성균관대학교 통계분석학회 P-SAT 제 1호 발행일 | 2022. 11. 04 발 행 | 데이터마이닝 팀

데마신문

데이터에서 반짝이는 인사이트를 얻는 그날까지 ☆야망☆

사용자 맞춤 뉴스 추천 시스템

"모두 환영합니다~" 깃헙까지 자료를 보러온 당신은 이미 만점☆

김현우 | 김준서 | 김수빈 | 변석주 | 서희나

이 내용이 흥미로운 당신… 신입 학회원 지원에 망설이지 마세요

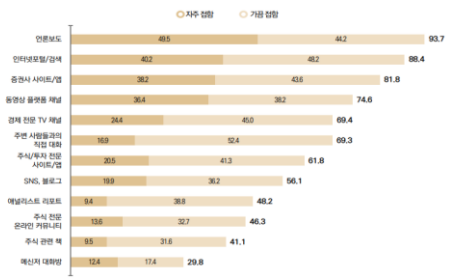
주제 선정

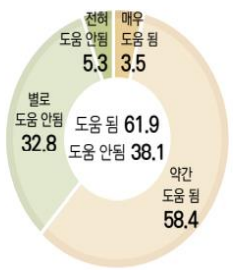
01 주제 선정

미국 연방준비제도의 기준 금리 인상

주식 시장의 혼란 가중

주식 투자자들의 인식 및 행동 조사





투자자들은 언론보도·인터넷 보도를 통해 유용한 투자 정보를 수집

주식 시장에서 전체적으로 혼조세가 나타나 빠르고 정확한 투자 정보가 더욱 중요해짐!

주제 선정

투자자들은 언론보도나 인터넷 보도를 통해

본인에게 유용한 투자 정보를 수집

투자자에게 유용할 만한 기업 관련 뉴스를

추천하는 시스템을 구현하자!

분석에 이용될 기업 선정

업종별 상위10개 기업 데이터 수집

KRX 정보데이터시스템

KRX 한국거래소 KOREA EXCHANGE

- 한국거래소의 기본 통계자료 활용
- 10월 23일 기준 KOSPI에 상장된 기업 대상
- 전 종목 거래량, 시가 총액 수집

KIND(기업공시채널)

대한민국 대표 기업공시채널 KIND

- 한국거래소의 전자 공시 홈페이지
- 10월 23일 기준 KOSPI에 상장된 법인 목록 수집

기업 리스트

코스피 상장 기업 중 업종별 상위 10개 기업 채택

종목코드	법인명	상장일	시가총액	거래량	업종명
432320	KB스타리츠	2022	2940억	209140	서비스업
403550	쏘카	2022	1669억	9245	서비스업
126720	수산인더스트리	2022	3082억	72572	건설업
⋮	⋮	⋮	⋮	⋮	⋮
000700	유수홀딩스	1956	3774억	1572625	서비스업

9개의 기업 업종

- KRX 정보데이터시스템과 KIND에서 수집한 데이터를 **종목코드** 기준으로 병합 후,
  병합된 데이터의 9개의 기업 업종을 10개로 세분화
- 각 업종 별 시가총액 상위 10개 기업을 선정하여 최종 분석 대상 기업으로 선정

데이터 수집

업종별 상위10개 기업 관련 정형 & 비정형 데이터 수집

정형-전자공시시스템 API

DART

- OpenDartReader 모듈 활용
- 연간 재무제표 손익계산서 (2019-2021)
- 사업보고서 – 직원 정보 크롤링

비정형-빅카인즈

BIGKinds
NEWS BIGDATA & ANALYSIS

- 뉴스 키워드와 관련된 인물, 장소, 조직의 관계망 분석을 제공
- 9월 21일~ 10월 29일 기업 기사 데이터 수집

뉴스 본문 데이터

뉴스 본문 URL을 통한 뉴스 본문 수집

중앙일보 한국경제
MT 머니투데이 朝鮮日報
매일경제 서울신문

10개 언론사 홈페이지의 웹구조를 각각 확인

BeautifulSoup

뷰티플 수프를 통해 HTML을 파싱 했을 때
정상적으로 데이터가 로드 되는 언론사

중앙일보, 매일경제 등 8개의 언론사

BeautifulSoup
+
Selenium

웹사이트가 동적으로 구성되어
정상적으로 HTML이 파싱되지 않거나
홈페이지에서 크롤링을 허용하지 않는 경우

조선일보,한국경제

뉴스 본문 데이터

뉴스 본문 URL을 통한 뉴스 본문 수집

수집한 뉴스 본문 데이터 中

- 뉴스 사설
- 단순 주가 등락만 알려주는 증권 기사
- 단편적인 내용 여러가지로 구성된 긴 기사

→ 본문이 담고 있는 정보가 적다 판단하여 제거

15000여개의 기사를 이용해 분석 진행

중앙일보 한국경제 MT 머니투데이 朝鮮日報 매일경제 서울신문

Beautiful Soup

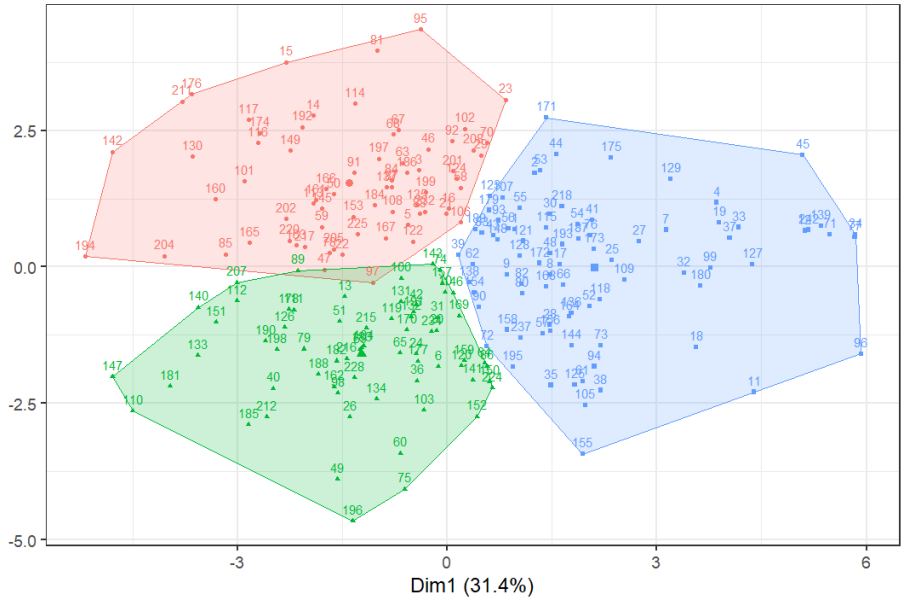
뷰티플 수프를 통해 HTML을 파싱 했을 때
정상적으로 데이터가 로드 되는 언론사

Beautiful Soup + Selenium

웹사이트가 동적으로 구성되어
정상적으로 HTML이 파싱되지 않거나
홈페이지에서 크롤링을 허용하지 않는 경우

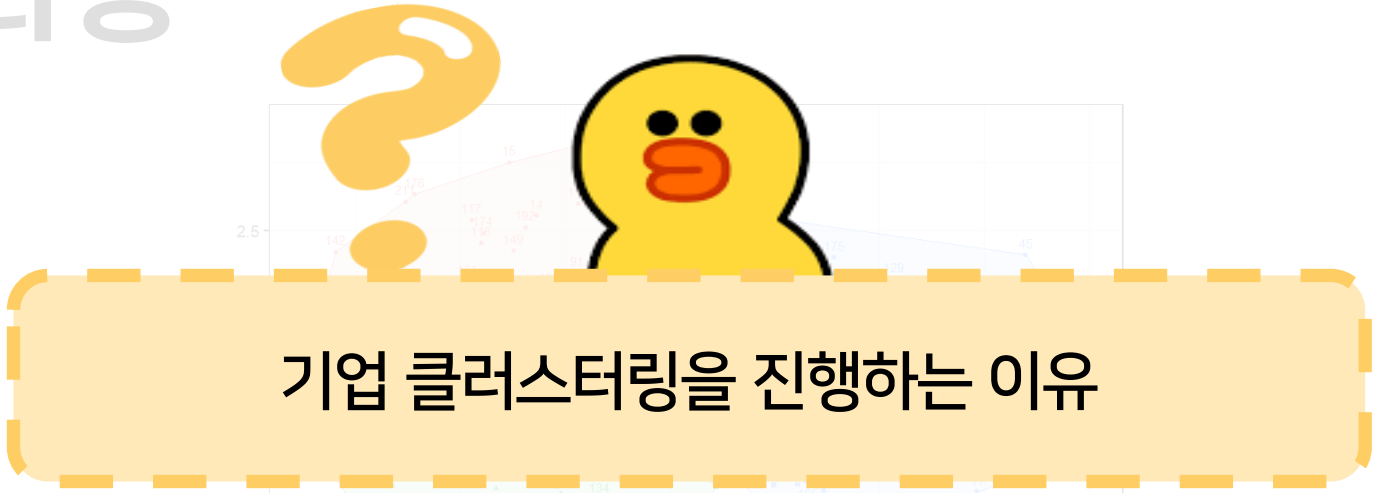
데마가 처리했으니
안심하라구
10개 언론사 홈페이지의 웹구조를 각각 확인

클러스터링



- **유사한 속성**을 갖는 데이터를 일정한 수의 군집으로 그룹핑하는 비지도 학습
  - 수집한 기업 정형 데이터를 이용해 기업 클러스터링을 진행

클러스터링



기업 클러스터링을 진행하는 이유

- 선정한 기업에 대한 개괄적인 이해
- 기업 기사 추천 시스템에서 **콜드 스타트** 문제 → 초기 설정값으로 활용하기 위함
- 유사한 속성을 갖는 데이터를 일정한 수의 군집으로 그룹핑하는 비지도 학습
- 수집한 기업 정형 데이터를 이용해 기업 클러스터링을 진행

중심점 기반 클러스터링

- 100개의 기업들을 **모두 클러스터에 할당**하는 것이 목표
- EDA 진행 결과 타 기업들에 비해 수치들이 높았던 삼성전자의 경우

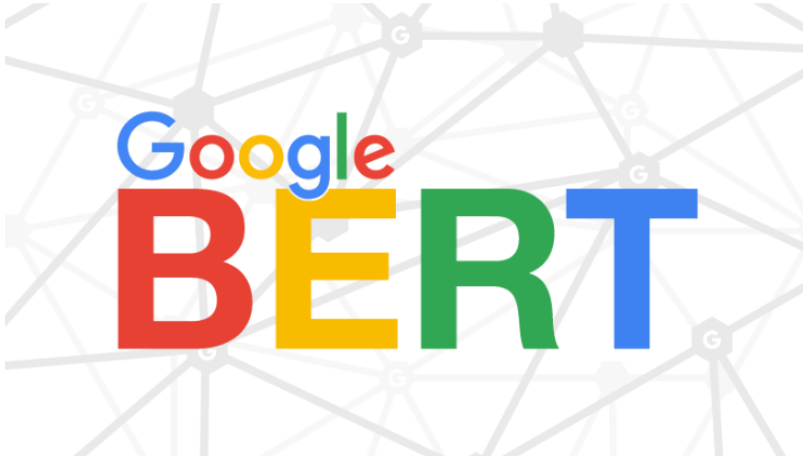
클러스터링에서 제외 이후 생성된 군집에 적절하게 할당

클러스터링 결과

클러스터1	클러스터2	클러스터3	클러스터4	클러스터5	클러스터6	클러스터7
LG화학	SK이노베이션	케이티엔지	SK 텔레콤	삼성바이오	SK 하이닉스	LG엔솔
삼성물산	고려아연	케이티	HMM	셀트리온	현대자동차	SK
현대모비스	LG 생활건강	대한항공	LG 이노텍	카카오	기아	포스코케미칼
LG전자	아모레퍼시픽	서울도시가스	현대글로벌	카카오뱅크	포스코 홀딩스	엔씨소프트
…	…		롯데케미칼	우리금융지주	한국전력공사	…
삼성화재보험	강원랜드		LG 유플러스	하이브		코웨이
LG 디스플레이	롯데쇼핑		GS리테일	넷마블		현대오토에버

딥러닝 모델 Text Summarization - SKT KoBart Model

Bert Model



Bidirectional **E**ncoder **R**epresentations from **T**ransformers

GPT Model

Generative **P**re - **T**raining

수집한 뉴스 본문의 길이가 너무 길다고 판단되어

딥러닝을 통한 뉴스 본문을 시도해보고자 함

Bert Model & GPT Model 결합하다!!

✓ Bert Model

양방향으로 언어 시퀀스의 토큰들을
Attention을 통해 모두 반영하여
문자를 Encoding 하는 모델

↓

백터화된 입력 시퀀스 상의 다른 단어를 연결

✓ GPT Model

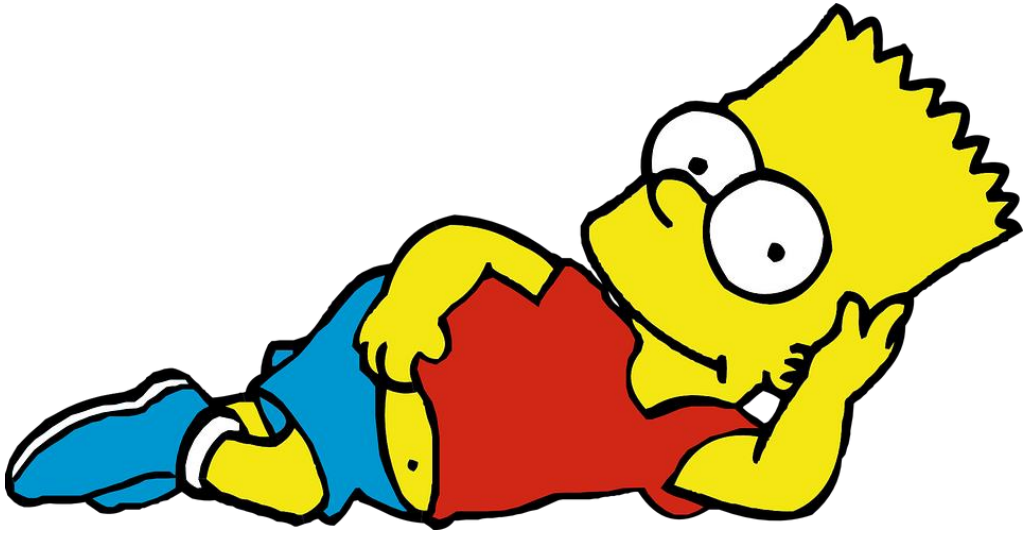
입력을 바탕으로 새로운 출력을
만들어내는 생성 모델

↓

이전까지 모델에 들어온 labeling 된
데이터를 바탕으로 새롭게 출력

Bart 모델

Bidirectional과 Auto-Regressive Transformer를 합친 모델



✓ **Noising의 유연성**

길이 변화와 같은 임의의 변형이라도

기존 텍스트에 바로 적용 가능

✓ **Text 생성 fine-tuning 효과·효율적**

Abstractive Dialogue, Question Answering,

Summarization Task에서 SOTA의 성능을 달성

SOTA: 현재 가장 성능이 뛰어난 기술을 뜻하는 의미

문서 요약 Dataset

Aihub 문서 요약 데이터 + SKT Kobart 데이터

신문기사 뉴스 텍스트 요약 데이터셋 40만 데이터

```
{"id": "신문 ID",
"abstractive": "요약 모델",
 "extractive": [0, 1, 2],
 "article_original": [원문],
 "media": "신문사명"}
```

→

News	Summary
article_original [1]	Abstractive[1]
article_original [2]	Abstractive[2]
…	…
article_original [n]	Abstractive[n]

✓ Json 형식의 데이터

뉴스 원문과 요약 데이터를 Json 형태로 저장해서 제공 → tsv 형식으로 변환 필요

✓ Preprocess

Json 형식을 파악하고 Python을 통해서 tsv로 저장

→ Train / Test / Valid set 으로 저장

Kobart Fine Tuning

✓ **하이퍼 파라미터 지정**

- Batch size를 10으로 지정 → 28950/epoch 학습
- Epoch 을 10으로 지정, 학습률 3e-5로 지정

✓ **Train & Validation 진행**

- Epoch 당 2시간 52분 정도 학습 진행
- val_loss → 최적 ckp와, epoch 5, 10의 ckp 저장

최종 Kobart Model

- **긴 문장 데이터** Kobart 모델 간 요약 성능이 비슷 / **짧은 데이터**에 있어서는 **성능 차이**가 발생
- Kobart 모델은 **기존의 평가지표로는 최적의 모델 선택 어려움** → Heuristic한 평가 진행
- epoch = 6 (val_loss = 1.183) 모델을 최종 모델로 선정

```
import torch
from transformers import PreTrainedTokenizerFast
from transformers.models.bart import BartForConditionalGeneration

model = BartForConditionalGeneration.from_pretrained('./junseo_kobart/kobart_summary_epoch_06')
tokenizer = PreTrainedTokenizerFast.from_pretrained('gogamza/kobart-base-v1')
```

최종 Kobart Model

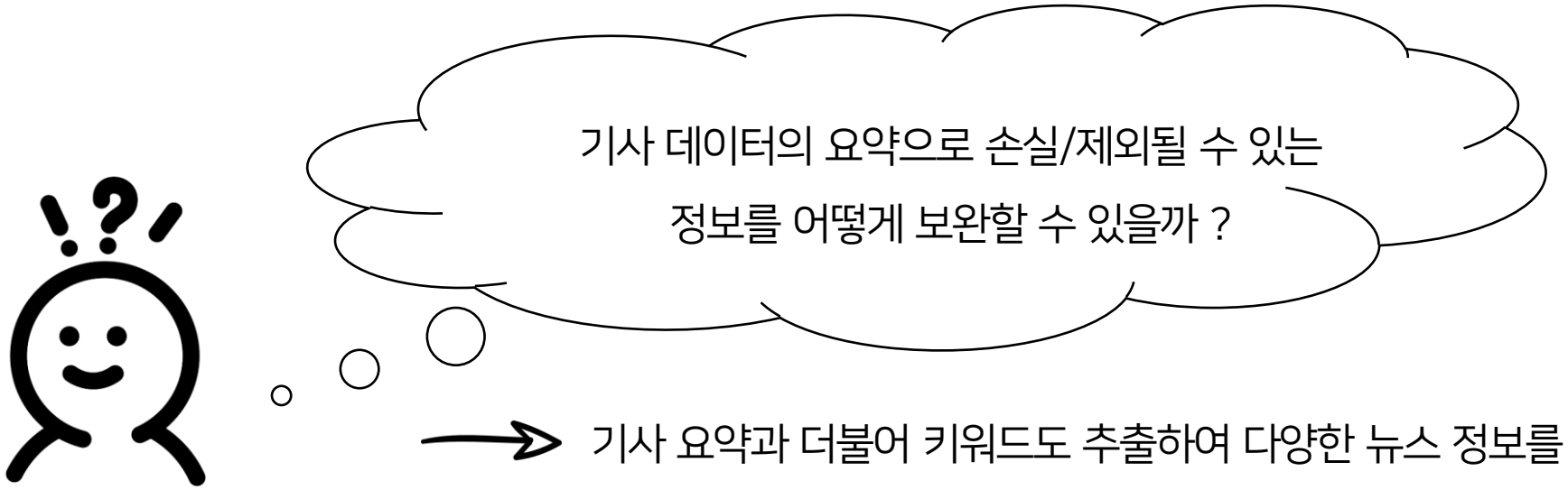
원문 데이터

카카오뱅크는 지난해 1년 동안 1조 649억원의 영업수익과 2569억 원의 영업이익을 거뒀는데, 올해는 3분기 만에 누적 영업수익 1조1211억원, 영업이익 2674억원을 기록했다. 9개월 만에 지난해 12개 월 동안의 성과를 넘어선 셈이다. 뱅킹 부문 역시 견고한 성장을 보이고 있다. 수신 잔액은 지난해 말 약 30조 원에서 9월 기준 약 34조6000억원으로 늘었으며, 저원가성 예금이 꾸준히 확대돼 62.1%의 비중을 기록했다. 여신 잔액은 같은 기간 약 25조9000억원에서 27조5000억원 수준으로 증가했다. 중저신용자 대출과 전월세보증금·주택담보대출이 성장을 견인한 것으로 분석된다.
또한 무보증 중저신용자 대상 대출 잔액은 3조288억원으로 전년 말 2조4643억원 대비 22.9% 증가했다. 중저신용대출 잔액 비중 또한 전년 말 17%에서 23.2%까지 6%포인트(p) 넘게 상승했다. 카카오뱅크는 지난 2017년 7월 출범 이후 약 5년 만에 9월 말 기준 1978만 명의 고객을 확보한 인터넷은행으로 성장했다. 지난해 말 1799만명, 올해 상반기 1917만명 이후 3개월 만에 60여만명의 고객이 추가로 유입되면서 카카오뱅크는 연내 2000만명 고객을 돌파할 것으로 전망하고 있다. 카카오뱅크가 모두의 은행으로 자리잡으면서 고객들의 나이도 다양해지고 있다. 지난해 3분기 ▲10대 7% ▲20대 27% ▲30대 27% ▲40대 23% ▲50대 이상 16% 였던 연령별 고객 비중이 올해 3분기에는 ▲10대 8% ▲20대 24% ▲30대 25% ▲40대 23% ▲50대 이상 19%로 변화했다. 청소년 가입자 확대는 10대 전용 금융 서비스인 '카카오뱅크 미니(mini)'가 기인했다는 게 카카오뱅크 측의 설명이다. 올 3분기 기준 mini의 누적 가입 고객 수는 약 150만명으로 지난해 3분기 100만명 가량보다 약 50% 늘었다. '카카오뱅크 mini'는 만 14~18세 청소년을 고객으로 한 결제, 송금, 충전 등이 가능한 선불전자지급수단 서비스다. '티머니 충전·조회' 기능 등을 탑재할 예정이다.중장년층 고객들의 유입도 계속되고 있다. 올해 신규 카카오뱅크 가입자 중 절반인 50%가 40대 이상이다. 신용정보를 조회하고 또 올릴 수 있는 '내 신용정보'서비스를 이용한 50대 이상 고객 수는 126만명에 이르며, '휴면예금·보험금 찾기'를 써본 고객 수도 72만명이다.
카카오뱅크는 올 4분기에도 주택담보대출 상품 확대와 개인사업자 뱅킹 출시, 인증 사업 진출 등에 주력할 계획이다. 우선 지난 2월 출시한 주택담보대출은 지속적인 대상 및 지역 확대로 누적 약정금액 8070억원을 기록했다. 카카오뱅크는 취급 지역을 수도권 및 5대 광역시에서 전국으로 넓히고, 대상 주택 금액의 한도를 없애는 등의 고객 접점을 늘리는 노력을 해왔다고 설명했다. 이에 지난 9월 처음으로 월 취급액 1500억 원을 돌파하기도 했다.
카카오뱅크는 지난 1일에는 개인사업자 뱅킹 서비스도 출시했다. 카카오뱅크 측은 단순 개인사업자 대출 상품뿐 아니라 수신 상품(통장)과 지급결제(카드)까지 망라한 풀뱅킹 서비스를 제공하며 개인사업자들에 인기를 끌 것으로 기대하고 있다. 세금 관리와 신용 관리, 매출 관리와 관련된 다양한 개인사업자 서비스를 추가할 예정이다.
인증 사업 역시 예정돼 있다. 지난 10월 방송통신위원회로부터 '본인확인기관'으로 지정됐으며, 연내 '공인전자문서중계자'와 '전자서명인증사업자' 라이선스까지 취득한다는 계획이다. 본인확인기관을 포함해 세 가지 라이선스를 모두 확보하면, 고객들은 카카오뱅크 애플리케이션(앱)을 통해 행정안전부, 국세청과 같은 정부기관 사이트에 로그인할 수 있다. 또 공문서를 신청하고, 신원확인이 필요한 서비스를 이용할 수 있다. 또한 지방세 고지서를 카카오뱅크 앱에서 받아보는 것도 가능하다.

DATA1

카카오뱅크는 지난해 1년 동안 1조 649억원의 영업수익과 2569억 원의 영업이익을 거뒀는데,
올해는 3분기 만에 누적 영업수익 1조1211억원, 영업이익 2674억원을 기록하여
9개월 만에 지난해 12개 월 동안의 성과를 넘어서고 뱅킹 부문 역시 견고한 성장을 보이고 있다.

키워드 추출



→ 기사 요약과 더불어 키워드도 추출하여 다양한 뉴스 정보를 제공하자!

- 문장 내 단어와 단어의 관계를 기반으로 작동하는 WORDRANK를 이용해 키워드를 추출하기로 함
  - WORDRANK는 구글의 PAGERANK를 단어에 적용한 알고리즘

PageRank & WordRank

✓ PageRank

PageRank는 웹에서

문서와 문서 간에 연결관계

$$P_i = \frac{1-d}{n} + d\sum_{k=1}^{n} L_{ij}\frac{P_j}{m_j}$$

✓ WordRank

WordRank는

단어와 단어의 연결관계

$$PR(u) = c * \sum_{v \in B_u} \frac{PR(v)}{N_v} + (1-c) * \frac{1}{N}$$

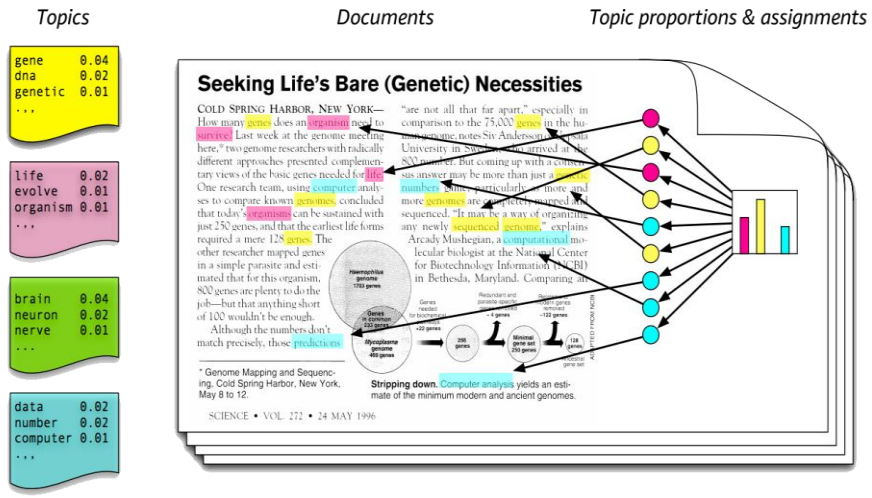
키워드 추출 결과

index	keywords
1	['데이터센터', '화재', 'SK']
2	['대북 제재', '북한']
3	['매출', '전년', '대비', '방산', '증가', '성장', '동기', '실적', '사업', '연구원']
4	['리츠', '고배', '하락']
5	['영업이익', '전망치', '실적', '하향', '3분기']
6	['필수소비재', '최근']
⋮	⋮

단어와 단어의 연결관계를 고려한 WORDRANK를 이용한 키워드 추출이 잘 되었음을 확인 가능

LDA Latent Dirichlet Allocation

Topic Modeling의 기법 중 하나



- 토픽 별 단어의 분포
- 문서 별 토픽의 분포

→ 모두 추정

디리클레 분포

- 베타 분포가 확장된 것
- 0과 1 사이의 값을 가지는 다변수 확률 변수의 베이지안 모형에 사용됨

$$Dir(x_1, x_2, \cdots, x_k;\ \alpha_1, \alpha_2, \cdots \alpha_k) = \frac{1}{B(\alpha_1, \alpha_2, \cdots \alpha_k)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

$$B(\alpha_1, \alpha_2, \cdots \alpha_k) = \frac{\prod_i^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)}$$

LDA Latent Dirichlet Allocation

LDA의 목표

- W(단어)를 관측 → Z(토픽)값을 정함
- 디리클레 분포를 업데이트
- 가능한 모든 경우 Z값 중 **가장 가능도 높은 Z값**

→ 문헌 내 각각의 단어들이 어디에 배정되는지 추론 가능

전체 과정을 베이즈 정리로 요약

$$P(Z|W) = \frac{P(W|Z)P(Z)}{P(W)}$$ 일 때

P(Z|W)를 최대로 하는 Z값을 찾는 것이 목표

- 문헌 내 모든 단어 개수 N개 → W,Z 모두 N차 벡터
- N값 ↑ → 계산 기하급수적으로 **복잡**해짐

→ Gibbs Sampling을 통한 빠르고 쉬운 연산

Gibbs Sampling

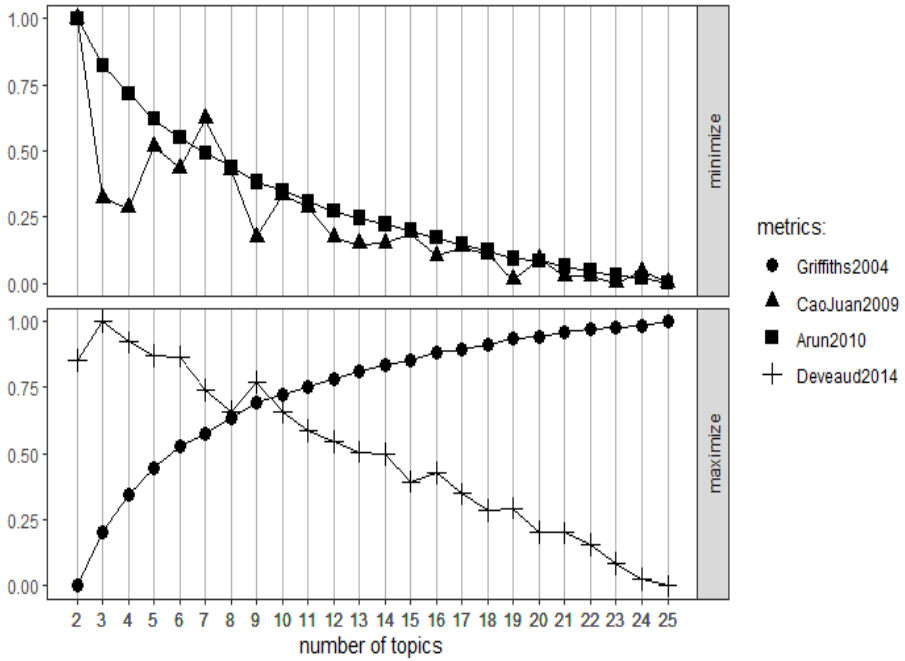
N차의 자료를 1차 자료 N개가 모인 것으로 가정

→ 나머지 N-1개를 고정

→ 한 차원에 대해서 자료 샘플링

→ N개의 차원에 대해 자료를 샘플링 후 합침

LDAtuning



✓ LDAtuning

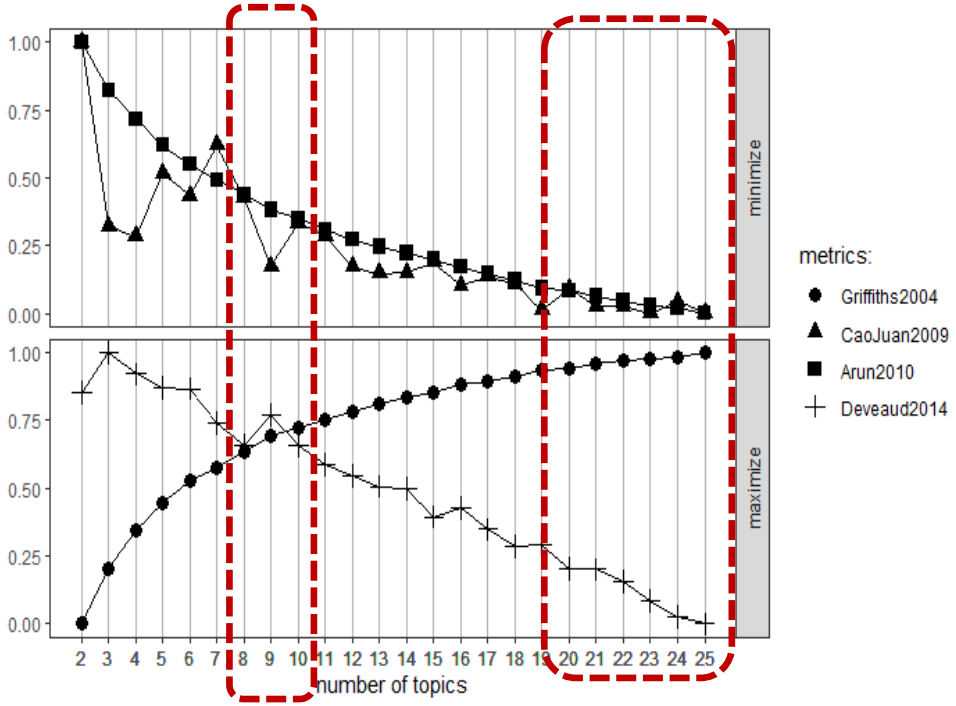
- LDA모델의 최적의 주제 수 K를 찾아주는 R 라이브러리
- Perplexity나 Coherence 이외의 LDA 관련 논문에서 제안 된 방법들로 하이퍼파라미터 튜닝 가능

✓ 최적의 주제 수 K 하이퍼파라미터 튜닝

- 주제 수 K의 범위는 충분히 해석 가능한 주제 수인 2~25개 사이로 설정
- LDA 모델의 반복수를 1000, Burn-in을 500으로 설정 후 진행

LDAtuning

최적의 주제 수 K 튜닝 결과



✓ 튜닝 결과 해석

- Deveaud2014는 계속 작아지는 추세를 보이다 K=9에서 다시 증가 → 감소하는 경향을 보임
- Deveaud2014를 제외한 나머지 지표는 토픽 개수가 커짐에 따라 최대화/최소화 되는 경향을 보임

✓ 최종 LDA 모델 주제 수 K 선정

- 값들이 적절하게 최적화되는 K=9와 K=25 중 선정
- 유일하게 최대화가 되지 않는 Deveaud2014 지표를 고려하고, 토픽에 대한 해석까지 고려하여 최종 주제 수를 K=9로 선정

추천시스템

✓ 콘텐츠 기반 추천

- 해당 **컨텐츠에 대한 정보**만을 이용해 추천 실시
- 사용자의 **과거 소비 컨텐츠 특성 분석**

→ **유사한 특성**을 지닌 컨텐츠를 사용자에게 추천

협업 필터링의 경우 시스템 초기에

사용자-아이템 관계에서 정보가 부족한

콜드스타트 문제가 발생 가능

프로젝트 특성상 콜드스타트 문제에서 벗어날 수 없음

→ **컨텐츠 기반 추천 시스템**을 build

추천시스템 설계 초기값 설정

기업	토픽 1	…	토픽 9
삼성전자	0.089	…	0.109
LG에너지솔루션	0.086	…	0.101
SK하이닉스	0.088	…	0.096
삼성바이오로직스	0.085	…	0.095
삼성SDI	0.085	…	0.101
LG화학	0.081	…	0.1101
⋮	⋮	⋮	⋮

선호 기업 선정

사용자	토픽 1	…	토픽 9
User 1	0.087	…	0.102

앞서 구한 기업 별 평균 주제 가중치를 바탕으로 사용자가 선호하는 기업을 선정

→ 이 값들을 평균 내어 **사용자 주제 선호 가중치의 초기값**으로 사용

추천시스템 설계 초기값 설정 및 Test용 기사 추출

기사	토픽 1	…	토픽 9	기업
News 1	0.049	…	0.058	삼성전자, LG
News 2	0.053	…	0.091	KT, LG유플러스
News 3	0.461	…	0.064	NAVER, 카카오
News 4	0.207	…	0.101	롯데렌탈
News 5	0.122	…	0.079	F&F
News 6	0.053	…	0.514	삼성전자, SK
⋮	⋮	⋮	⋮	⋮

기업	토픽 1	…	토픽 9
삼성전자	0.089	…	0.109
LG에너지솔루션	0.086	…	0.101
SK하이닉스	0.088	…	0.096
삼성바이오로직스	0.085	…	0.095
삼성SDI	0.085	…	0.101
LG화학	0.081	…	0.1101
⋮	⋮	⋮	⋮

- 기사 별 주제 가중치를 통해 **기업 별 평균 주제 가중치**를 얻어냄
- 해당 기업 관련 뉴스들을 추출하여 **주제 가중치를 평균** 낸 후 사용

추천시스템 설계 최종데이터셋 소개

사용자	토픽 1	…	토픽 9
User 1	0.087	…	0.102
User 2	0.156	…	0.093
⋮	⋮	⋮	⋮

✓ **사용자의 선호 주제 가중치 행렬**

- 회원가입 과정에서 사용자가 선택한 선호 기업을 바탕으로 생성
- 이후 사용자가 읽은 뉴스에 따라 사용자 선호 주제 가중치를 업데이트

기사	토픽 1	…	토픽 9
News 1	0.049	…	0.058
News 2	0.053	…	0.091
News 3	0.461	…	0.064
⋮	⋮	⋮	⋮

✓ **기사 별 주제 가중치 행렬**

- LDA를 통해 얻어낸 기사 별 주제 가중치
- Test를 위해 7개 클러스터에서 각 3개 기업을 추출,
  기업당 5개의 뉴스를 추출하여 총 105개의 기사로 Test 진행

추천 시스템을 위한 유사도 계산

✓ 코사인 유사도

두 벡터의 유사도를 측정하는
가장 기본적인 방법

$$cos\theta = \frac{A \cdot B}{|A||B|}$$

(두 벡터의 내적을 각 벡터의 길이로 normalize)

✓ KL-Divergence

두 확률분포의
정보 엔트로피 차이를 계산하여
확률 분포간 차이를 측정하는 방법

주제 가중치는 LDA를 통해 추론된
다항 분포의 파라미터이므로 사용 가능

추천시스템 설계 콘텐츠 기반 추천 결과

선호기업 '카카오', 'SK'로 설정 시 추천 결과

코사인 유사도 추천 결과

기사	관련 기업	제목
News 1	카카오	카카오 먹통에 '화재 주의보' 떨어진…
News 2	F&F	수급난 속 연기금의 약한 매수세…
News 3	네이버	네이버와 카카오의 차이는 데이터 …
News 4	F&F	신기사 전환' F&F파트너스, 'F&F …
News 5	롯데렌탈	롯데렌터카, 10월 가을맞이 '단기…

Jensen-Shannon Divergence 추천 결과

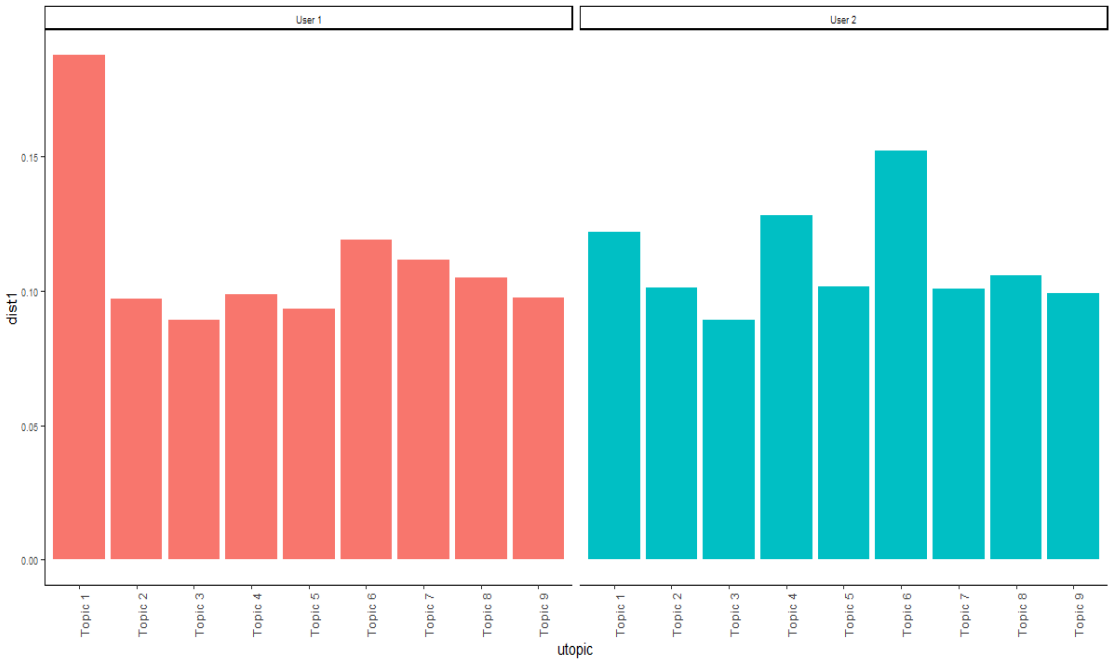
기사	관련 기업	제목
News 1	F&F	수급난 속 연기금의 약한 매수세…
News 2	카카오	카카오 먹통에 '화재 주의보' 떨어진…
News 3	F&F	신기사 전환' F&F파트너스, 'F&F …
News 4	롯데렌탈	롯데렌터카, 10월 가을맞이 '단기…
News 5	네이버	네이버와 카카오의 차이는 데이터 …

코사인 유사도와 Jensen-Shannon Divergence를 사용한 추천 알고리즘이

사용자의 선호 주제 가중치와 비슷한 기사를 잘 추천하지 못하는 문제 발생

왜 추천이 되지 않는가?

① 초기값 설정의 문제



✓ **Flat한 사용자 선호 주제 가중치**

앞서 설명한 방식으로 사용자 선호 주제 가중치를 구했을 때

평균을 내는 방식으로 구하게 되면서 전체적으로

사용자 주제 가중치가 Flat한 형태를 띄게 되었음

✓ **Flat한 분포가 가지는 문제점**

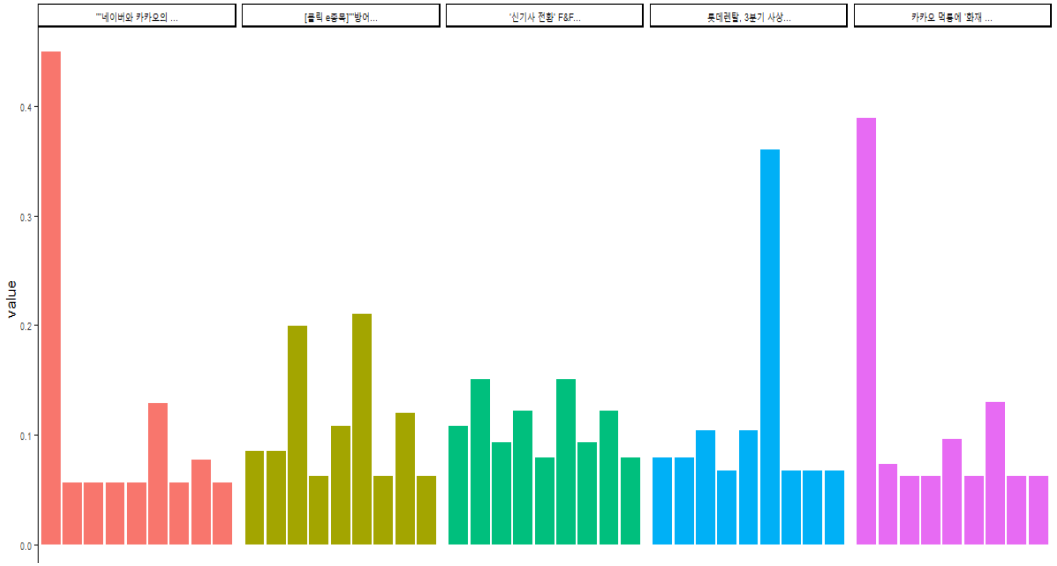
베이지안 입장에서 보았을 때,

분포가 Flat해질수록 **분포에서 정보가 사라지게 됨**

→분포가 Non-Informative해지면서 정교한 추천이 어려워짐

왜 추천이 되지 않는가?

② 극단적인 형태를 가지는 기사 별 주제 가중치



✓ **극단적인 기사 별 주제 가중치**

반면 선호 기업과 관련 있는 뉴스의 경우에는
주제 가중치가 극단적으로 나타나는 것을 확인

✓ **사용자 주제 가중치와 성격이 다른 기사 별 주제 가중치**

- 사용자의 주제 선호 가중치는 Flat / 기사의 주제 가중치는 극단적
- 기존의 코사인 유사도나 Jensen-Shannon Divergence를 사용

→비교적 Flat한 가중치를 갖는 기사를 사용자에게 추천하게 됨

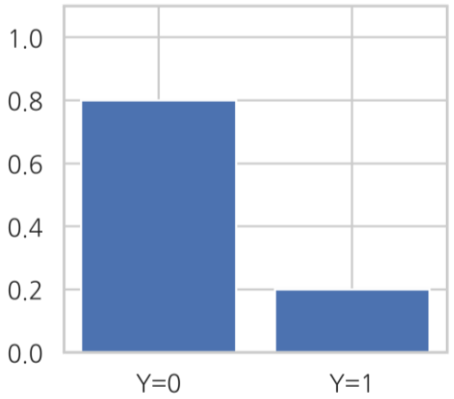
왜 추천이 되지 않는가?

이를 해결하기 위해선...

① **초기값 설정** 시 사용자 선호 주제 분포가 Flat한 형태를 띄지 않게 하도록 **가중치 부여**

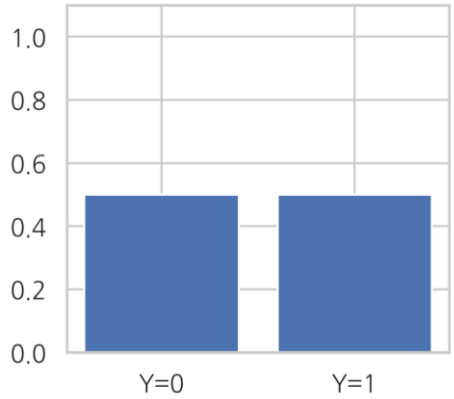
② **주제 가중치가 극단적인 기사에 높은 가중치**를 부여하여 정교한 추천을 할 수 있게 함

Attention이 왜 더 잘 작동하는가?



A=[0.8, 0.2]

$|A| = 0.8246 \ldots$



B=[0.5, 0.5]

$|B| = 0.7071 \ldots$

확률 분포를 어떤 벡터라고 가정하고 길이를 구해보면

분포가 좀 더 극단적인 경우에 벡터의 길이가 더 긴 것을 확인할 수 있음

→ 분포가 좀더 극단적인 것에 대한 Measure로 사용 가능!

왜 Attention이 더 잘 작동하는가?

✓ 코사인 유사도

두 벡터의 유사도를 측정하는
가장 기본적인 방법

$$cos\theta = \frac{A \cdot B}{|A||B|}$$

(두 벡터의 내적을 각 벡터의 길이로 normalize)

✓ Attention Score

Query와 Key의 내적을 통해
두 벡터의 유사도를 구함

$$A \cdot \mathrm{B} = |A||B|cos\theta$$

(각 벡터의 길이와 두 벡터의 사잇각으로 표현)

왜 Attention이 더 잘 작동하는가?

Attention Score를 구할 때

내적을 통해 유사도를 계산하므로

벡터의 길이에도 영향을 받게 됨

↓

극단적인 분포를 띄는 경우에

좀 더 높은 가중치를 주어 유사도 측정 가능!

✓ Attention Score

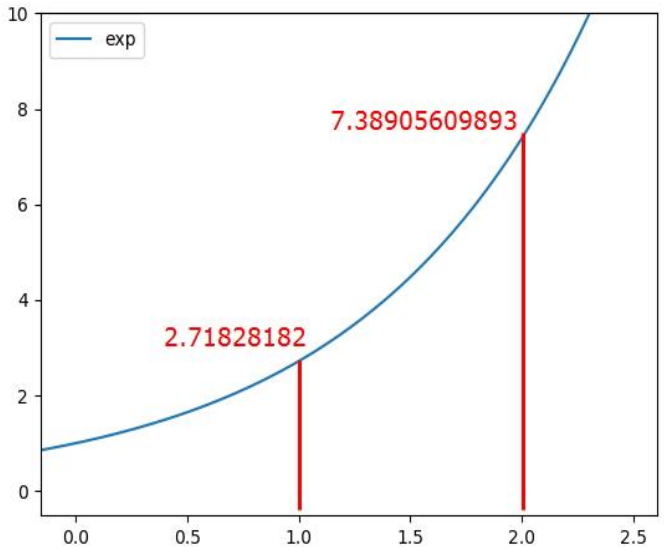
Query와 Key의 내적을 통해

두 벡터의 유사도를 구함

$$A \cdot \mathrm{B} = |A||B|cos\theta$$

(각 벡터의 길이와 두 벡터의 사잇각으로 표현)

Attention이 왜 더 잘 작동하는가?



$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

✓ **Softmax Function**

- Attention Distribution을 구할 때 사용
- $y = e^x$는 지수함수

→ 가중치로 **단순 비중을 사용할 때 보다 더 강하게 가중치**를 줄 수 있음

✓ **Attention의 의미**

- Attention Score와 Attention Distribution을 구할 때, 벡터에서 좀더 **극단적인(특징적인) 값**들을 추출하는 것
- 이 과정에서 극단적인 분포에 좀 더 높은 가중치를 주게 되어 데이터로부터 특징들을 효과적으로 추출할 수 있음

추천시스템 설계 Attention Distribution 기반 추천 결과

선호 기업 '카카오', 'SK' 설정 시 추천 결과

기사	관련 기업	제목
News 1	카카오	카카오 먹통에 '화재 주의보' 떨어진…
News 2	네이버	네이버와 카카오의 차이는 데이터 …
News 3	카카오	12시간 먹통된 카톡 다시 수면위로…
News 4	카카오	판교 데이터센터 화재로 카카오 …
News 5	카카오	카카오, 사상초유 '먹통에 공매도…

→Topic 3와 관련

선호 기업 '삼성바이오로직스', '셀트리온' 설정 시 추천 결과

기사	관련 기업	제목
News 1	LG	韓, 바이든 IRA로 장기적으로는 …
News 2	LG	전기차 생태계 빨아들이는 美 IRA…
News 3	삼성바이오	진격의 거인'된 삼바, 압도적 생산…
News 4	롯데렌탈	롯데렌터카, 10월 가을맞이 '단기…
News 5	네이버	식약처, 삼바 제조 모더나 코로나 …

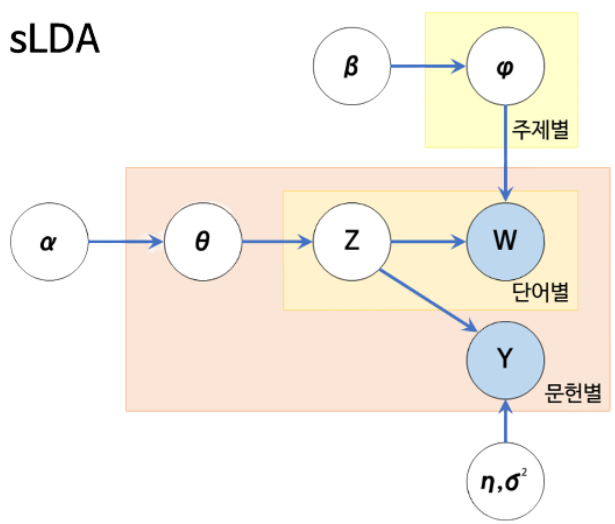
→Topic 4와 관련

Attention Distribution 기반 추천 결과

코사인 유사도와 Jensen-Shannon Divergence를 사용한 이전 알고리즘보다 훨씬 더 나은 성능을 보임을 확인

추후 시스템 보완점

③ S-LDA를 활용한 더 많은 서비스 제공



✓ S-LDA(Supervised LDA)

- 기존 LDA의 구조에서 응답변수 Y를 추가
- 기존 LDA에서 다음과 같은 가정이 추가

$$Y \sim N(\eta \cdot E(Z), \sigma^2)$$

✓ S-LDA를 활용한 서비스 런칭

사용자로부터 뉴스에 대한 평점을 수집한다면

- 평점 정보를 추가적으로 활용한 좀 더 세부적인 토픽모델링
- 새로운 기사에 대한 평점 예측을 통한 감성분석
- 예측 평점을 초기값으로 한 행렬 분해 기반 협업 필터링을 통한 뉴스 추천 정교화

→ 좀 더 풍부한 서비스를 제공할 수 있을 것으로 기대