
0	17283	House Republicans Fret About Winning Their Hea...	New York Times	Carl Hulse	2016-12-31	2016.0	12.0	NaN	WASHINGTON — Congressional Republicans have...
1	17284	Rift Between Officers and Residents as Killing...	New York Times	Benjamin Mueller and Al Baker	2017-06-19	2017.0	6.0	NaN	After the bullet shells get counted, the blood...
2	17285	Tyrus Wong, 'Bambi' Artist Thwarted by Racial ...	New York Times	Margalit Fox	2017-01-06	2017.0	1.0	NaN	When Walt Disney's "Bambi" opened in 1942, cri...
3	17286	Among Deaths in 2016, a Heavy Toll in Pop Musi...	New York Times	William McDonald	2017-04-10	2017.0	4.0	NaN	Death may be the great equalizer, but it isn't...
4	17287	Kim Jong-un Says North Korea Is Preparing to T...	New York Times	Choe Sang-Hun	2017-01-02	2017.0	1.0	NaN	SEOUL, South Korea — North Korea's leader, ...
...	...	...	...	...	...	...	...	...	...
49995	73465	Rex Tillerson Says Climate Change Is Real, but ...	Atlantic	Robinson Meyer	2017-01-11	2017.0	1.0	NaN	As chairman and CEO of ExxonMobil, Rex Tillers...
49996	73466	The Biggest Intelligence Questions Raised by t...	Atlantic	Amy Zegart	2017-01-11	2017.0	1.0	NaN	I've spent nearly 20 years looking at intellig...
49997	73467	Trump Announces Plan That Does Little to Resol...	Atlantic	Jeremy Venook	2017-01-11	2017.0	1.0	NaN	Donald Trump will not be taking necessary st...
49998	73468	Dozens of For-Profit Colleges Could Soon Close	Atlantic	Emily DeRuy	2017-01-11	2017.0	1.0	NaN	Dozens of colleges could be forced to close ...
49999	73469	The Milky Way's Stolen Stars	Atlantic	Marina Koren	2017-01-11	2017.0	1.0	NaN	The force of gravity can be described using a ...

50000 rows × 10 columns



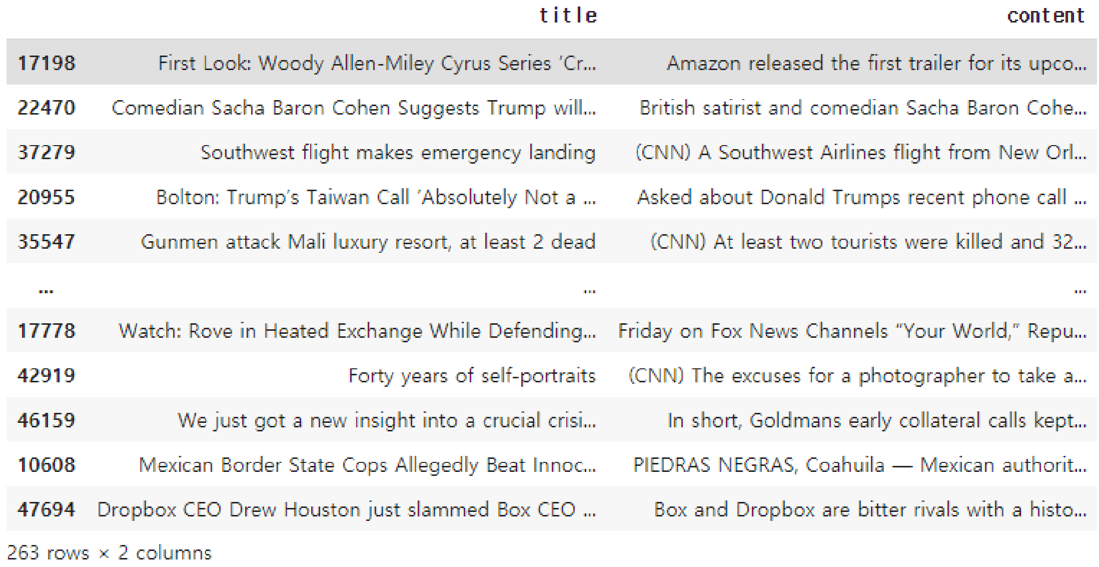
```
Breitbart           23781
CNN                 11488
New York Times       7803
Business Insider     6757
Atlantic              171
Name: publication, dtype: int64
```



```
2016.0    28451
2017.0    17908
2015.0     3326
2013.0      212
2014.0       76
2012.0       26
2011.0        1
Name: year, dtype: int64
```

Preprocessing

- 요약하기 전 기사 본문 부분(content) 부분의 길이를 확인하여 길이가 너무 긴 문장들은 제외 시켰다. 길이는 기준을 3000으로 놓고 그 이하의 것들만 선정하였더니 26201개의 데이터가 추출되었다. 개수가 너무 많으므로 임의로 1%만 추출하여 263개의 데이터에 대해 요약을 진행할 것이다.
- 이 중 제목과 기사내용만 필요하여 그 두가지 column만 추출하였고, 기사 본문 부분(content)에서 쓸모 없는 공백, 쓸모 없는 기호 (‘) 를 제거해 주었다.



	title	content
17198	First Look: Woody Allen-Miley Cyrus Series ’Cr...	Amazon released the first trailer for its upco...
22470	Comedian Sacha Baron Cohen Suggests Trump will...	British satirist and comedian Sacha Baron Cohe...
37279	Southwest flight makes emergency landing	(CNN) A Southwest Airlines flight from New Orl...
20955	Bolton: Trump’s Taiwan Call ’Absolutely Not a ...	Asked about Donald Trumps recent phone call ...
35547	Gunmen attack Mali luxury resort, at least 2 dead	(CNN) At least two tourists were killed and 32...
...	...	...
17778	Watch: Rove in Heated Exchange While Defending...	Friday on Fox News Channels “Your World,” Repu...
42919	Forty years of self-portraits	(CNN) The excuses for a photographer to take a...
46159	We just got a new insight into a crucial crisi...	In short, Goldmans early collateral calls kept...
10608	Mexican Border State Cops Allegedly Beat Innoc...	PIEDRAS NEGRAS, Coahuila — Mexican authorit...
47694	Dropbox CEO Drew Houston just slammed Box CEO ...	Box and Dropbox are bitter rivals with a histo...

263 rows × 2 columns

Model

- 두가지 모델을 사용
- 1. transformer의 pipeline.

-> Pipeline은 추론을 위해 모델을 사용하는 방법이다. 작업할 수 있는 API를 제공하는 데 그 중 summarization 기능을 사용하였다. Summarization은 미세 조정 된 Bart 모델을 사용한다. 여기서는 content 내용을 요약한 후 Summary라는 column에 삽입해 주었다.

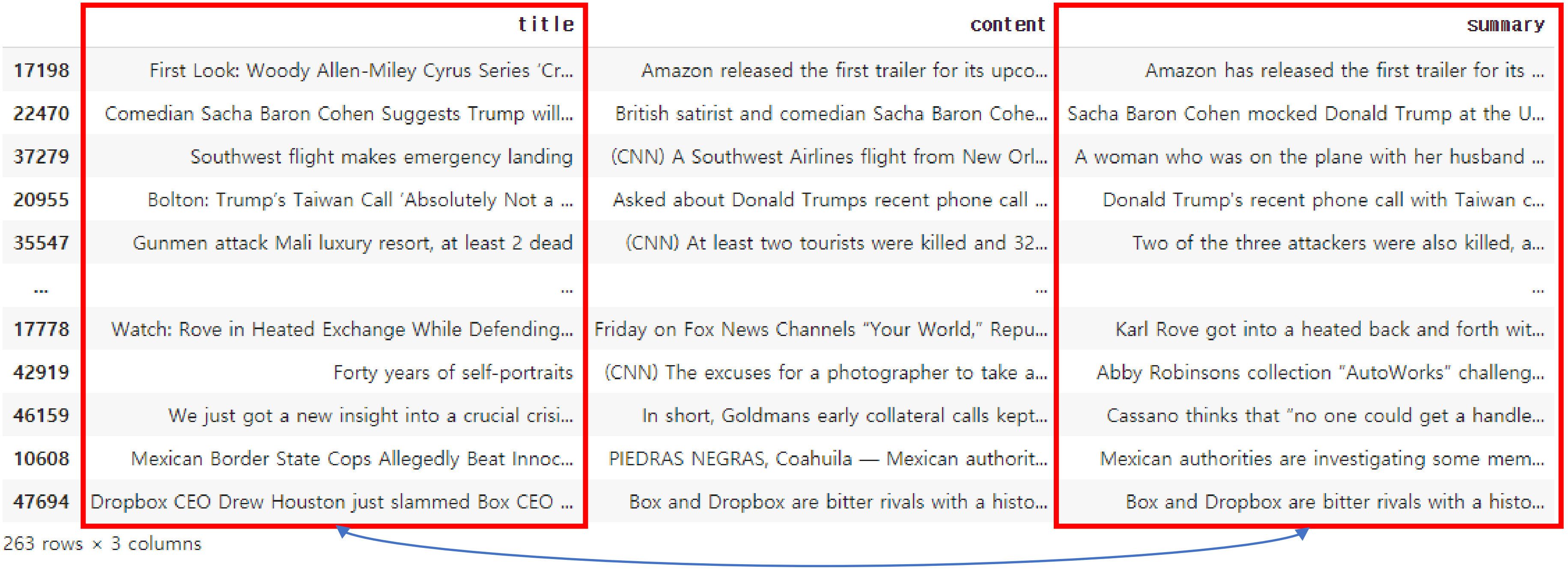
- 2. Google에서 만든 T5.

-> T5는 Text-To-Text Transfer Transformer의 약자로, 통합된 텍스트 입력 – 텍스트 출력으로 NLP를 처리하는 방법이다. 여기서는 위에 방법과 동일하게 content 내용을 요약한 후 Summary_T5라는 column에 삽입해 주었다. 또한 model과 tokenizer을 t5-base로 사용해 주었다.

Training & Fine Tuning

- 1. Pipeline summarization 학습 및 미세조정.

-> summarization 모델을 그대로 사용한다. 요약한 결과를 기사 제목과 비교할 것이기 때문에 parameter을 추가하여 제목과 비슷한 길이로 요약을 생성한다.



	title	content	summary
17198	First Look: Woody Allen-Miley Cyrus Series 'Cr...	Amazon released the first trailer for its upco...	Amazon has released the first trailer for its ...
22470	Comedian Sacha Baron Cohen Suggests Trump will...	British satirist and comedian Sacha Baron Cohe...	Sacha Baron Cohen mocked Donald Trump at the U...
37279	Southwest flight makes emergency landing	(CNN) A Southwest Airlines flight from New Orl...	A woman who was on the plane with her husband ...
20955	Bolton: Trump's Taiwan Call 'Absolutely Not a ...	Asked about Donald Trumps recent phone call ...	Donald Trump's recent phone call with Taiwan c...
35547	Gunmen attack Mali luxury resort, at least 2 dead	(CNN) At least two tourists were killed and 32...	Two of the three attackers were also killed, a...
...	...	...	...
17778	Watch: Rove in Heated Exchange While Defending...	Friday on Fox News Channels "Your World," Repu...	Karl Rove got into a heated back and forth wit...
42919	Forty years of self-portraits	(CNN) The excuses for a photographer to take a...	Abby Robinsons collection "AutoWorks" challeng...
46159	We just got a new insight into a crucial crisi...	In short, Goldmans early collateral calls kept...	Cassano thinks that "no one could get a handle...
10608	Mexican Border State Cops Allegedly Beat Innoc...	PIEDRAS NEGRAS, Coahuila — Mexican authorit...	Mexican authorities are investigating some mem...
47694	Dropbox CEO Drew Houston just slammed Box CEO ...	Box and Dropbox are bitter rivals with a histo...	Box and Dropbox are bitter rivals with a histo...

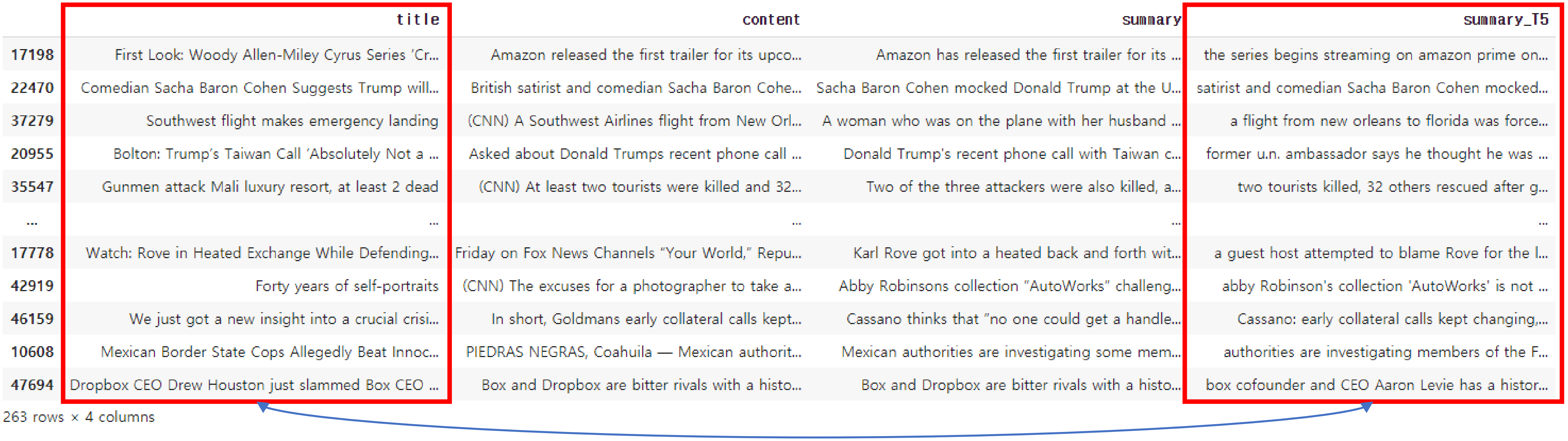
263 rows × 3 columns

두개 비교예정

Training & Fine Tuning

- 2. T5 transformer 학습 및 미세조정.

-> model과 tokenizer 둘 다 t5-base를 사용한다. length_penalty를 설정하여 길이에 따라 가중치를 주는 방식을 사용하였고, num_beams를 설정하여 beam탐색 시에 탐색 수를 지정해주었다.



	title	content	summary	summary_T5
17198	First Look: Woody Allen-Miley Cyrus Series 'Cr...	Amazon released the first trailer for its upco...	Amazon has released the first trailer for its ...	the series begins streaming on amazon prime on...
22470	Comedian Sacha Baron Cohen Suggests Trump will...	British satirist and comedian Sacha Baron Cohe...	Sacha Baron Cohen mocked Donald Trump at the U...	satirist and comedian Sacha Baron Cohen mocked...
37279	Southwest flight makes emergency landing	(CNN) A Southwest Airlines flight from New Orl...	A woman who was on the plane with her husband ...	a flight from new orleans to florida was force...
20955	Bolton: Trump's Taiwan Call 'Absolutely Not a ...	Asked about Donald Trumps recent phone call ...	Donald Trump's recent phone call with Taiwan c...	former u.n. ambassador says he thought he was ...
35547	Gunmen attack Mali luxury resort, at least 2 dead	(CNN) At least two tourists were killed and 32...	Two of the three attackers were also killed, a...	two tourists killed, 32 others rescued after g...
...	...	...	...	...
17778	Watch: Rove in Heated Exchange While Defending...	Friday on Fox News Channels "Your World," Repu...	Karl Rove got into a heated back and forth wit...	a guest host attempted to blame Rove for the l...
42919	Forty years of self-portraits	(CNN) The excuses for a photographer to take a...	Abby Robinsons collection "AutoWorks" challeng...	abby Robinson's collection 'AutoWorks' is not ...
46159	We just got a new insight into a crucial crisi...	In short, Goldmans early collateral calls kept...	Cassano thinks that "no one could get a handle...	Cassano: early collateral calls kept changing,...
10608	Mexican Border State Cops Allegedly Beat Innoc...	PIEDRAS NEGRAS, Coahuila — Mexican authorit...	Mexican authorities are investigating some mem...	authorities are investigating members of the F...
47694	Dropbox CEO Drew Houston just slammed Box CEO ...	Box and Dropbox are bitter rivals with a histo...	Box and Dropbox are bitter rivals with a histo...	box cofounder and CEO Aaron Levie has a histor...

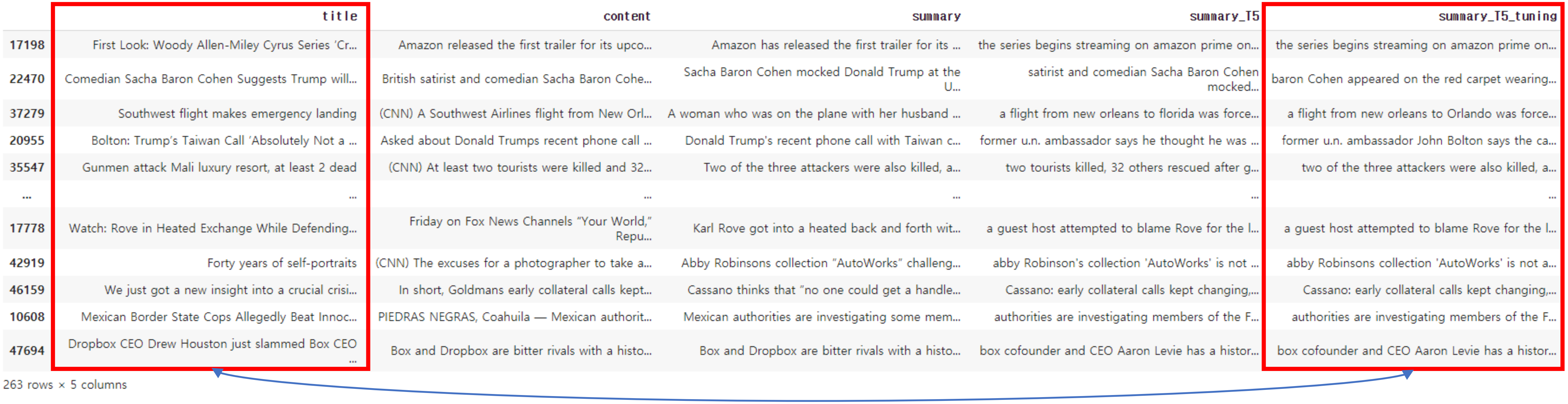
263 rows × 4 columns

두개 비교예정

Training & Fine Tuning

- 3. T5 transformer parameter 변경.

-> length_penalty는 숫자가 커질 수록 나오는 요약의 길이가 길어지므로 수치를 낮춰 보았고, num_beams는 확률이 높은 단어를 개수를 늘려 탐색하면 더 좋은 결과가 나올거라 생각하고 수치를 올려보았다. 또한 repetition_penalty를 주어 같은 단어가 반복시에 가중치를 주는 방식을 채택하였다.



	title	content	summary	summary_T5	summary_T5_tuning
17198	First Look: Woody Allen-Miley Cyrus Series 'Cr...	Amazon released the first trailer for its upco...	Amazon has released the first trailer for its ...	the series begins streaming on amazon prime on...	the series begins streaming on amazon prime on...
22470	Comedian Sacha Baron Cohen Suggests Trump will...	British satirist and comedian Sacha Baron Cohe...	Sacha Baron Cohen mocked Donald Trump at the U...	satirist and comedian Sacha Baron Cohen mocked...	baron Cohen appeared on the red carpet wearing...
37279	Southwest flight makes emergency landing	(CNN) A Southwest Airlines flight from New Orl...	A woman who was on the plane with her husband ...	a flight from new orleans to florida was force...	a flight from new orleans to Orlando was force...
20955	Bolton: Trump's Taiwan Call 'Absolutely Not a ...	Asked about Donald Trumps recent phone call ...	Donald Trump's recent phone call with Taiwan c...	former u.n. ambassador says he thought he was ...	former u.n. ambassador John Bolton says the ca...
35547	Gunmen attack Mali luxury resort, at least 2 dead	(CNN) At least two tourists were killed and 32...	Two of the three attackers were also killed, a...	two tourists killed, 32 others rescued after g...	two of the three attackers were also killed, a...
...	...	...	...	...	...
17778	Watch: Rove in Heated Exchange While Defending...	Friday on Fox News Channels "Your World," Repu...	Karl Rove got into a heated back and forth wit...	a guest host attempted to blame Rove for the l...	a guest host attempted to blame Rove for the l...
42919	Forty years of self-portraits	(CNN) The excuses for a photographer to take a...	Abby Robinsons collection "AutoWorks" challeng...	abby Robinson's collection 'AutoWorks' is not ...	abby Robinsons collection 'AutoWorks' is not a...
46159	We just got a new insight into a crucial crisi...	In short, Goldmans early collateral calls kept...	Cassano thinks that "no one could get a handle...	Cassano: early collateral calls kept changing,...	Cassano: early collateral calls kept changing,...
10608	Mexican Border State Cops Allegedly Beat Innoc...	PIEDRAS NEGRAS, Coahuila — Mexican authorit...	Mexican authorities are investigating some mem...	authorities are investigating members of the F...	authorities are investigating members of the F...
47694	Dropbox CEO Drew Houston just slammed Box CEO ...	Box and Dropbox are bitter rivals with a histo...	Box and Dropbox are bitter rivals with a histo...	box cofounder and CEO Aaron Levie has a histor...	box cofounder and CEO Aaron Levie has a histor...

263 rows × 5 columns

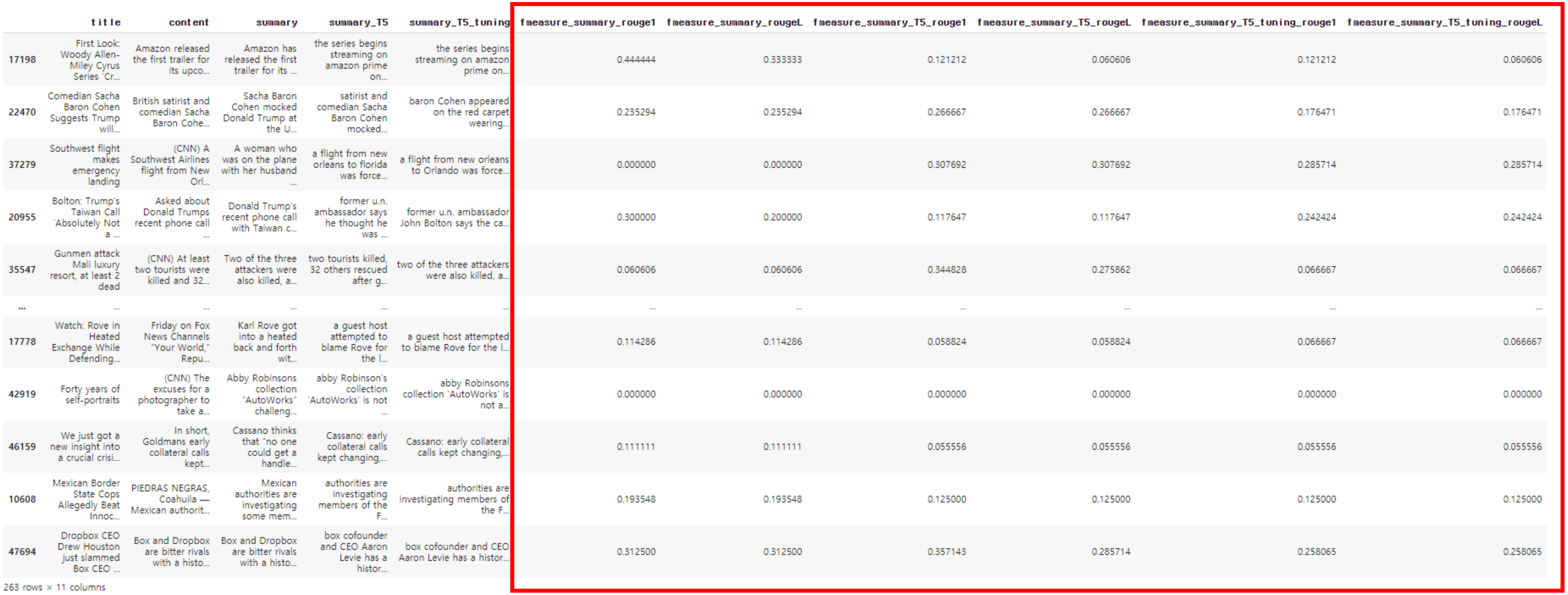
두개 비교예정

Evaluation

- Rouge-score이란, Recall-Oriented Understudy for Gisting Evaluation의 약자로 텍스트 요약 모델의 성능 평가 지표이다. 이를 사용하여 평가를 할 것이다. 성능 지표로 사용하기 위해 Recall과 Precision을 구하는데, 비슷한 평가 방식인 BLEU는 precision을 기준으로 평가를 하고, Rouge-score은 recall 부분 까지 같이 고려를 하여 평가하는 방식이다.
- Recall은 참조 요약본의 단어 중 몇 개가 모델 생성 요약본과 겹치는 지에 대한 점수고, Precision은 반대로 모델 생성 요약본의 단어가 참조 요약본과 얼마나 겹치는 지에 대한 점수이다. 여기서는 참조 요약본을 title로, 모델 생성 요약본을 summary, summary_T5, summary_T5_tuning으로 사용하였다.
- Recall과 Precision을 모두 고려하는 것이 fmeasure(f1 score)로 두 개의 조화평균으로 구할 수 있다. 여기서 성능 판단을 할 때 fmeasure 값으로 성능 판단을 하였다.
- Rouge1 과 RougeL에 대하여 score을 도출하였는데, Rouge1은 1-gram, 즉 단어 단위로 두 요약문에 대해서 비교하는 것이다. RougeL은 최장 공통 부분 수열을 가지고 평가하는 방식인데, 즉 연속된 단어열 중 가장 길이가 긴 것으로 비교하는 방식이다.

Evaluation

- 전부 fmeasure(f1 score)값으로 rouge1과 rougeL으로 나누어 평가하였다.

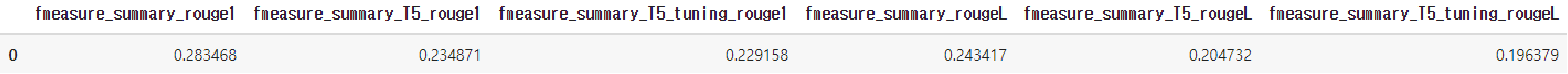


	title	content	summary	summary_T5	summary_T5_tuning	fmeasure_summary_rouge1	fmeasure_summary_rougeL	fmeasure_summary_T5_rouge1	fmeasure_summary_T5_rougeL	fmeasure_summary_T5_tuning_rouge1	fmeasure_summary_T5_tuning_rougeL
17198	First Look: Woody Allen-Miley Cyrus Series 'Cr...	Amazon released the first trailer for its upco...	Amazon has released the first trailer for its ...	the series begins streaming on amazon prime on...	the series begins streaming on amazon prime on...	0.444444	0.333333	0.121212	0.060606	0.121212	0.060606
22470	Comedian Sacha Baron Cohen Suggests Trump will...	British satirist and comedian Sacha Baron Cohe...	Sacha Baron Cohen mocked Donald Trump at the U...	satirist and comedian Sacha Baron Cohen mocked...	baron Cohen appeared on the red carpet wearing...	0.235294	0.235294	0.266667	0.266667	0.176471	0.176471
37279	Southwest flight makes emergency landing	(CNN) A Southwest Airlines flight from New Orl...	A woman who was on the plane with her husband ...	a flight from new orleans to florida was force...	a flight from new orleans to Orlando was force...	0.000000	0.000000	0.307692	0.307692	0.285714	0.285714
20955	Bolton: Trump's Taiwan Call 'Absolutely Not a ...	Asked about Donald Trumps recent phone call ...	Donald Trump's recent phone call with Taiwan c...	former u.n. ambassador says he thought he was ...	former u.n. ambassador John Bolton says the ca...	0.300000	0.200000	0.117647	0.117647	0.242424	0.242424
35547	Gunmen attack Mali luxury resort, at least 2 dead	(CNN) At least two tourists were killed and 32...	Two of the three attackers were also killed, a...	two tourists killed, 32 others rescued after g...	two of the three attackers were also killed, a...	0.060606	0.060606	0.344828	0.275862	0.066667	0.066667
...	...	...	...	...	...	...	...	...	...	...	...
17778	Watch: Rove in Heated Exchange While Defending...	Friday on Fox News Channels "Your World," Repu...	Karl Rove got into a heated back and forth wit...	a guest host attempted to blame Rove for the l...	a guest host attempted to blame Rove for the l...	0.114286	0.114286	0.058824	0.058824	0.066667	0.066667
42919	Forty years of self-portraits	(CNN) The excuses for a photographer to take a...	Abby Robinsons collection "AutoWorks" challeng...	abby Robinson's collection 'AutoWorks' is not ...	abby Robinsons collection 'AutoWorks' is not a...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
46159	We just got a new insight into a crucial crisi...	In short, Goldmans early collateral calls kept...	Cassano thinks that "no one could get a handle...	Cassano: early collateral calls kept changing,...	Cassano: early collateral calls kept changing,...	0.111111	0.111111	0.055556	0.055556	0.055556	0.055556
10608	Mexican Border State Cops Allegedly Beat Innoc...	PIEDRAS NEGRAS, Coahuila — Mexican authorit...	Mexican authorities are investigating some mem...	authorities are investigating members of the F...	authorities are investigating members of the F...	0.193548	0.193548	0.125000	0.125000	0.125000	0.125000
47694	Dropbox CEO Drew Houston just slammed Box CEO ...	Box and Dropbox are bitter rivals with a histo...	Box and Dropbox are bitter rivals with a histo...	box cofounder and CEO Aaron Levie has a histor...	box cofounder and CEO Aaron Levie has a histor...	0.312500	0.312500	0.357143	0.285714	0.258065	0.258065

263 rows × 11 columns

Evaluation

- 어떤 방식이 가장 요약이 잘 되었는지 판단하기 위해 각각의 score에 대해서 column마다 평균값을 구하였다.



	fmeasure_summary_rouge1	fmeasure_summary_T5_rouge1	fmeasure_summary_T5_tuning_rouge1	fmeasure_summary_rougeL	fmeasure_summary_T5_rougeL	fmeasure_summary_T5_tuning_rougeL
0	0.283468	0.234871	0.229158	0.243417	0.204732	0.196379

- Rouge1과 RougeL 두 개 모두 pipeline을 사용한 요약 생성이 가장 성능이 높게 나왔다. Parameter tuning 한 것이 기본으로 T5를 사용한 것 보다 성능이 낮은 것도 봐야할 대목이다.
- 데이터와 미세 조정에 따라 성능 차이가 많이 나겠지만, 시도한 방법 중에서는 pipeline summarization API를 이용한 방법이 가장 좋은 것으로 나타났다.

References

- About Preprocessing & model :

  https://www.thepythoncode.com/article/text-summarization-using-huggingface-transformers-python

- About Training & Fine Tuning :

  https://huggingface.co/transformers/task_summary.html

- About Evaluation :

  https://pypi.org/project/rouge-score/

  https://huffon.github.io/2019/12/07/rouge/

  https://ai-information.blogspot.com/2019/04/text-generation-evaluation-06-rouge.html

자기 평가표

항목	점수	평가 근거
서론	2/2	하고자 하는 것, 평가근거, 기대효과 등을 설명하였음.
데이터	2/2	데이터 출처, 구성 등을 설명하였음.
전처리	2/3	데이터 자르기, 노이즈 등을 제거하였음.
모형	3/3	수업시간에 배운 Pipeline 뿐만 아니라 T5 transformer을 추가 사용하였음.
학습 및 미세조정	2/3	Parameter 조정 정도 추가하였고 수업에서 소개한 방법을 그대로 사용하였음.
평가	3/3	수업에 배운 BLEU 말고 Rouge score를 추가로 조사하여 사용하였음.
합계	14/16	