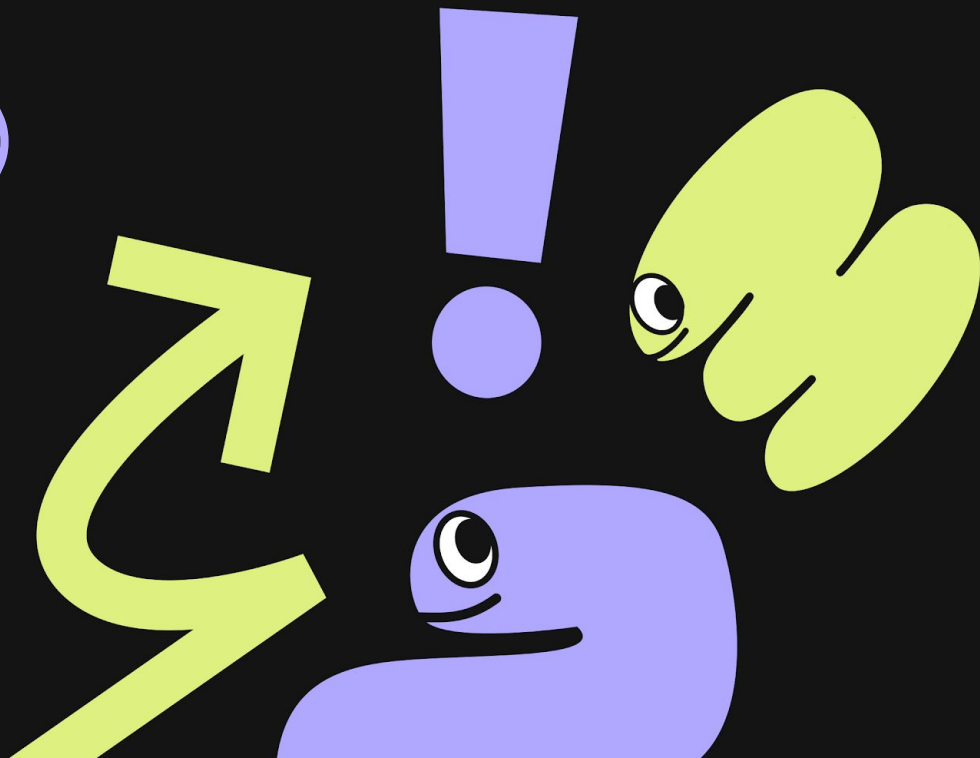


PyConKR 2024

인공지능과 파이썬으로 금융 데이터 분석해보기 with 자연어처리 (NLP)

Analyzing financial data with
AI & Python with NLP

김대현 (Daehyun Kim)



김대현 (Daehyun Kim)

Ai Engineer & Researcher, Student

bigdarkgold@gmail.com

MODULABS Researcher. (MODUAI Lab)

KakaoTech BootCamp. (1st, GenAI Tech)

GopherCon Korea 2024 Organizing Committee

- with GDG Golang Korea

Hankuk University of Foreign Studies.

- Computer Engineering, DS

Former Robotics Engineer, JRC.



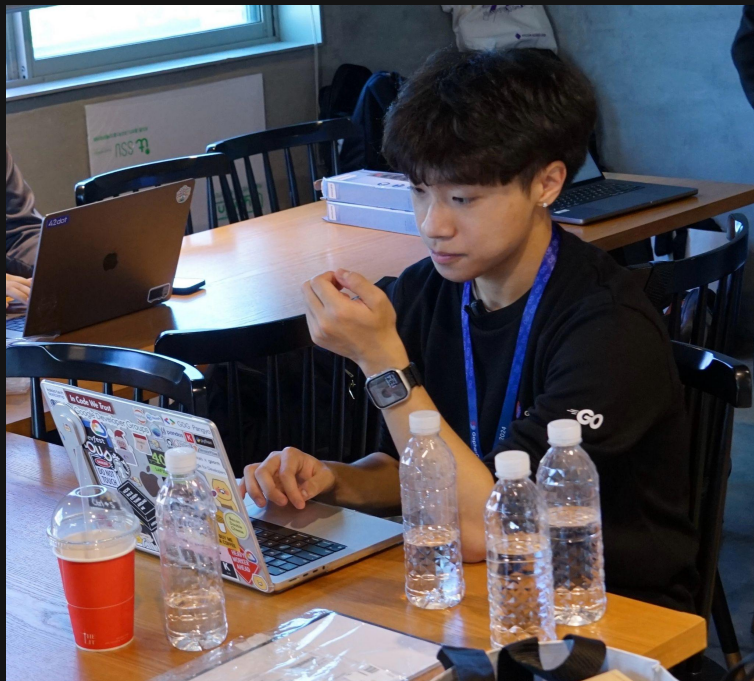
Daehyun Kim



Daehyun-Bigbread



developer._toby



In this talk, we will treat..

Agenda

Problem Recognition

Dev & Research Purpose

Data Collect

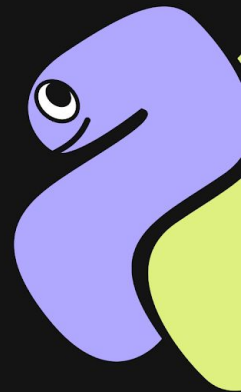
TF-IDF, N-Gram

KoBERT

Denosing Method

Data Implementation

Positive Index Reliability



PART 1

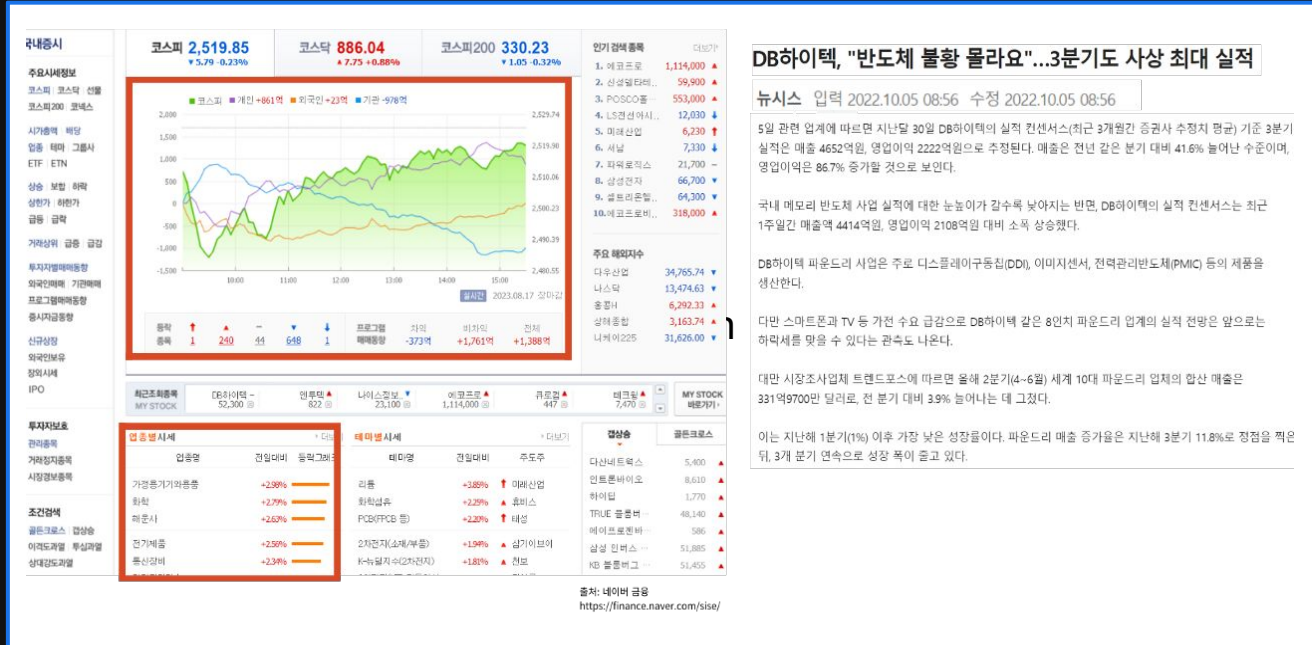
Problem Recognition & Purpose

문제인식 & 목적



Problem Recognition

문제 인식



The screenshot shows a financial dashboard with the following elements:

- Market Indices:** 코스피 2,519.85 (▼ 5.79 -0.23%), 코스닥 886.04 (▲ 7.75 +0.88%), 코스피200 330.23 (▼ 1.05 -0.32%).
- Line Chart:** A line chart showing market performance from 8:00 to 15:00. A red box highlights the chart area.
- Table of Changes:**

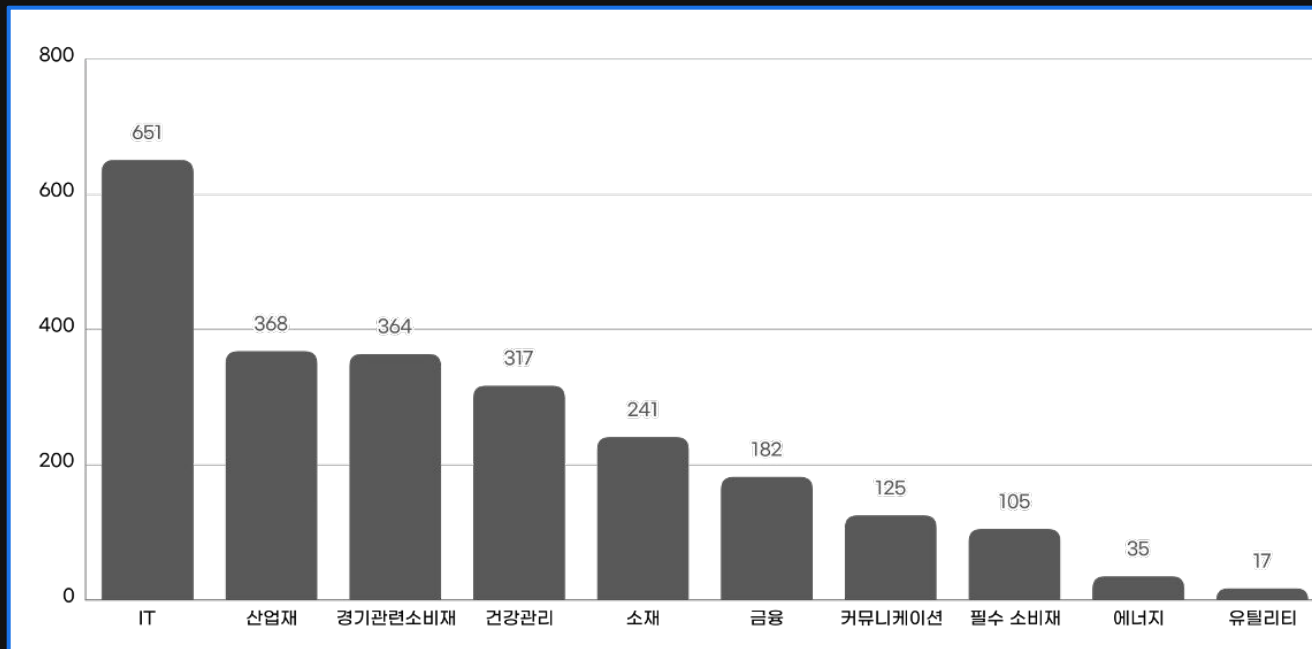
종목	↑	▲	-	▼	↓	프로그램	변동	가치	비율	변화
종목	1	240	44	649	1	해당종목	-37%	+1,761억	+1,388%	
- News Article:** DB하이텍, "반도체 불황 몰라요"...3분기도 사상 최대 실적. The article discusses DB Hi-tech's record performance in Q3 despite semiconductor market challenges.
- Table of Stocks:**

업종	연일대비	동일(2분기)	대기업	연일대비	주도주
가공정기차용종	+2.88%		리움	+3.88%	미래산업
화학	+2.99%		화학공업	+2.29%	휴비스
의약	+2.67%		PCB(PCB 등)	+2.20%	태성
전기차용	+2.56%		2차전지(소재/부품)	+1.94%	삼기티브이
통신장비	+2.34%		K-뉴딜지(국산전자)	+1.81%	천보

문제점: 복잡한 리서치 화면 및 뉴스 및 커뮤니티를 통한 투자 결정의 어려움
 Complex screens and difficulties in making investment decisions through news & communities

Problem Recognition

문제 인식



문제점: 2400개 종목을 10개 부문으로만 구분하여 의사결정이
어려운 점

It is difficult to make decisions by dividing 2400 items into 10 categories

Dev & Research Purpose

개발 & 연구 목적 요약



주요 정보

Easy Access



정확한 투자 동향

Grasp



유망한 분야

Analysis

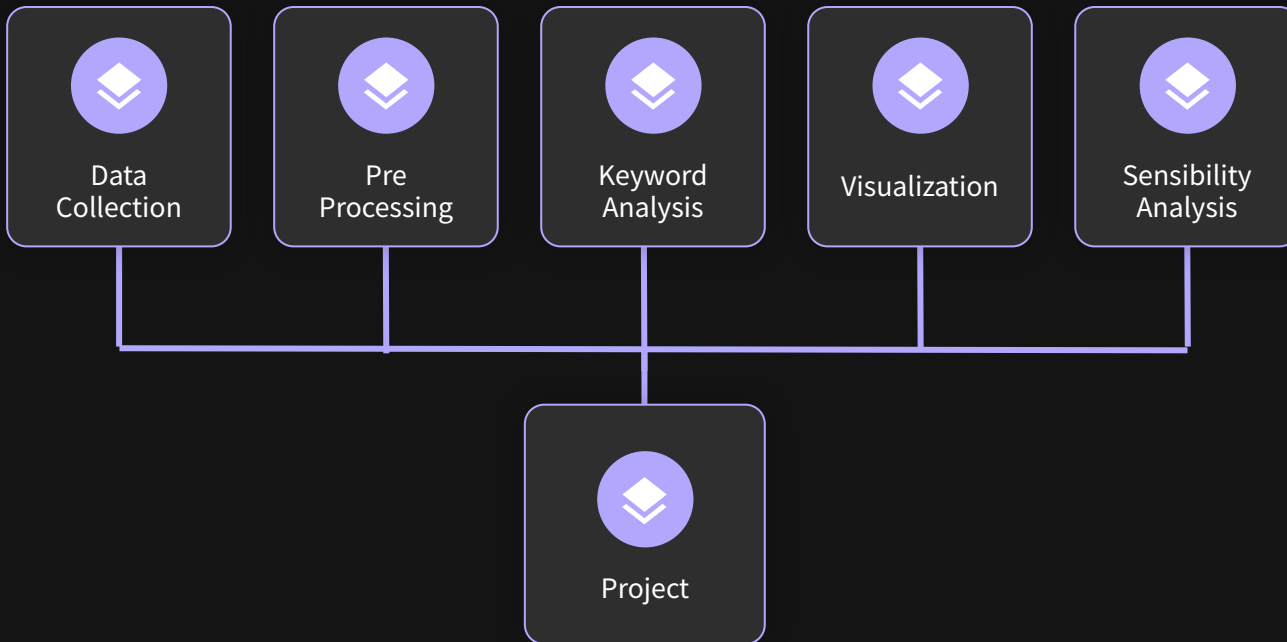


상관 관계 분석

Public Opinion

Overall of Project

프로젝트의 전체 구조



PART 2

Technology & Techniques

사용된 기술 & 기법



Dev Design & Research Method

개발 설계 & 연구 방법



Data Collection

데이터 수집

```
import pandas as pd
```

```
#한국거래소에서 종목 코드 받아옵니다. [0]은 헤더를 첫번째 행으로 지정하기 위해 사용 했습니다.
```

```
code_df = pd.read_html('http://kind.krx.co.kr/corpgeneral/corpList.do?method=download&searchType=13')[0]
```

```
code_df
```

회사명	종목코드	업종	주요제품	상장일	결산월	대표자명	홈페이지	지역
0	시넥트릭스	산업용 기계 및 장비 임대업	원발(머릿머, OA장비, 건설장비)	2014-08-21	12월	손성일	http://www.ajnet.co.kr	서울특별시
1	BNK금융지주	기타 금융업	금융지주회사	2011-03-30	12월	반대연	http://www.bnkfg.com	부산광역시
2	DSR	1차 보험금 수. 제조업	합성섬유로브	2013-05-15	12월	홍석범	http://www.dsr.com	부산광역시
3	GS	기타 금융업	지주회사/부동산 임대	2004-08-05	12월	허태수, 홍순기 (차지 대표이사)	NaN	서울특별시
4	HDC현대산업개발	건설 건설업	위주주택, 자재공시, 일반건축, 토목 등	2018-06-12	12월	최익훈, 정지희, 김희연 (감사 대표이사)	http://www.hdc-dvp.com	서울특별시
...
2817	카이바이오텍	의약품 제조업	병사정 진단 및 치료 의약품	2022-12-23	12월	김영덕	http://www.kabiotech.com	경남북도
2818	코스텍시스템	특수 목적용 기계 제조업	반도체 웨이퍼 이송장비, 분당장비	2022-01-21	12월	배용호	http://www.kosteks.com	경기도
2819	티입기술	소프트웨어 개발 및 운영업	ILS(종합군수지원), IETM(전자식 기술교범), CBT(민간사교육장), 기술연계	2021-06-24	12월	주영호	http://tinet.co.kr	경상남도
2820	테크연	연구 및 조형장치 제조업	LED조명장치	2018-12-21	12월	이재현, 박철(한지 대표이사)	http://www.techn.co.kr	대구광역시
2821	한국미래올리콜사	기타 보험금융 제조업	손나노스텝, 데이터센터 등	2019-10-28	12월	이훈정	http://www.kmpc.co.kr	경기도

2822 rows x 9 columns

```
df = pd.read_csv("sector별/필수소비재_sector.csv", encoding="cp949")
df
```

Company	items
0	MH에탄올
1	모나리자
2	선진
3	오뚜기
4	오리온
...	...
91	오에스피
92	이지출딩스
93	정다운
94	제주맥주
95	지어소프트

96 rows x 2 columns

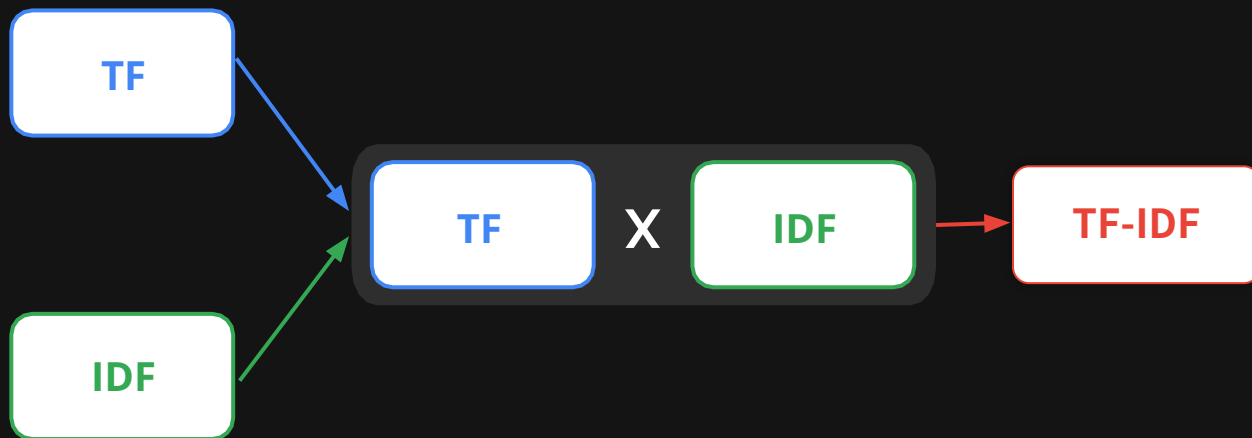
Mobile Solution & Platform, Internet GIS, Mobi...

공식 데이터 수집 및 결측치, 중복값 제거

Official data collection and missing values, deduplication

TF-IDF (단어 빈도 - 역문서 빈도)

Term Frequency - Inverse Document Frequency

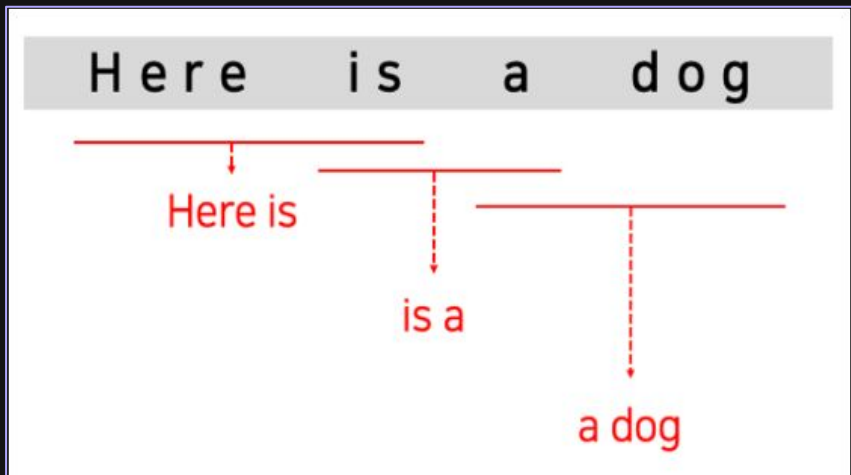


TF: 문서 내 단어의 빈도
frequency of words in the document

IDF: 전체 문서 집합에서의 단어 출현 빈도
frequency of word appearance in the entire document set

N-Gram (SLM 언어 모델)

SLM Language Models



Example of cutting reference units in an N-gram language model (Tomohiro Odakata, 2012)



단어 or 문장에 확률값을 할당

입력된 문장을 n개 단위로 잘라 분석

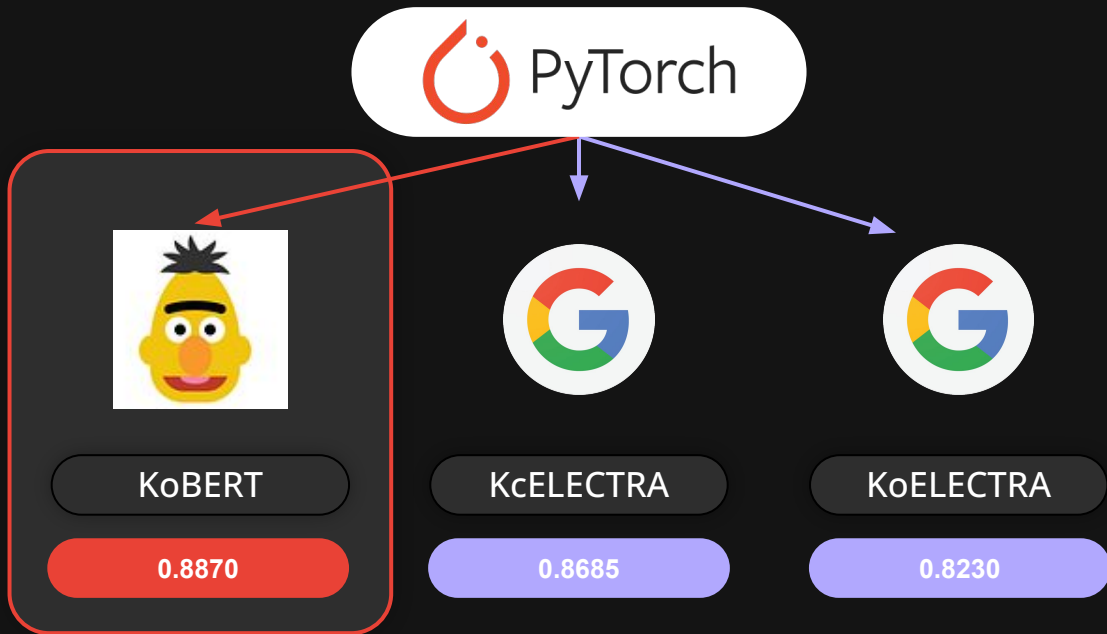
Assign probability values to words or sentences

Cut input sentences into n units and analyze them

모델별 성능 평가

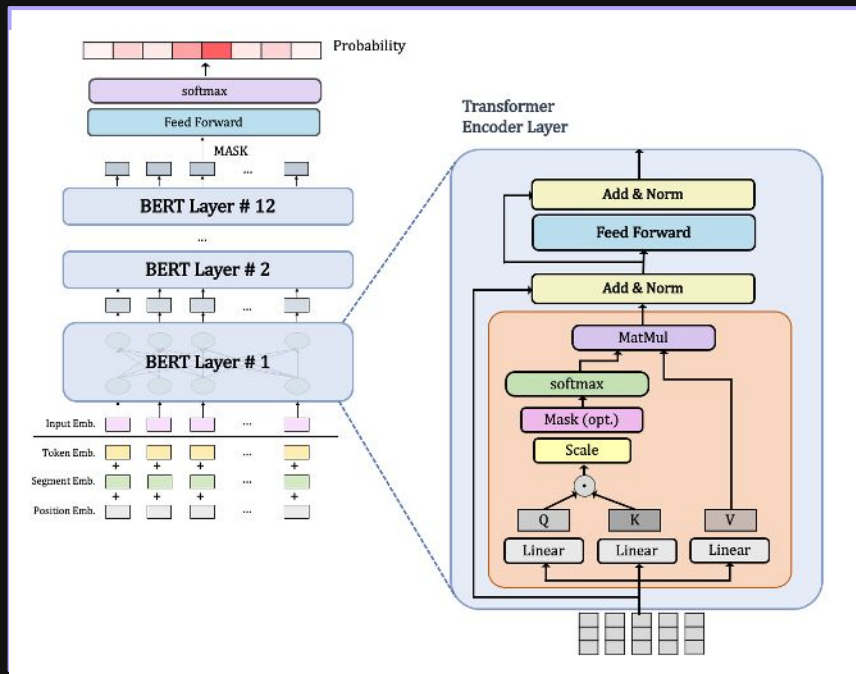
Performance Evaluation (By PyTorch)

금융 문장 데이터셋으로 모델별 성능 평가
Evaluate performance by model with financial sentence dataset

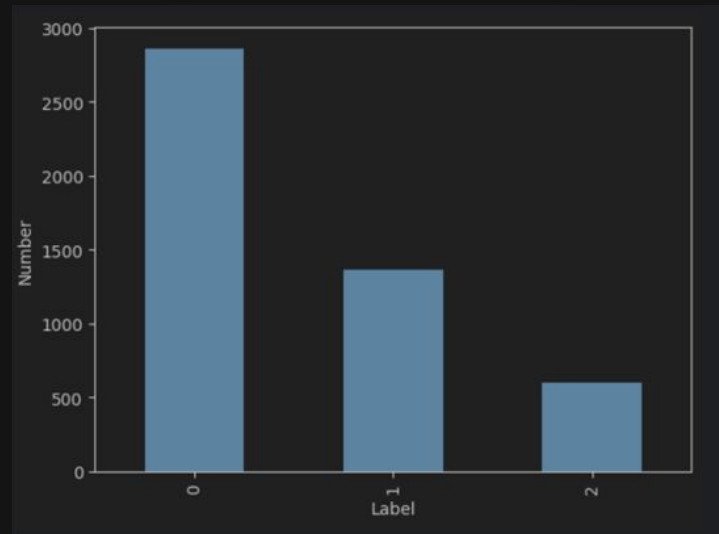


KoBERT

Korean-BERT - Pretrained Model



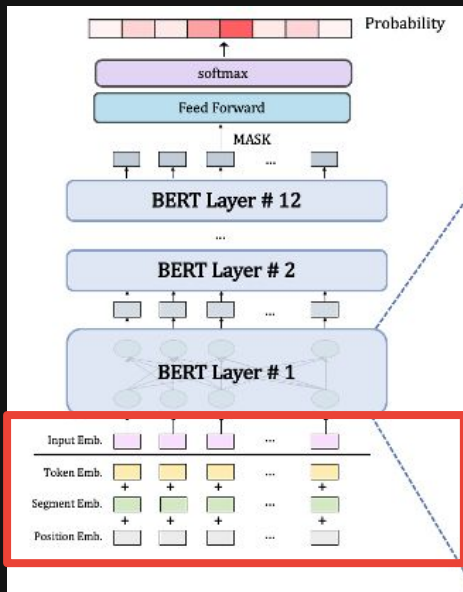
Emotional Analysis Results



0: Positive(긍정)
1: Netural(중립)
2: Negative(부정)

KoBERT

Korean-BERT - Input Embedding



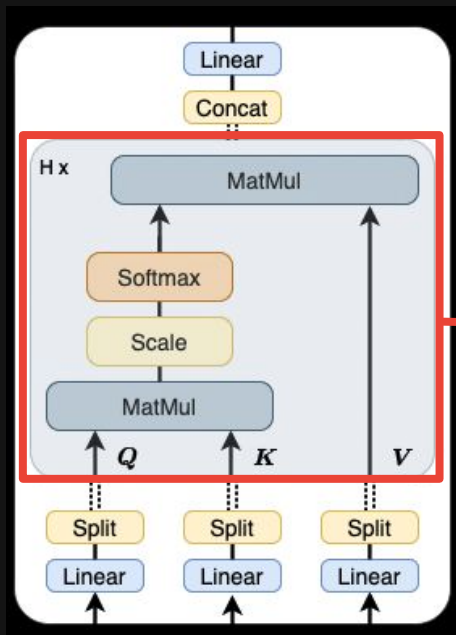
Tokenizer
(문장을 단어 단위로 분리)

“오늘”, “주식”, “시장”, “이”, “크게”, “상승했습니다”, “.”

Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#ing}$	$E_{[SEP]}$
Segment Embeddings	+	+	+	+	+	+	+	+	+	+	+
	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

KoBERT

Korean-BERT - Self-Attention



BERT Encoder Layer

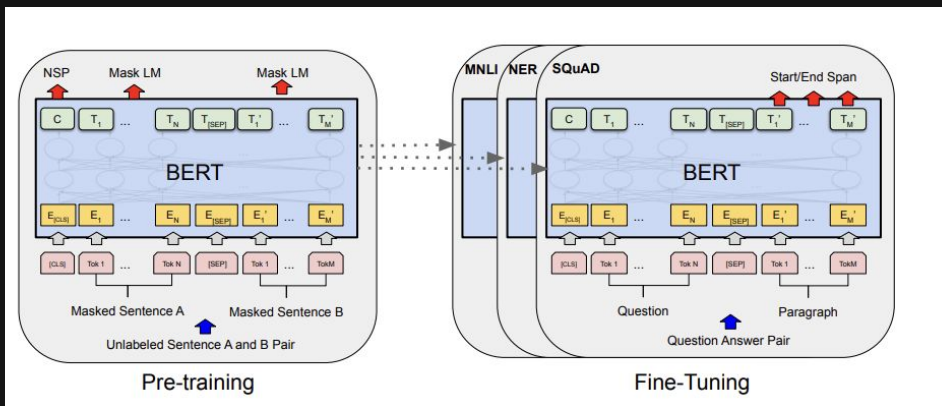
Encoder에서 Self-Attention 메커니즘



KoBERT

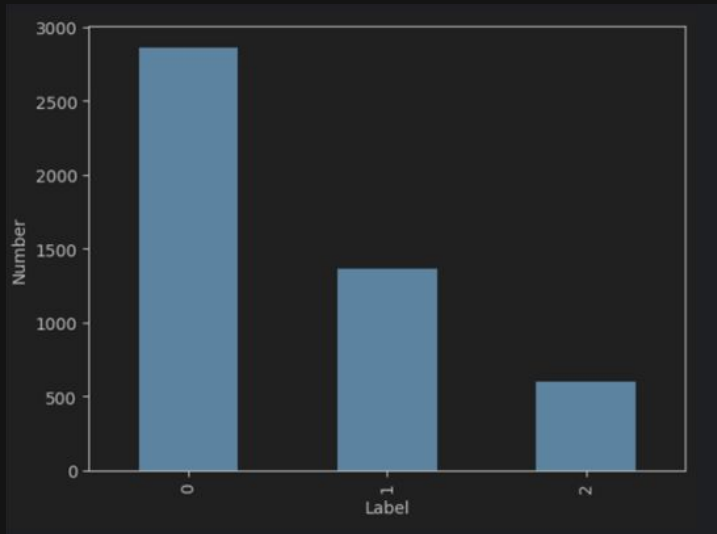
Korean-BERT - Pre-training & Fine-Tuning

Pre-training (사전 학습)



MLM, NSP 방식 사용

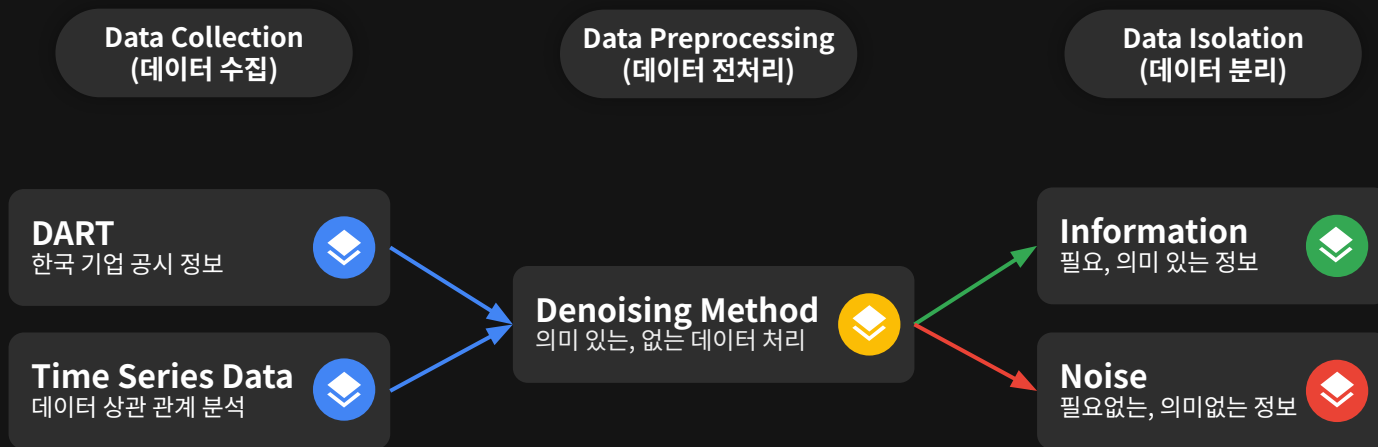
Fine-tuning (Supervised Learning - 지도학습)



0: Positive(긍정)
 1: Netural(중립)
 2: Negative(부정)

Financial Data

금융 데이터 처리 방법

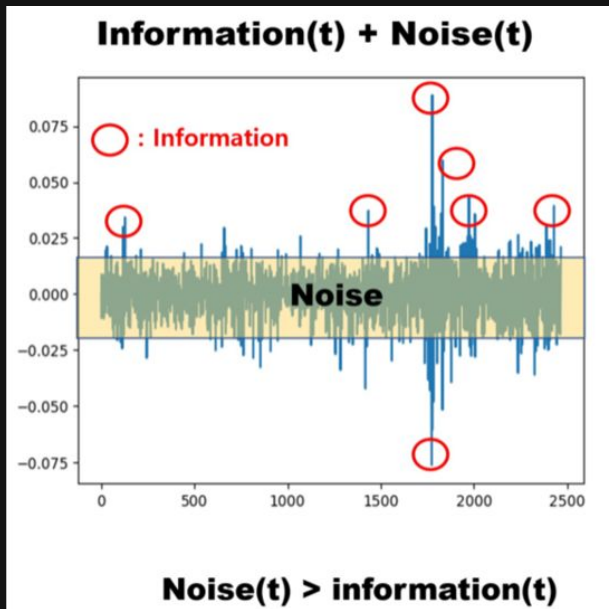


Denoising Method

금융 데이터 전처리

Denoising

Reason for Denoising
(노이즈 제거 사유)



무의미한 가격 정보 제거 - 주식 폭등, 폭락

Eliminating Pointless Price Information
ex) Stock Soars, Slams



의미 없는 정보 제거 - 분석과 무관한 뉴스

Remove meaningless information
ex) news unrelated to analysis

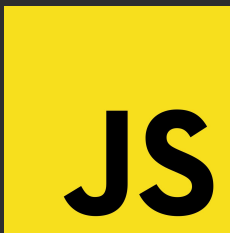
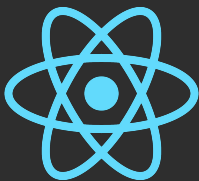


정확, 의미 있는 정보를 알기 위함
Accurate & Meaning Information

Data Implementation

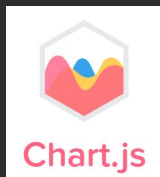
데이터 표현 방법

Web Page (FE)



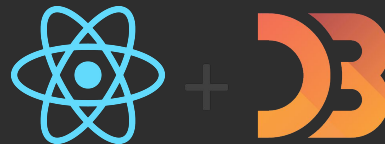
React & JavaScript

Visualizing Data



JSON, D3, Chart.js

Graph Implement



React + D3, (Dynamic)
Node.Express (BE)

Data Implementation

데이터 표현 방법

Web Page (FE)

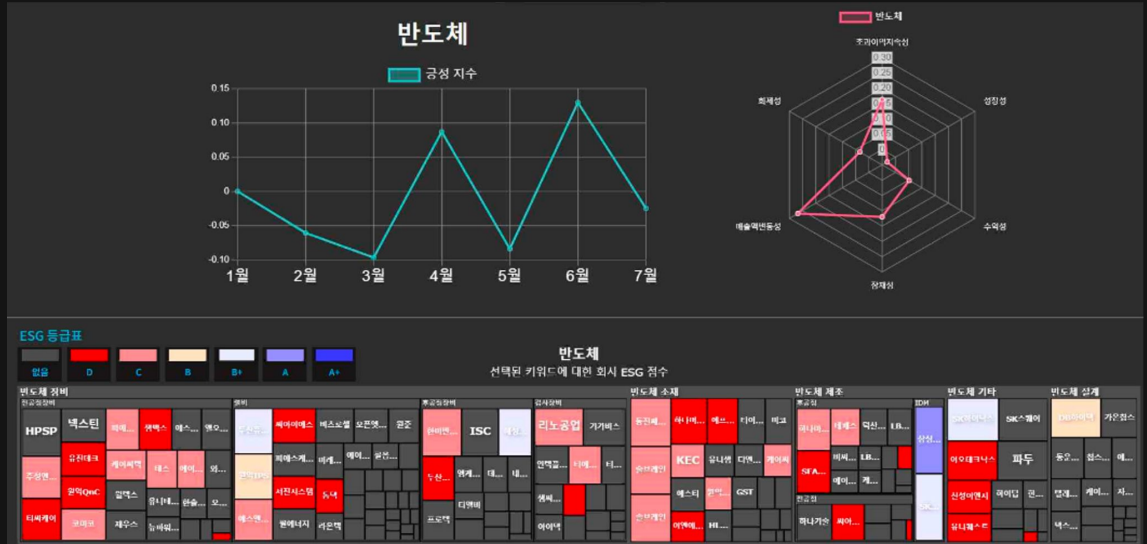
Mapping 101

쉬운 투자 정보 한눈에 보기는?
Mapping 101

Category

- 경기 관련 소재 (Consumer goods, metals, technology)
- 금융 (Finance)
- 산업재 (Industrial goods)
- 소재 (Material)
- IT (IT, mobile technology)
- 커뮤니케이션서비스 (Communication service)
- 물류 소재 (Logistics and transport goods)
- 에너지 (Energy)
- 유틸리티 (Utility)

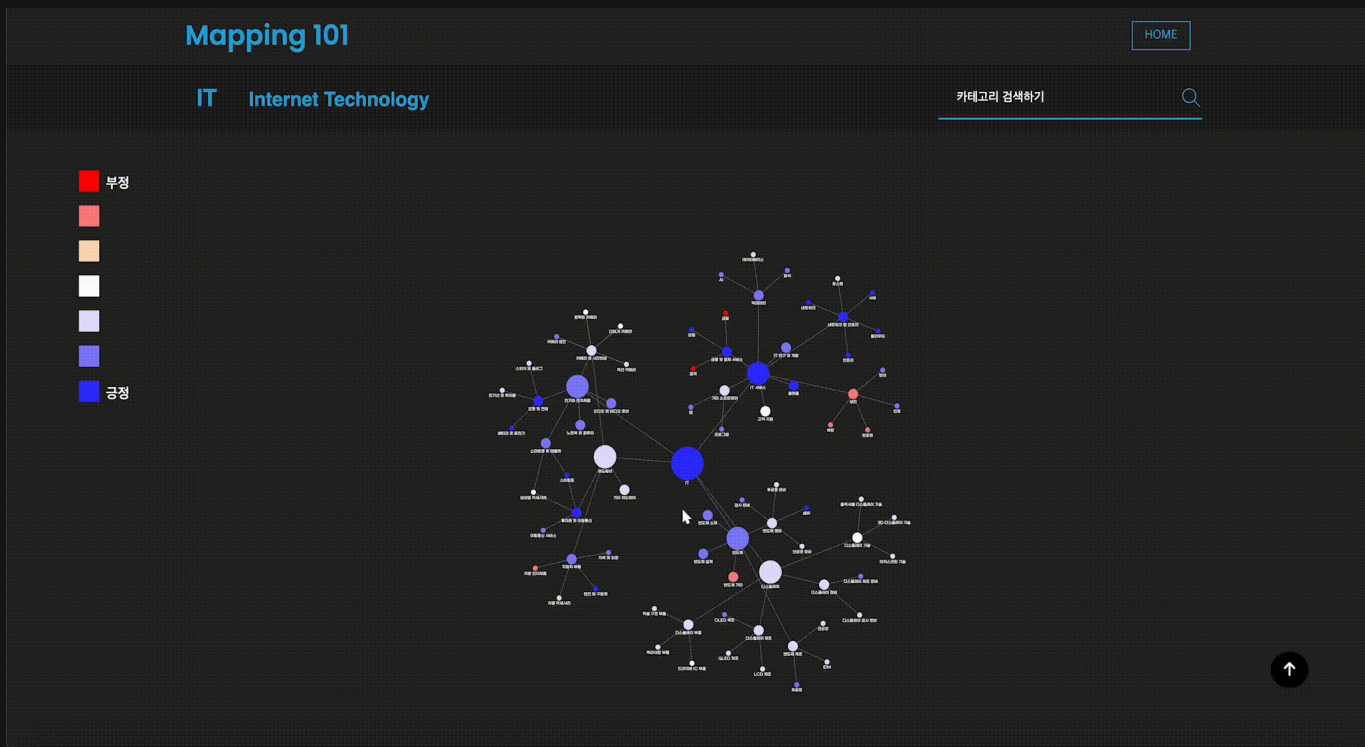
Visualizing Data



Data Implementation

데이터 표현 방법

Visualizing Data



PART 3

Development Results & Conclusion

개발 결과 & 결론



Dev Result & Effects

개발 결과 & 효과 예상



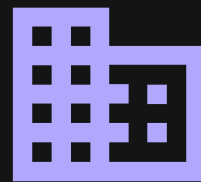
월별 양 지수

Visualization



재무지표, ESG

Corporate Analysis



타 업종

Comparison

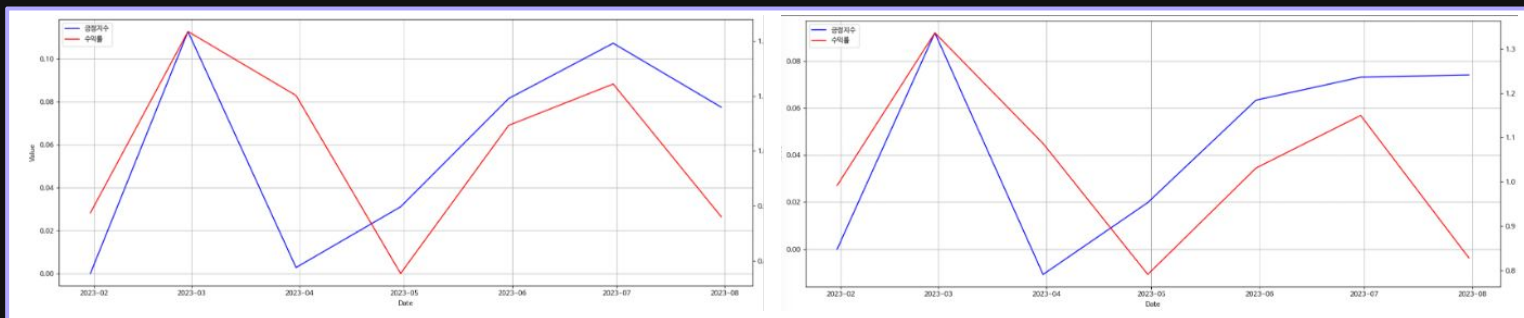


키워드 &
금융정보

Dimensional Analysis

Positive Index Reliability

긍정 지수 신뢰도



1. 실제 주가의 **노이즈 제거**, 긍정 지수 & 수익률 **상관관계 분석**
noise cancellation of Stock Prices positive index & yield correlation analysis
2. 시장의 긍정반응과 부정반응이 더 높은, 더 낮은 수익률로 연결된다는 것
positive and negative market reactions lead to higher and lower returns
3. **긍정 지수**가 급격히 상승하거나 하락할 때 **수익률이 따라오는 경향**
Yields tend to follow when the positive index rises or falls sharply

Conclusion & Discussion

결론 & 의논점



성장가능성 과
평가 정보를 제공.

Provides **growth potential** and
evaluation information



최신 트렌드 및
시장 동향을 분석

analyzing the **latest trends**
and **market trends**.



데이터 수집 제한,
시각화에 미흡

limited data collection,
lack of Visualization

Reference

관련 레퍼런스

- 강장구, 권경윤, and 심명화. "개인투자자의 투자심리와 주식수익률." *재무관리연구*3 (2013): 35-68.
- 김유신, 김남규, and 정승렬. "뉴스와 주가: 빅데이터 감성분석을 통한 지능형 투자의 사결정모형." *지능정보연구, 한국지능정보시스템학회*2 (2012).
- 김영민, 정석재, and 이석준. "소셜 미디어 감성분석을 통한 주가 등락 예측에 관한 연구." *Entrue Journal of Information Technology* 13.3 (2014): 59-69.
- 고재창, 조근태, and 조윤희. "키워드 네트워크 분석을 통해 살펴본 기술경영의 최근 연구동향." *지능정보연구*2 (2013): 101-123.
- 조수지, 김흥규, and 양철원. "기업 재무분석을 위한 한국어 감성사전 구축." *한국증권 학회지*2 (2021): 135-170.
- Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
- 안성원, and 조성배. "뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측." *한국정 보과학회 학술발표논문집*1C (2010): 364-369.
- 성태웅, et al. "기업정보 기반 지능형 밸류체인 네트워크 시스템에 관한 연구." *지능정 보연구*3 (2018): 67-88.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)** (pp. 5998-6008).
- Salton, G., & Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. In *Information Processing & Management*, 24(5), 513-523.
- Cavnar, W. B., & Trenkle, J. M. (1994). N-Gram-Based Text Categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval* (pp. 161-175).
- Odakata, T. (2012). Example of Cutting Reference Units in an N-Gram Language Model. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)* (pp. 1235-1244).

 Thank You! 



Daehyun Kim



Daehyun-Bigbread



developer._.toby

김대현 (Daehyun Kim)

Ai Engineer & Researcher, Student