

KUBIG CONTEST NLP 분반

금융 뉴스 요약 & 메일링 프로젝트

15기 김제성, 김지후, 남정재, 이병주, 이승은





CONTENTS

01. 주제 설명

02. Dataset

03. 모델

04. 메일링 서비스 결과

05. Future Work

네이버 금융 뉴스 요약 모델 생성 & 메일링 서비스



AI Hub 뉴스 기사 요약 데이터

	passage	summary
0	이건희 삼성그룹 회장 사망 이후 불거진 상속 세제 개편 논의가 급물살을 탈지 주목된...	홍 경제부총리 겸 기획재정부 장관이 이건희 회장의 사망 이후 불거진 상속 세제 개편...
1	17일 국무총리실 산하 김해 신공항 검증위원회가 내놓은 '사실상 백지화'결론에 대해...	사실상 백지화 결론에 대해 경제성 평가를 빼고 검증했다는 전문가의 지적에 김 위원장...
2	12일 오전 11시 30분경 서울 소공동 롯데백화점 본점 샤넬 매장 앞에는 매장에 ...	다수의 명품 브랜드에 이어 샤넬이 7개월 만에 가격을 또 인상한다는 소식이 전해지자...
3	이탈리아 롬바르디아주의 한 병원에 근무하는 간호사 모니카 마리오티는 밤마다 악몽에 ...	세레나 박사에 따르면 이탈리아 코로나19 보건 종사자 약70%가 우울증을 앓고 있는...
4	세 곡 모두 2000년대 들어 발표된 곡으로 전통 트로트라기보다는 댄스가 가미된 ...	대중문화평론가 이 씨는 현미의 밤안개 등을 트로트로 분류하기는 어렵지만 수용층이 넓...
...
10795	병원 탐방 서울밝은세상안과 눈은 일상에서 가장 많이 혹사당하는 신체 기관이다.\n ...	이 대표원장은 다른 질환이 있으면 백내장 수술을 해도 원하는 효과를 기대하기 힘들다...
10796	정부가 4차 추가경정예산을 통해 만 13세 이상 국민에 통신비 2만원을 지원하겠다고...	만 13세 이상 국민에게 통신비 2만원을 정부가 지원하겠다고 밝히자 예정처는 일부 ...
10797	손 떨림 걱정이 없어 오차 가능성을 크게 줄였다.\n 하지만 2세대 수술보다 치...	연세사랑병원이 국내 최초로 도입한 3D 맞춤형 인공관절 수술은 수술시간이 가장 짧은...
10798	1917년 러시아 혁명으로 핀란드가 독립하자 귀국했지만 너무 오랫동안 사용하지 않...	핀란드 내전 당시 적위대를 물리치고 핀란드 건국의 초석을 놓은 그를 데려온 것은 신...
10799	직장인 최주은(28·가명)씨의 주식계좌는 밤낮없이 일한다.\n 국내 주식과 미국 ...	코로나19로 세계 증시가 폭락한 뒤 반등을 보이자 해외 주식 투자에 거리낌 없이 뛰...

뉴스의 본문과 2~3문장 요약 데이터

02 Dataset

NAVER 금융

종목명·지수명 입력

통합검색

로그인

금융 홈

국내증시

해외증시

시장지표

리서치

뉴스

MY

연기, 냄새 걱정없는 생선오븐

인기 많은데 이유가 있죠! 겹바속속 생선오븐의 완성

지금 구매하기 >

AD X

뉴스

실시간 속보

주요뉴스

뉴스포커스

시황·전망

기업·종목분석

해외증시

채권·선물

공시·메모

환율

많이 본 뉴스

포토뉴스

TV뉴스

투자정보

공시정보

장종특징주

증시일정

뉴스검색

뉴스로 보는 증시일정

오늘의 증시일정

금융홈 > 뉴스 > 주요뉴스

속보

산한금투, 리테일 채권판매금액 10...

< >

가장 많이 본 뉴스

더보기

[특징주] 카카오 그룹 동반 ...

[특징주] '매파' 파월에 삼성...

[특징주] 도화엔지니어링, '...

[특징주] 고려신통정보, 美 ...

[특징주] 박셀바이오, 신규 ...

[특징주] 세아베스틸지주, ...

[특징주] 금비, 유럽 천연가...

[특징주] 네오위즈 'P'의 거...

인기검색어

더보기

삼성전자 58,600 ▼

카카오 72,200 ▼

한국조선해양 97,100 ▼

LG전자 96,200 ▼

LG화학 604,000 ▼

주요뉴스

파월 한마디에 무너졌다...일본·대만도 2%대 급락[Asia 마감]

[日 닛케이225 2.66%↓, 대만 증시 2.31%↓...중화권 혼조세, '상하이' 오르고 '선전·홍콩' 하락]인플레이션을 잡기 위해 금...

머니투데이 | 2022-08-29 17:23:01

[마켓뷰] 미국궤 '검은 월요일'...코스피·코스닥 2% 넘게 급락

외국인, 유가증권시장서 열흘 만에 '팔자' 日니케이225, 홍콩H, 美 선물도 하락파월 매파적 발언...시장 참여자 "다음 달 자이언트스텝 ...

조선비즈 | 2022-08-29 16:49:01

[유럽개장] '파월 충격'에 일제히 내림세...유로스톡스50 1.2%↓

제롬 파월 미국 연방준비제도(Fed) 의장 [이미지출처=연합뉴스] 29일 유럽 증시가 장 초반 일제히 내림세를 나타내고 있다. 29일 오후...

아시아경제 | 2022-08-29 16:31:44

[시황종합] 파월 쇼크에 코스피·코스닥 2%대 하락...환율 1350원 돌파

코스피 2426선·코스닥 779선 마감...달러·원 환율 장중 1350.8원◎ News1 이지원 디자인(서울=뉴스1) 이기림 기자 = 국...

뉴스1 | 2022-08-29 16:07:00

[코스닥 마감] 킹달러·매파 발언에 코스닥 종목 86% 내렸다

전거래일보다 2.81% 내린 779.89에 거래 마쳐개인·외국인 쌍끌이 사자에도 기관 1598억 '팔자' 美금리인상 가능성에 반도체주 부...

이데일리 | 2022-08-29 16:01:01

뉴스로 보는 증시일정

오늘의 증시일정

코스닥 마감] 킹달러·매파 발언에 코스닥 종목 86% 내렸다

전거래일보다 2.81% 내린 779.89에 거래 마쳐개인·외국인 쌍끌이 사자에도 기관 1598억 '팔자' 美금리인상 가능성에 반도체주 부...

이데일리 | 2022-08-29 16:01:01

외환마감] 파월의 '매'도 무섭구나...환율, 13년 4개월래 1350원 돌파

환율, 19.1원 올라 1350.4원 넘어1380원이 다음 번 저항선 될 듯외환당국 경계감 커졌으나 환율 급등세 못 막아달러인덱스 10...

이데일리 | 2022-08-29 15:56:02

美파월이 부른 '검은 월요일'...日·대만 2% 넘게 급락 [Asia오전]

제롬 파월 미 연준 의장/AFPBBNews=뉴스129일 오전 아시아 주요 증시에 '검은 월요일' 공포가 짙게 드리우고 있다. 제롬 파월...

머니투데이 | 2022-08-29 11:48:59

일일펀드동향] 국내 채권형 펀드, 하루 동안 2020억 순유입

국내 채권형 펀드에 하루 동안 약 2000억원이 들어왔다. 29일 금융투자협회에 따르면 지난 25일 기준 상장지수펀드(ETF)를 제외한 ...

아시아경제 | 2022-08-29 10:30:17

오전시황] 파월 '매파본능'에 코스피 와르르...2%대 급락 2420선 '위태'

제롬 파월 미국 연방준비제도(연준) 의장./연합뉴스[서울경제] 제롬 파월 미국 연방준비제도(연준) 의장의 '매파' 연설에 코스피가 와르...

서울경제 | 2022-08-29 09:38:01

외환브리핑] 10원 급등한 역외환율...연고점 테스트 vs 외환당국 개입이나

미 파월 매파 발언에 달러인덱스 109선뉴욕 선물 지수도 하락세 지속주요국 긴축 강도 세져...위험자산 회피 심리존 윌리엄스 미국 뉴욕 연...

이데일리 | 2022-08-29 08:17:01

굿모닝 증시] '파월 '매파' 연설에 하락 출발 예상'...낙폭 축소 전망도

[이미지 출처=로이터연합뉴스] 제롬 파월 미국 연방준비제도(Fed) 의장이 잭슨홀 미팅 연설에서 매파 연설 여파가 이어지고 있다. 파월 ...

아시아경제 | 2022-08-29 08:06:09

1 |

08월29일(월)

08월28일(일)

08월27일(토)

08월26일(금)

08월25일(목)

네이버 금융 뉴스

05

02 Dataset

```
import pandas as pd
raw = {'url':url_list,
       'title':title_list,
       'text':text_list}
```

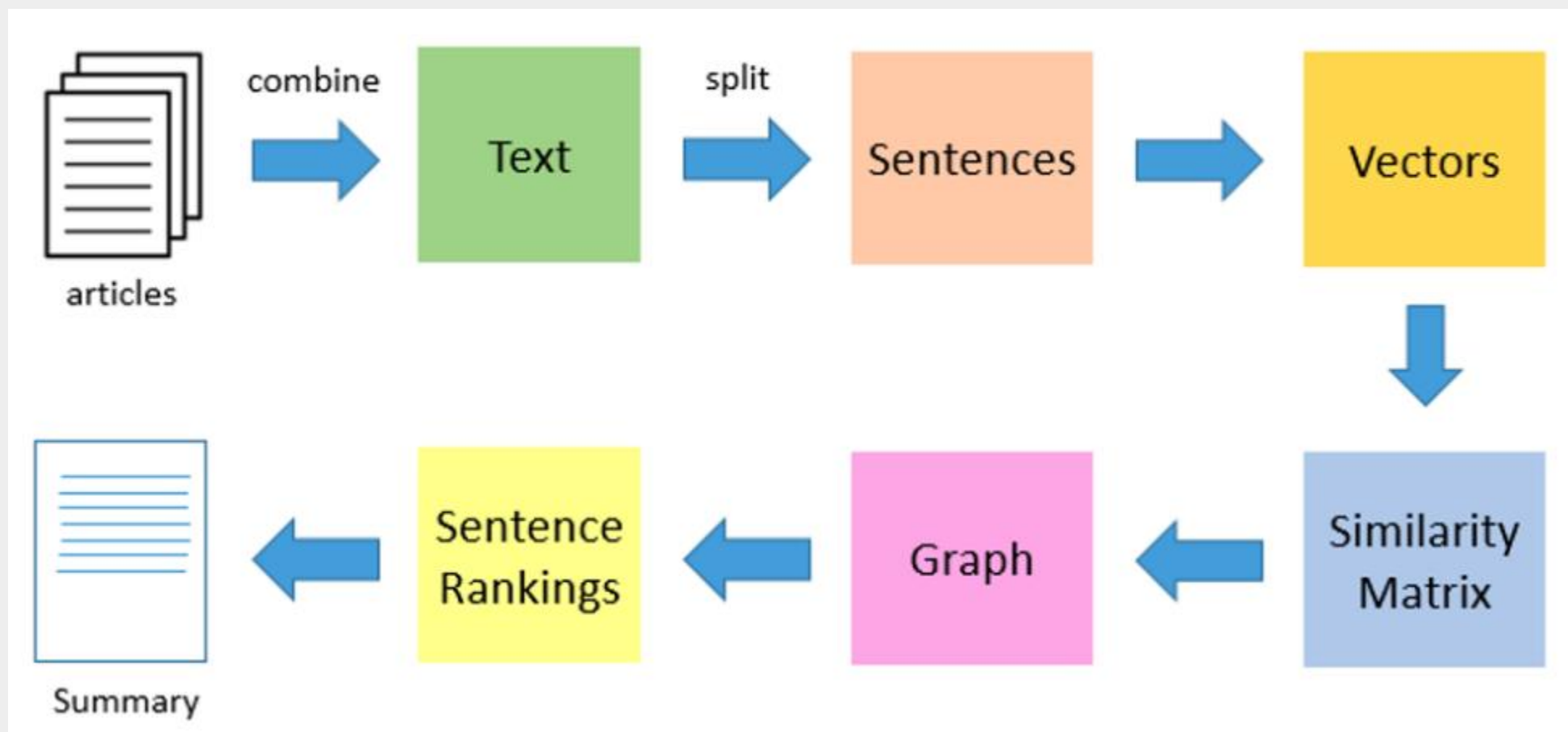
```
data=pd.DataFrame(raw)
data.head()
```

	url	title	text
0	https://finance.naver.com/news/news_read.naver...	잭슨 홀 리스크·유럽 침체우려에도 나스닥 소폭 상승 출발 [데일리 국제금융시장]	다우 0.065%·S&P 0.069% 상승10년 물 국채 3.06%로 올라유로존 8...
1	https://finance.naver.com/news/news_read.naver...	[Asia마감] 긴축 공포에 짓눌린 시장...닛케이 1.19%↓	23일 아시아 주요국 증시가 일제히 하락 마감했다. 통화긴축, 경기침체에 대한 우려...
2	https://finance.naver.com/news/news_read.naver...	[마켓뷰] 코스피·코스닥 1% 넘게 하락...치솟는 美 달러화 가치에 투자금 '썰물'	유럽, 500년 만에 최악의 가뭄...경기 우려 심화유로화 가치 20년 만에 최저치로,...
3	https://finance.naver.com/news/news_read.naver...	[마감시황]코스피, 강달러 지속에 1%대 하락...2430선 마감	[서울=뉴시스] 조성우 기자 =코스피가 전 거래일(2462.50)보다 27.16포인트...
4	https://finance.naver.com/news/news_read.naver...	[유럽개장]장 초반 하락세...獨 0.35%↓	23일 유럽 증시가 장 초반 하락세를 보이고 있다.이날 오후 4시20분(한국시간) ...

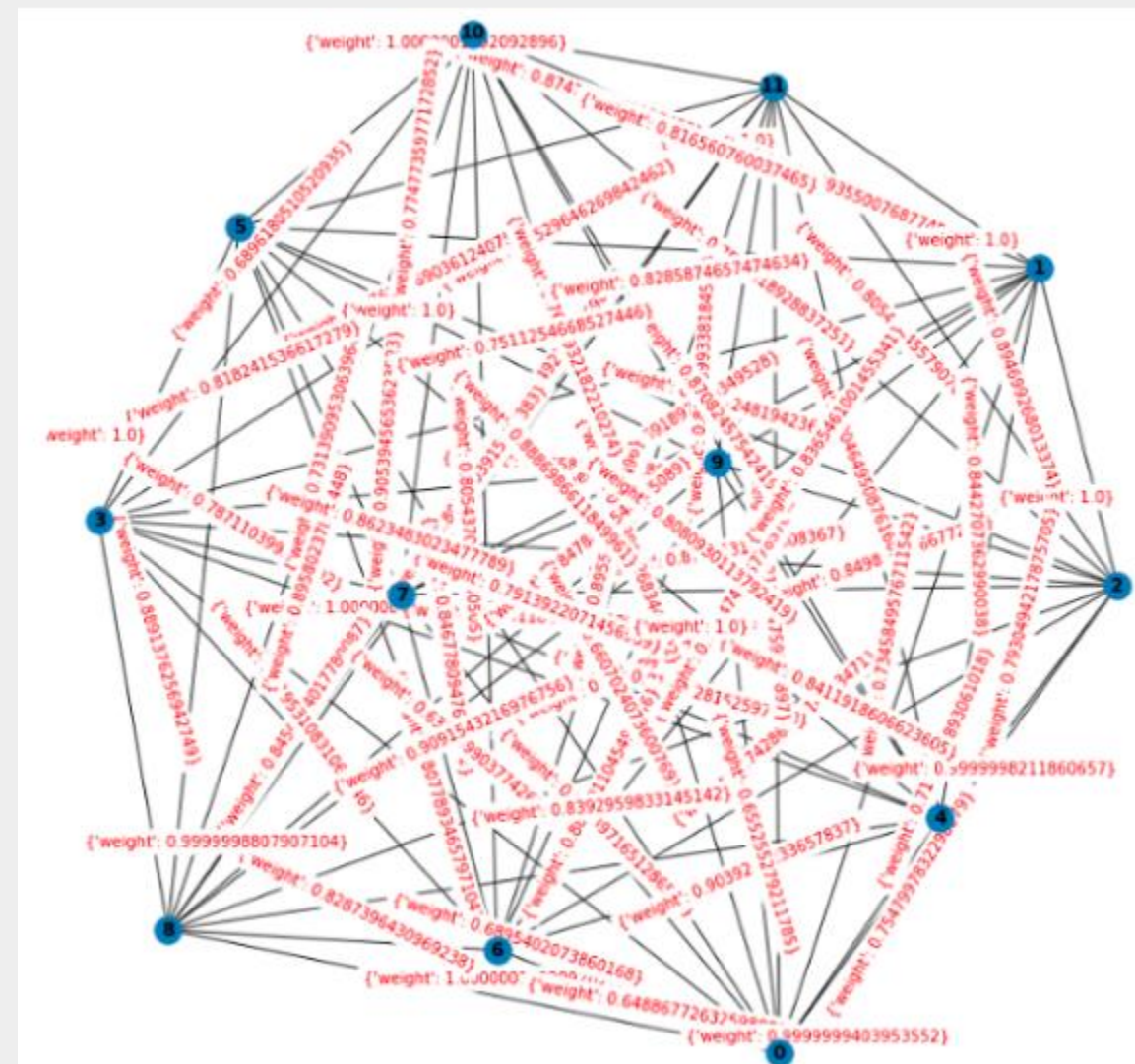
selenium 사용해서 크롤링

03 모델

Textrank 추출요약



그래프 기반의 랭킹 모델로 샘플의 한 문장마다
순위를 매겨서 요약문을 추출하는 방법



$$S(V_i) = (1 - d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

문장과 문장 사이의 가중치를 구하고
텍스트랭크 점수를 계산하여
내림차순으로 정렬
상위 n개의 문장으로 요약

03 모델

Textrank 추출요약

```
tfidf= TfidfVectorizer()  
cnt_vec= CountVectorizer()  
graph_sentences=[]  
  
def build_sent_graph(sentences):  
    tfidf_mat = tfidf.fit_transform(sentence).toarray()  
    graph_sentence = np.dot(tfidf_mat, tfidf_mat.T)  
    return graph_sentence  
sent_graph = build_sent_graph(nouns)
```

sent_graph

```
array([[1.          , 0.03386508, 0.0351497 , ..., 0.05366189, 0.05922813,  
        0.02774264],  
       [0.03386508, 1.          , 0.04076138, ..., 0.          , 0.03434197,  
        0.07227906],  
       [0.0351497 , 0.04076138, 1.          , ..., 0.05023318, 0.16456149,  
        0.05145673],  
       ...,  
       [0.05366189, 0.          , 0.05023318, ..., 1.          , 0.          ,  
        0.          ],  
       [0.05922813, 0.03434197, 0.16456149, ..., 0.          , 1.          ,  
        0.01751071],  
       [0.02774264, 0.07227906, 0.05145673, ..., 0.          , 0.01751071,  
        1.          ]])
```

TF-IDF 모델로 문장의 단어를 보고
문장의 중요도를 판단

중요도가 높은 3개 문장으로 요약

정답 요약문을 보고 학습하는 것이
아니기 때문에 한계점 존재

03 모델

koBart



BERT와 GPT가 합쳐진 형태의 모델로, 입력 텍스트 일부에 노이즈를 추가하여 이를 다시 원문으로 복구하는 AutoEncoder 형태로 학습

KoBART는 BART의 text infilling 노이즈 함수를 이용하여 한국어 텍스트에 대해서 학습한 encoder-decoder 언어 모델

AI Hub 데이터로 학습, fine tuning 진행 후 모델 저장해 크롤링 된 뉴스로 테스트

03 모델

koBart - 추출 요약

추출 요약은 원문에서 중요한 핵심 문장 또는 단어구를 몇 개 뽑아서 이들로 구성된 요약문을 만드는 방법. 3개의 문장을 추출해 요약

예시) 기사 제목 : 코스피, 강달러 지속에 1%대 하락...2430선 마감

>> 23일 코스피는 전 거래일 대비 27.16포인트(1.10%) 내린 2435.34에 거래를 마쳤다.

이날 13.19포인트 하락 출발한 지수는 장 초반 2440선 위쪽에서 움직였지만 오후 2시를 기점으로 낙폭을 확대하면서 1% 넘게 떨어졌다.

원·달러 환율이 강세 흐름을 이어가면서 시장 전반의 투자심리가 위축됐다.

03 모델

koBart - 생성 요약

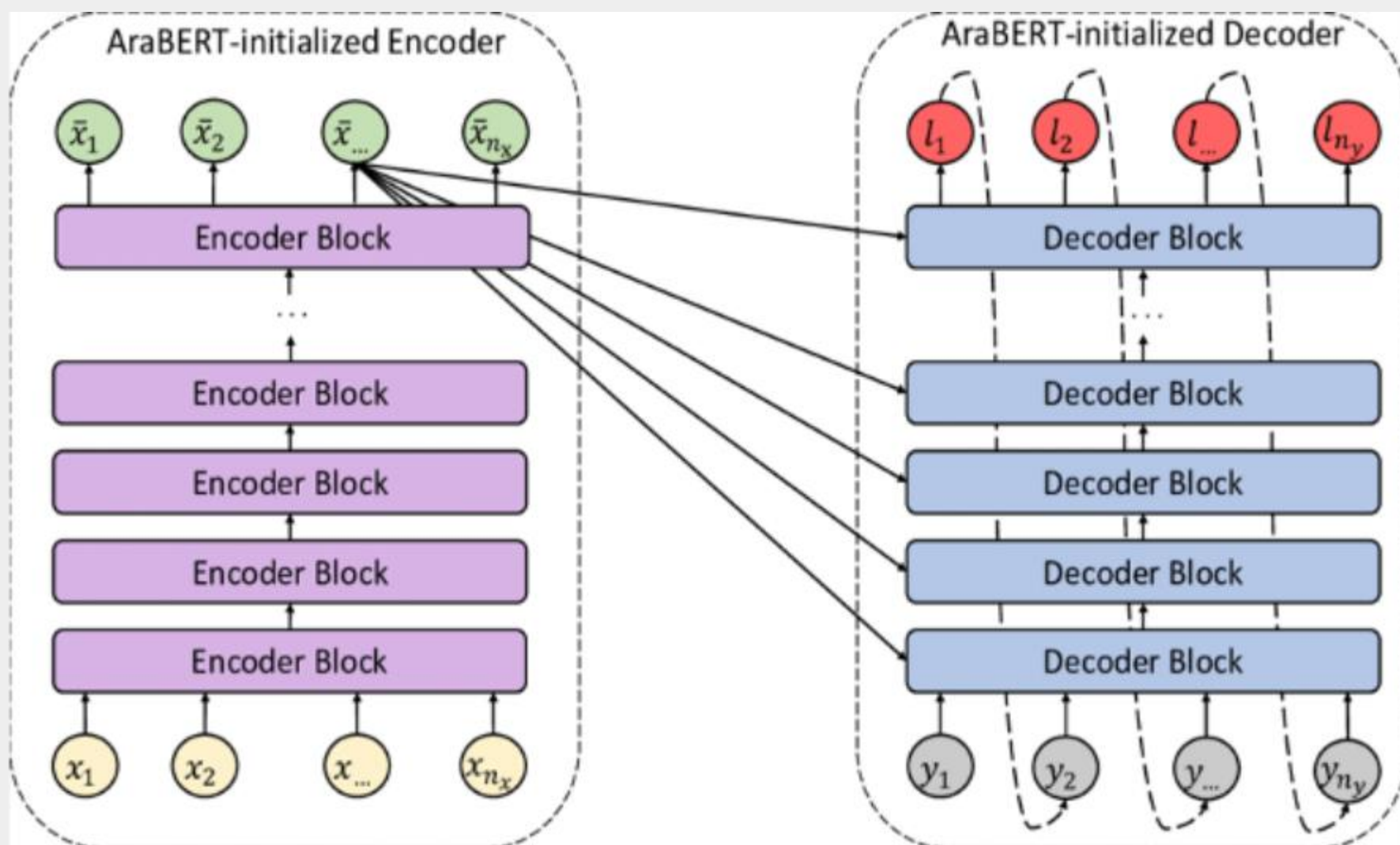
추상적 요약(생성 요약)은 원문에 없던 문장이라도 핵심 문맥을 반영한 새로운 문장을 생성해서 원문을 요약하는 방법.

예시) 기사 제목 : 코스피, 강달러 지속에 1%대 하락...2430선 마감

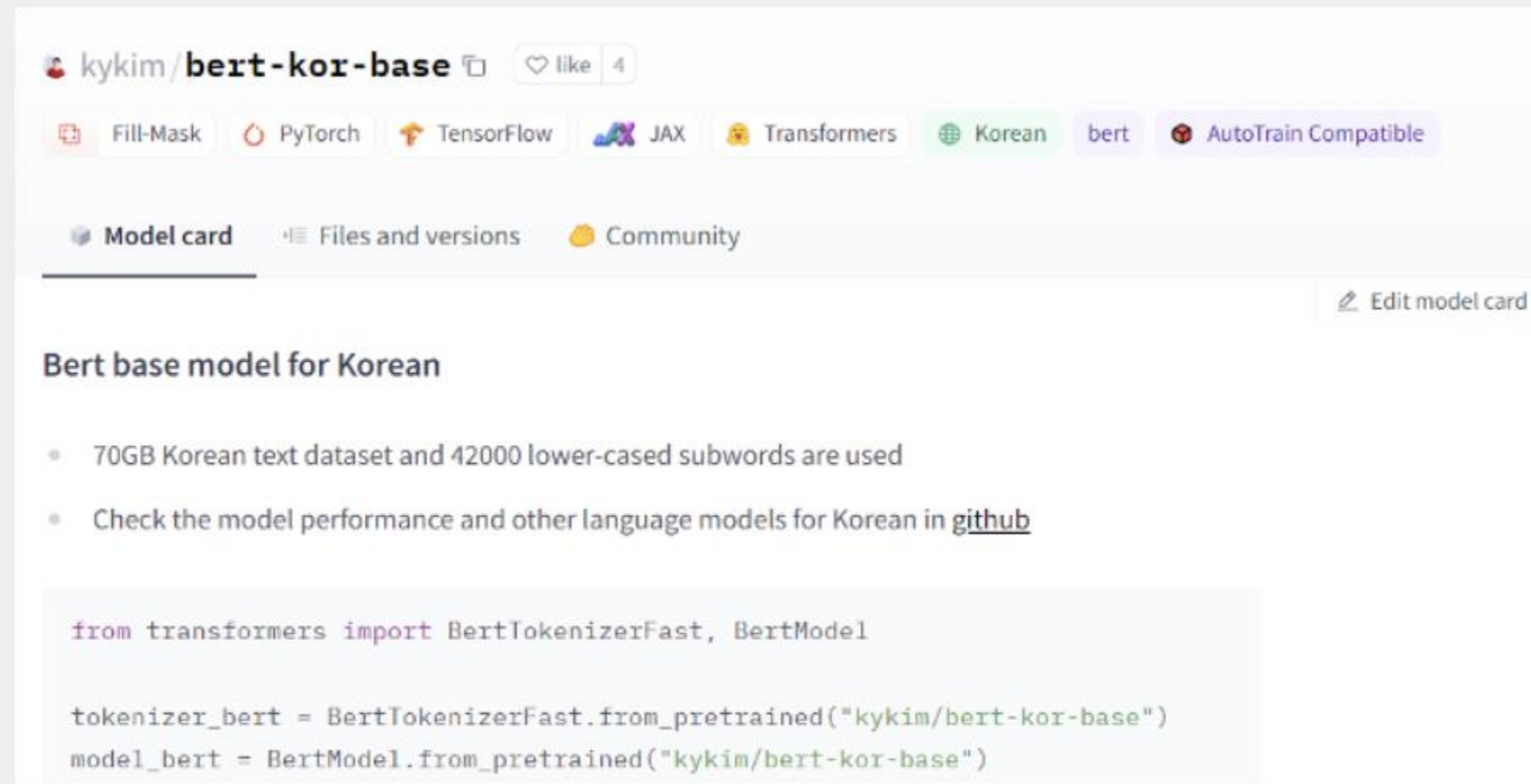
>> 코스피가 전 거래일보다 27.16포인트(1.10%) 내린 2435.34에 거래를 마쳤으며 원·달러 환율이 강세 흐름을 이어가면서 시장 전반의 투자심리가 위축되어 원·달러 환율은 전 거래일보다 5.7원 오른 1345.5원에 거래를 마쳤다.

03 모델

KoBertShared 생성 요약



Bert기반으로 한 EncoderDecoder 모델로,
Bert2Bert 모델에서 더 나아가
인코더와 디코더 간 파라미터를 서로 공유



bert-kor-base 한국어로 Pre-train된 모델을 활용

03 모델

KoBertShared 생성 요약 fine tuning

{'passage': '더불어민주당이 뉴딜 펀드 띄우기를 시작했다.Wn 이해찬 대표를 비롯한 민주당 지도부는 5일 서울 여의도 한국거래소에서 현장간담회를 열고 민간 참여를 당부했다.Wn 이 대표는 인사말에서 "(한국형 뉴딜 사업에) 2025년까지 약 160조원이 투자될 거라고 보는데 상당 부분은 정부 재정이지만, 민간 부분에서도 10% 가까이 투자를 하는 구상을 하고 있다"고 말했다.Wn 문재인 정부의 임기 말 중점 사업들에 뉴딜 펀드로 민간 자금 16조원 이상을 끌어오겠다는 것이다.Wn 민주당은 이날 간담회에 기획재정부, 금융위원회 등 정부 부처와 금융투자협회, 은행연합회, 경영자총연합회, 신용보증기금 등 금융권 주요 협회 관계자를 두루 불렀다.Wn 최영권 우리자산운용 대표, 김재익 KDB인프라자산운용 대표 등 펀드 운용사 대표들도 나와 뉴딜 펀드의 구체적 설계 방향을 제시하고 토론에 참여했다.Wn "개인뿐 아니라 연기금, 기관투자자 모두에게 좋은 '1석 3조' 펀드"(홍성국 민주당 의원)라는 메시지를 강조하려는 의도다.Wn 이호형 은행연합회 전무이사는 간담회 직후 통화에서 "부동산 대책 이후에 이런 (그린·디지털 뉴딜 등) 신산업 쪽으로 물꼬를 틀겠다는 상징적 의미의 행사였다"고 했다.Wn ……

'summary': '비공개 토론에서 뉴딜 펀드에 대한 몇 가지 의문이 제기됐는데 전례 없는 형태의 관제 인프라 펀드라 시중 유동성이 큰 지금 같은 상황이 유지될지가 논의의 핵심이었다.'



'pred': ['더불어민주당 지도부는 5일 서울 여의도 한국거래소에서 열린 한석, 등과 함께 함께 같이 함께 하는 구상을 하고 있다....., 1 1 석 석 으로 만들어졌다 다시 제로를가 시작했다 시작하면서 하며 또한 이러한 다양한 형태의 해가 해가 대통령 선거에 보조주가 들을 적인 요소가 요소 것들을 들이 것들도 일어나면 일어나고 일어날 일어났다 일어날 가능성이 판단이 하기 ']]

메모리 사용량 초과로 절반의 데이터를 활용하여 fine tuning을 진행

실제 요약과 구현한 모델의 요약 간의 rouge metric 을 사용하여 loss 함수 설정

다소 아쉬운 요약 결과를 얻음

04 메일링 서비스 결과



안녕하세요 🙌

고려대학교 데이터사이언스 학회 "KUBIG" 입니다!

2022년 하반기 🏆 KUBIG CONTEST 🏆 를 맞이하여

NLP 팀이 제공하는 📰 금융 주요 뉴스 요약 메일링 서비스 📧 입니다.

제목	요약
파월 잭슨 홀 연설 앞두고 나스닥 소폭 하락 출발 [데일리 국제금융시장]	파월, 동부시간 오전10시 연설7월 PCE 전년 6.3%, 전월 -0.1% 소비는 예상 밑돌아 0.1% 그쳐 제롬 파월 미 연방준비제도(Fed·연준) 의장의 잭슨 홀 미팅 연설을 앞두고 뉴욕증시의 주요 지수가 하락 출발했다. 26일(현지 시간) 미 경제 방송 CNBC에 따르면 이날 오전9시35분 현재 뉴욕증권거래소(NYSE)의 다우존스산업평균지수가 전날보다 58.12포인트(0.17%) 오른 3만3349.90에 거래되고 있다. 스탠더드앤드푸어스(S&P)500은 0.53포인트(0.013%) 상승한 4199.65인 반면 나스닥은 6.47포인트(-0.05%) 내린 1만 2632.80를 기록 중이다.

[\[마켓뷰\] 코스피, 美 잭슨 홀 경계감에 소폭 상승... 개인 홀로 차익 실현](#)

코스닥지수는 약세 전환하며 800선에 턱걸이로 마감했다. 외국인·기관 동반 순매수 삼전, 나흘 만에 '6만전자' 코스닥, 오후 들어 낙폭 키워 코스피지수가 소폭 상승 마감했다. 26일 서울 중구 하나은행 명동점 딜링룸 전광판에 코스피 지수가 전일대비 3.77포인트(0.15%) 상승한 2,481.03을 나타내고 있다.

[\[마감시황\]코스피, 잭슨홀 미팅 앞두고 고...2500선 턱밑까지 반등](#)

사상 처음 4번 연속 기준금리 인상에도 외인과 기관의 매수세가 이어진데 따른 것으로 풀이된다. 기사내용 요약코스피 한때 2496.76까지 상승외인·기관 쌍끌이 매도세에 반등11시 제롬 파월 연설에 주목 이승주 기자 = 잭슨홀 미팅을 앞두고 코스피가 3거래일 연속 반등세를 이어갔다. 26일 코스피는 전 거래일(2477.26)보다 3.77포인트(0.15%) 오른 2481.03에 장을 닫았다.

[\[코스닥 마감\]3거래일 만에 하락... 잭슨홀 경계심리 고조](#)

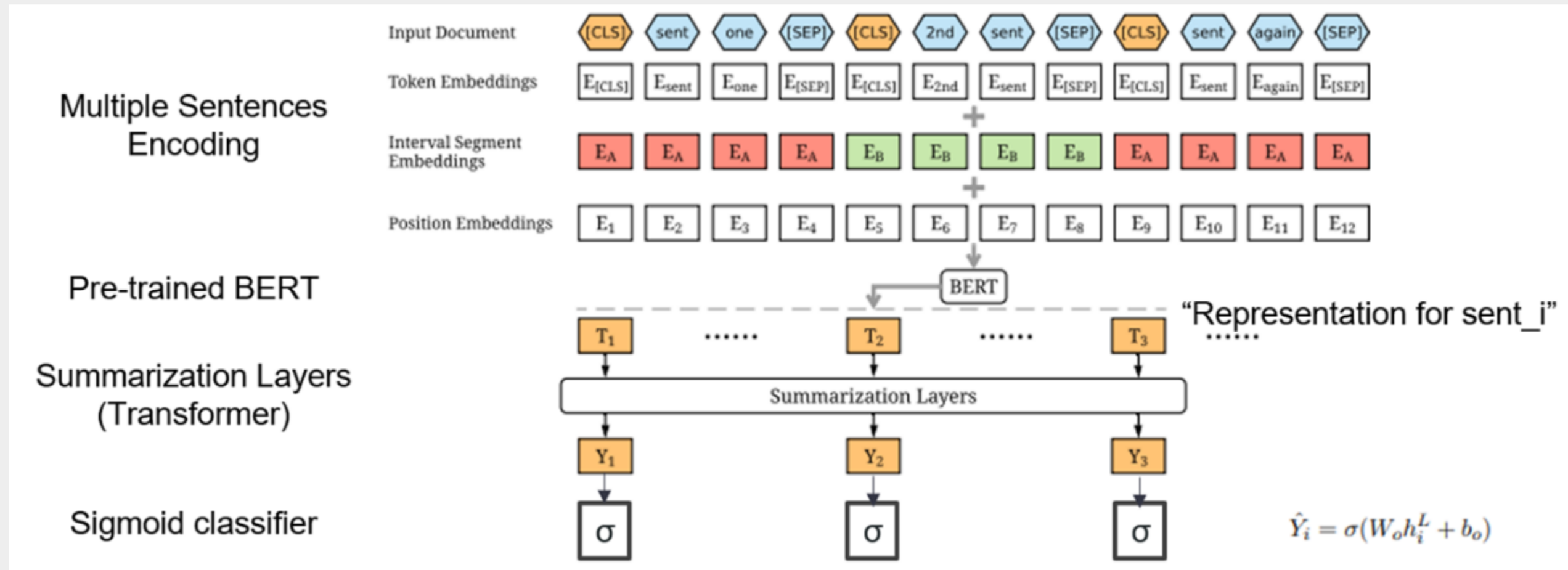
26일 마켓포인트에 따르면 코스닥 지수는 전거래일 대비 4.92포인트(0.61%) 하락한 802.45로 거래를 마쳤다. 이날 코스닥은 상승 출발했지만 오후 들어 하락 전환하며 810선이 붕괴됐다. 기관·외국인 '팔자'...개인 1085억 '사자'유통, 종이·목재 등 1%대 약세...운송·화학 강세셀트리온헬스케어 4%↓...현대바이오 15%대↑유럽 비료 공급망 축소에 비료株 강세 코스닥이 3거래일 만에 하락 마감했다.

[\[외환마감\]강달러에도 네고+구두개입...환율, 1300원대 3일째 하락](#)

잭슨홀 미팅 앞두고 달러화 강세 전환코스피 지수 외국인 순매수에 강보합기재부 구두개입성 발언에 당국 경계 원·달러 환율이 3거래일 연속 하락하면서 1330원대 초반으로 하락했다. 제롬 파월 연방준비제도(Fed) 의장의 잭슨홀 미팅 연설을 앞두고 달러인덱스는 108선에서 다시 상승폭을 키우고 있지만, 상단의 네고(달러 매도) 물량과 외환당국의 구두개입성 발언 등으로 하락 압력이 우위를 보였다. 글로벌 달러인덱스는 파월 의장의 연설이 우리시간으로 26일 밤 11시 예정된 가운데 108선에서 상승폭을 키우고 있다.

koBart 추출요약 모델 선택
당일 금융 뉴스의 제목과 요약을 이메일로 전송

05 Future Work



koBertSum이란?

- BERT 위에 inter-sentence Transformer 2-layers를 추가로 얹은 구조

- 요약 task를 위한 임베딩 layer로 활용하기 위해서는 여러 문장을 하나의 input으로 넣어주고, 각 문장에 대한 정보를 출력할 수 있도록 입력을 수정해줘야 함

>> 매 문장의 앞에 [CLS] 토큰 삽입

>> 매 문장마다 다른 segment 임베딩 토큰을 더해주는 interval segment 임베딩 추가

현재 사용하려는 데이터셋의 구조와 KoBertSum를 사용하려면 필요로 구축해야 하는 데이터셋의 구조 차이로 구현 못한 아쉬움

05 Future Work

- Google Colaboratory RAM 부족 및 런타임 중단으로 보다 큰 epoch, batchsize에서 모델을 훈련시켜보지 못한 아쉬움이 남음
- KoBART, KoBERTSUM, TextRank 등 모델을 합쳐 ensemble 형태로도 요약해보고 싶음
- 뉴스 기사처럼 정형화된 데이터 이외에도 블로그 글, 메신저 대화 등 구성이 다양한 데이터도 전처리해서 요약 모델을 만들어보고 싶음



Thank You