



***** 서울맛집공유
***** 플랫폼 서비스

미안하다 4랑한다

김채원, 명재정, 윤형석, 전동준, 허정윤

팀 구성원 및 역할



FRONT-END

- 네이버 식당 크롤링
- UI / UX 디자인



BACK-END

- 망고플레이트 식당 크롤링
- DB 설계 및 구축



DATA-SCIENCE

- 다이닝코드 식당 크롤링
- 모델링



DATA-ENGINEER

- 카카오 리뷰 크롤링
- 리뷰 SPARK로 전처리
- AIRFLOW 구현



DATA-ENGINEER

- 서울 식당 API 크롤링 및 전처리
- ELASTIC 구현
- AWS 환경구축

목차

먹갈암



1 프로젝트 기획 배경

1-1. 주제선정 이유

1-2. 시장조사

2 데이터

2-1. 데이터 수집

2-2. 데이터 저장

2-3. 데이터 전처리

2-4. ERD

3 데이터 분석

3-1. NLP흐름

3-2. Sample Data

3-3. Total Data

3-4. Theme 선정

4 서비스 시연

5 마무리

5-1. 한계점 및 기대효과

5-2. 참고자료





1. 프로젝트 기획배경

1-1. 주제선정 이유

1-2. 시장조사





1-1 주제 선정 이유

연남동 |

#연남동맛집
게시물 1,389,519

#연남동카페
게시물 1,360,126

#연남동
게시물 3,814,698

#성수
#성수동맛집
게시물 576,965

#성수동카페
게시물 998,728

seongsu_official
성수동 길라잡이

#성수맛집
게시물 372,363

서울 맛집
서울 맛집 / 강남, 잠실, 대학로, 홍대, 성수 · hye...

성 수
성수짬뽕박사 / 성수 성수동 성수역 맛집

#연남동맛집
게시물 1,389,520

팔로우

인기 게시물



1-2 시장 조사

먹잘알



사이트 이름	리뷰 데이터	키워드별 검색	마이페이지	리뷰 작성 조건	테마별 음식점 검색	지도로 음식점 검색	날씨를 이용한 음식 추천
망고플레이트	0	X	X	X	X	X	X
식신	0	0	X	X	0	0	X
다이닝	0	0	△	0	X	X	X
먹잘알	0	0	0	0	0	0	0



1-2 시장 조사_ 날씨별 음식점 조회

날씨	메뉴
비	파전, 빈대떡, 삼계탕, 볶음밥, 중식
눈	설렁탕, 도가니탕, 삼겹살, 레스토랑, 호박죽
맑음	레스토랑, 설렁탕, 비빔밥, 삼겹살, 치킨
구름많음	막걸리, 맥주, 베이커리, 아이스크림, 도넛
흐림	초밥, 딤섬, 쌀국수, 삼겹살, 곱창

조찬열, 정구임, 서양민, 최혜림(2017),
 「날씨 및 요일 특성이 음식점 메뉴 검색시스템 이용에 미치는 영향에 관한 실증 연구」,
 『스마트미디어저널』 제 6권 제 2호, 50p.-56p.



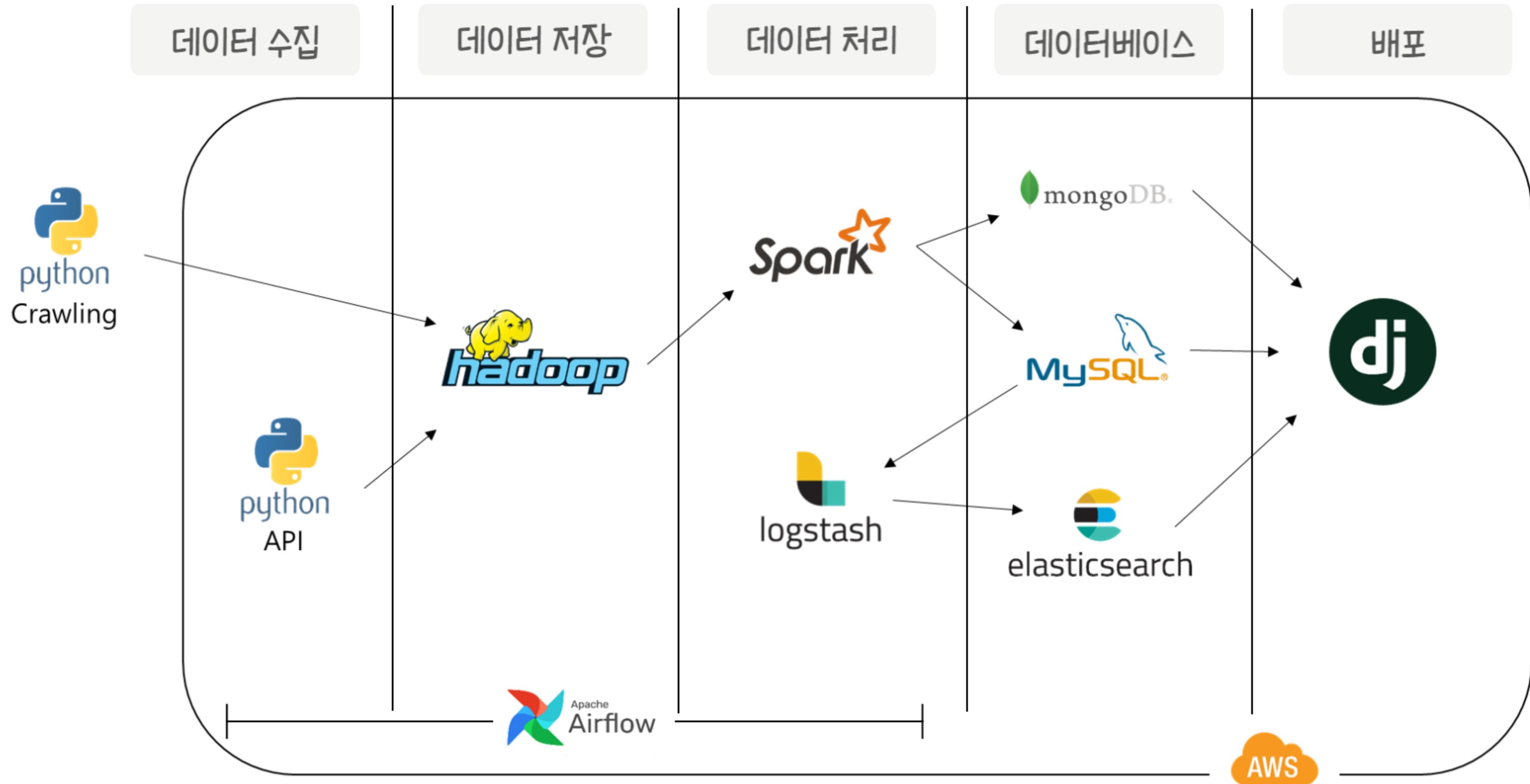
2. 데이터

- 2-1. 데이터 수집
- 2-2. 데이터 저장
- 2-3. 데이터 전처리
- 2-4. ERD





2 데이터_시스템 아키텍쳐





2-1 데이터 수집_데이터 명세서

데이터 분류	출처	자료유형	자료명	행	열	비고
API	서울 열린 데이터 광장	JSON	api_seoul0.json api_seoul1.json api_seoul2.json api_seoul3.json api_seoul4.json	473879	44	서울음식점
	공공데이터포털	CSV	weather.csv	6	5	기상청 초단기예보
	Kakao Developers	JSON		1	12	주소에 대한 좌표
	Kakaomaps API	Object				지도
Crawling	카카오맵	JSON	kakao_total.json	124500	3	
	네이버맵	JSON	naver_1000.json ... naver_51550.json	26121	9	
	다이닝코드	JSON	seoul0-1000.json ... seoul124000-end.json	124500	9	서울음식점
	망고플레이트	JSON	mango_seoul_total.json	74328	9	

데이터 수집



수집



API

서울 음식점 인허가
날씨



Crawling

카카오맵 음식점 리뷰
네이버 음식점 상세정보 및 리뷰
다이닝코드 음식점 상세정보 및 리뷰
망고플레이트 음식점 상세정보 및 리뷰

결과물

API

서울 음식점.json
날씨.csv



Crawling

카카오맵 음식점.json
네이버 음식점.json
다이닝코드 음식점.json
망고플레이트 음식점.json



하둡저장

2-2 데이터 저장

먹갈암



저장 데이터

처리

DB저장

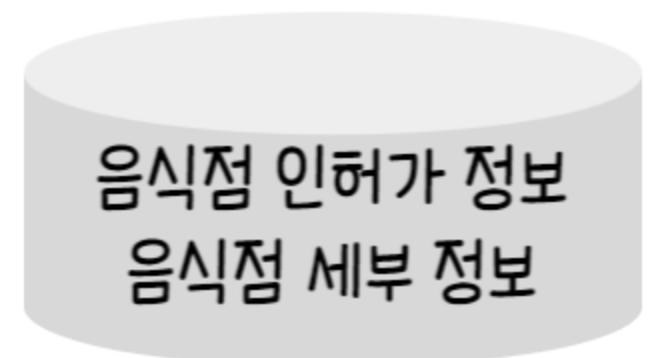


Spark

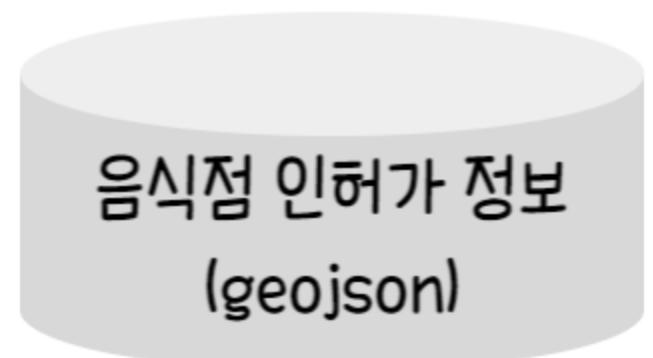
이상치 제거
결측치 제거
음식점 위치 좌표 추가
리뷰 특수문자 제거
리뷰 영어 제거
.....



MySQL®



mongoDB®



2-3 데이터 전처리

머글님
♥♥



DB 데이터

처리

저장



logstash



elasticsearch



2-3 데이터 전처리

배포

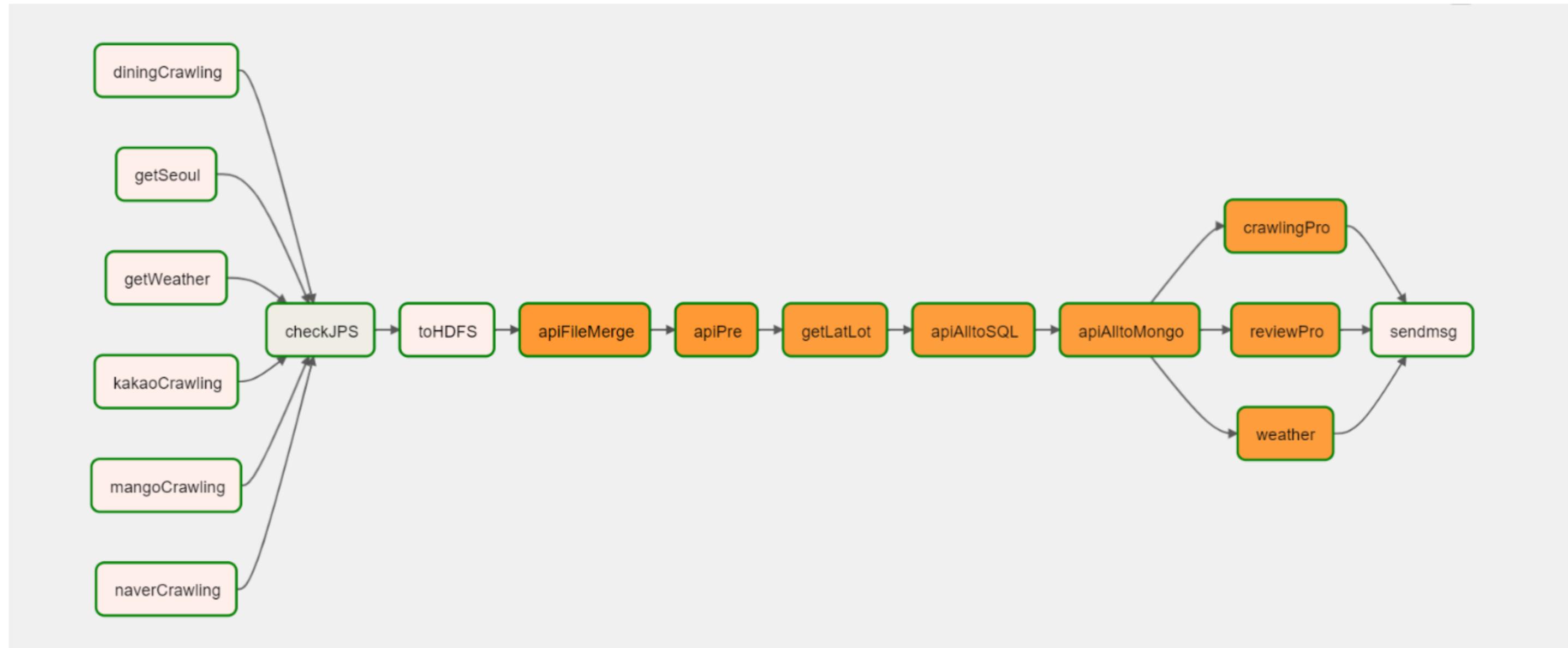


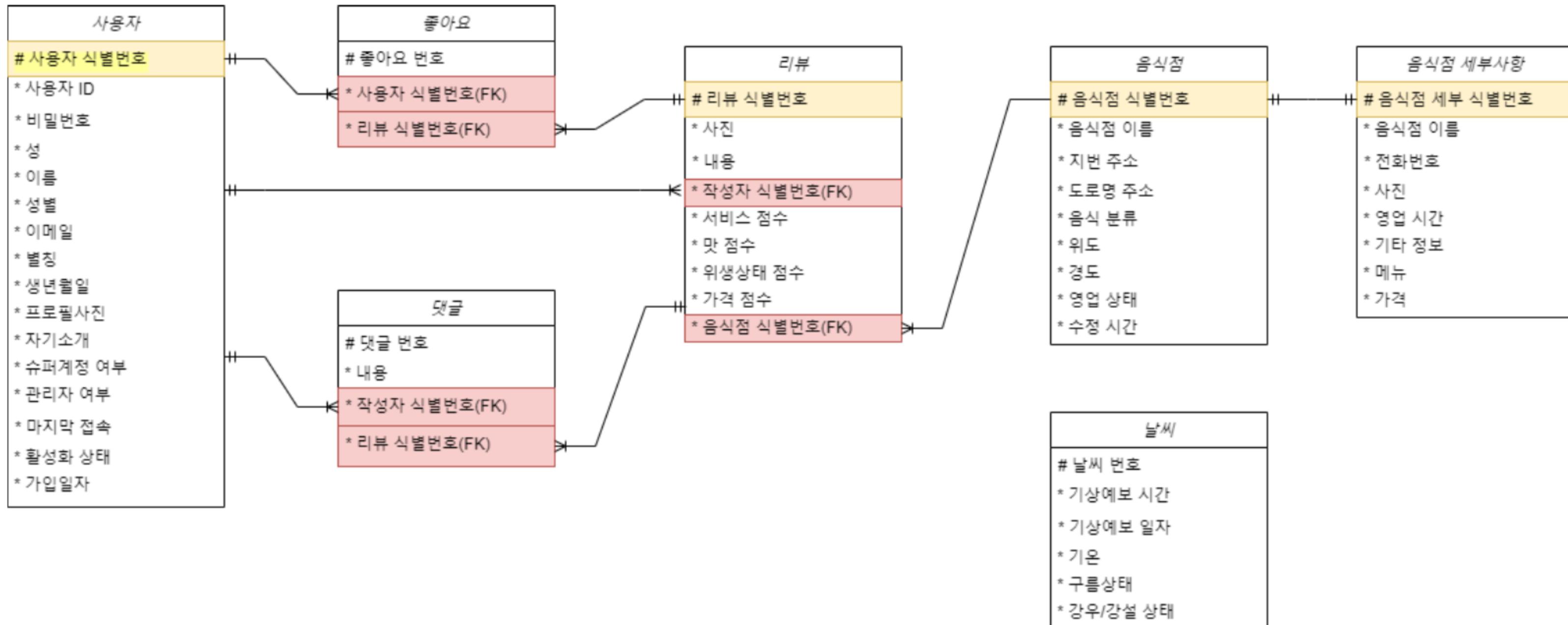
django





2-3 데이터 전처리_Airflow







3. 데이터 분석

- 3-1. NLP흐름
- 3-2. Sample Data
- 3-3. Total Data
- 3-4. Theme 설정



모델링 요소

LDA
토픽 모델링

Word2Vec

토픽 시각화



3-1 NLP 흐름





3-1

NLP 흐름

총 1086

총 1086개

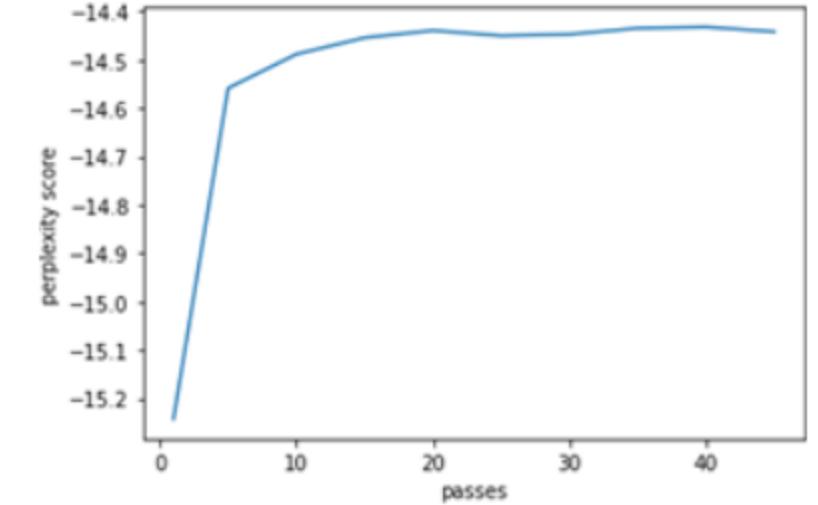
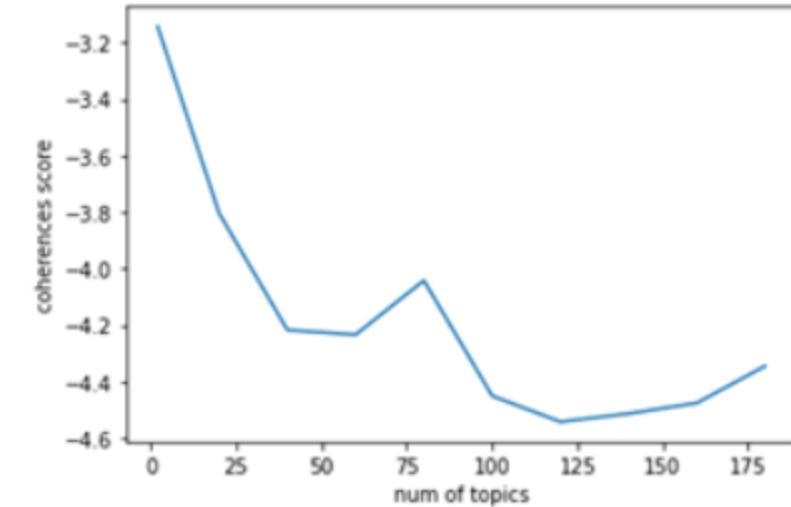


3-2 Sample Data_ 카카오맵 리뷰 1만개

1. Tokenize

```
[0,
 '0.020*"만두" + 0.014*"칼국수" + 0.011*"분위기" + 0.010*"인증" + 0.006*"음식" + 0.005*"막걸리" + 0.005*"많이" + 0.005*"케이크" + 0.004*"얼
(1,
 '0.036*"떡볶이" + 0.015*"김밥" + 0.009*"튀김" + 0.008*"추어탕" + 0.006*"음식" + 0.005*"맵다" + 0.005*"성선여대" + 0.005*"깔끔하다" + 0.005*
(2,
 '0.015*"쭈꾸미" + 0.010*"구이" + 0.007*"굽다" + 0.007*"생선" + 0.005*"심검살" + 0.004*"음식" + 0.004*"연탄" + 0.004*"죽" + 0.004*"길음" +
(3,
 '0.013*"만두" + 0.009*"냉면" + 0.008*"비싸다" + 0.006*"육" + 0.005*"만두진골" + 0.005*"음식" + 0.005*"커피" + 0.004*"분위기" + 0.004*"惩戒
(4,
 '0.026*"치킨" + 0.011*"튀김" + 0.009*"음식" + 0.007*"피자" + 0.005*"칼국수" + 0.005*"기름" + 0.005*"마리" + 0.005*"별별" + 0.005*"기분" +
(5,
 '0.007*"빈대떡" + 0.007*"고기" + 0.007*"서비스" + 0.007*"기분" + 0.006*"비싸다" + 0.006*"김밥" + 0.006*"음식" + 0.006*"쫄국수" + 0.005*"길
(6,
 '0.018*"짬뽕" + 0.017*"탕수육" + 0.014*"음식" + 0.009*"방문" + 0.008*"짜장면" + 0.007*"분위기" + 0.006*"깔끔하다" + 0.006*"아쉽다" + 0.006
(7,
 '0.029*"고기" + 0.011*"음식" + 0.010*"갈비" + 0.009*"서비스" + 0.009*"기성" + 0.006*"굽다" + 0.005*"냉면" + 0.005*"해주다" + 0.005*"추천"
(8,
 '0.033*"초밥" + 0.015*"꼬치" + 0.011*"스시" + 0.008*"신부" + 0.007*"언어" + 0.005*"서비스" + 0.005*"밥" + 0.005*"우동" + 0.005*"분위기" +
(9,
 '0.008*"김밥" + 0.008*"분위기" + 0.007*"안암" + 0.006*"꼬치" + 0.004*"해주다" + 0.004*"맥주" + 0.004*"음식" + 0.004*"모르다" + 0.004*"방문
```

2. Tuning



3. Assign_Topic

doc_topic_df.head()				
Doc_Num	Topic	Percentage	Id	s_name
0	0	7	0.912596	0 일식동경
1	1	8	0.914813	1 상해
2	2	2	0.699926	2 미쿠
3	3	5	0.887488	4 카페디퍼
4	4	2	0.687354	12 올리브커피숍

4. Word2Vec

```
model.wv.most_similar("김치")
[('깍두기', 0.9430533051490784),
 ('육수', 0.9200084209442139),
 ('바지락', 0.9093956351280212),
 ('국수', 0.9019865989685059),
 ('멸치', 0.9006192684173584),
 ('들깨', 0.8996758460998535),
 ('만두국', 0.8949522376060486),
 ('중국산', 0.8885930180549622),
 ('칼제비', 0.8880550861358643),
 ('끓이다', 0.8810204267501831)]
```

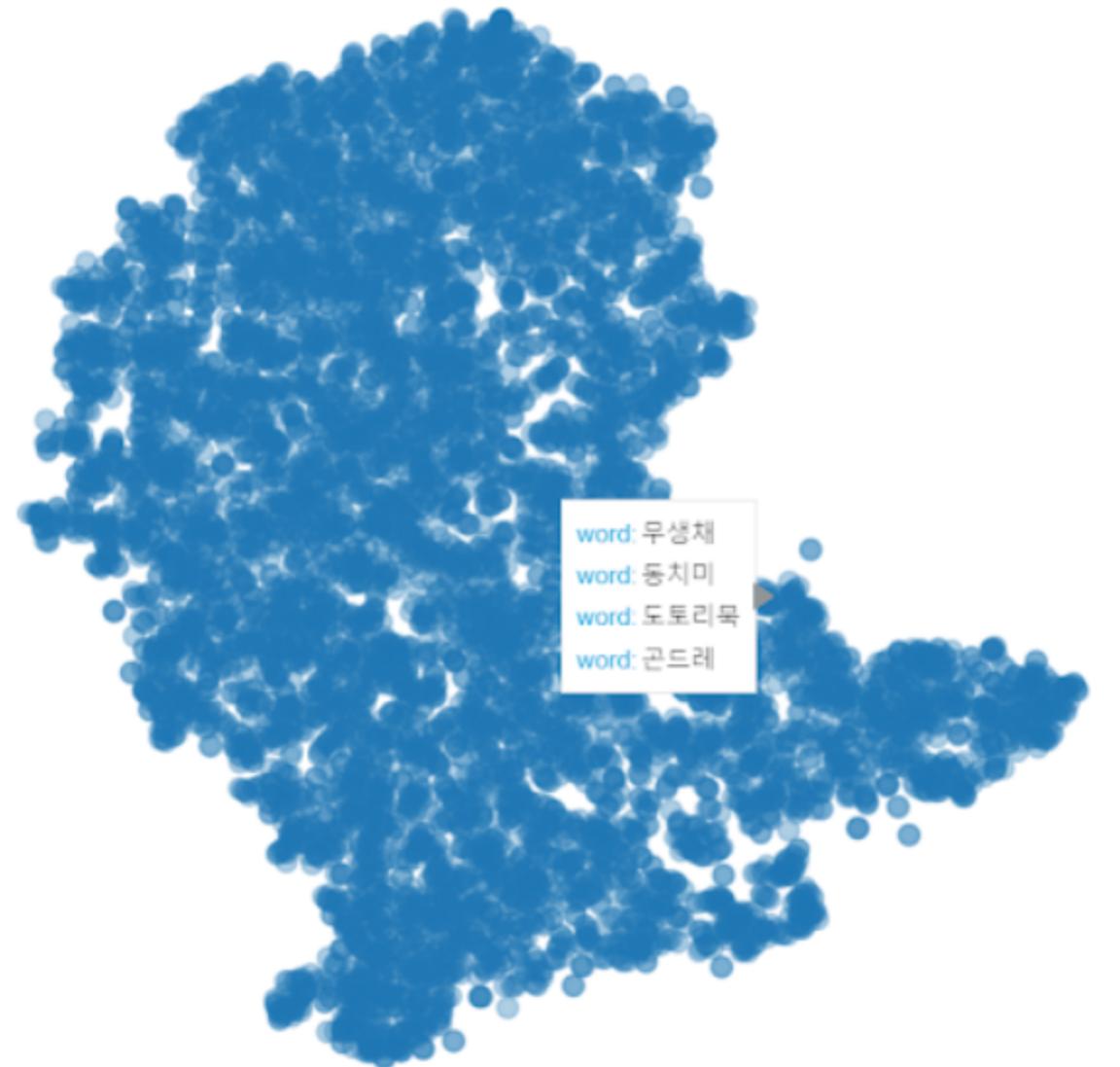
```
model.wv.most_similar("분위기")
[('즐기다', 0.9525651335716248),
 ('아늑하다', 0.9473084807395935),
 ('이쁘다', 0.945631742477417),
 ('데이트', 0.9426214694976807),
 ('편안하다', 0.942006528377533),
 ('넓다', 0.9373310208320618),
 ('와인', 0.9290252327919006),
 ('옥', 0.9253629446029663),
 ('음악', 0.9175732135772705),
 ('넉', 0.9153951406478882)]
```



3-2 Sample Data_ 카카오맵 리뷰 1만개

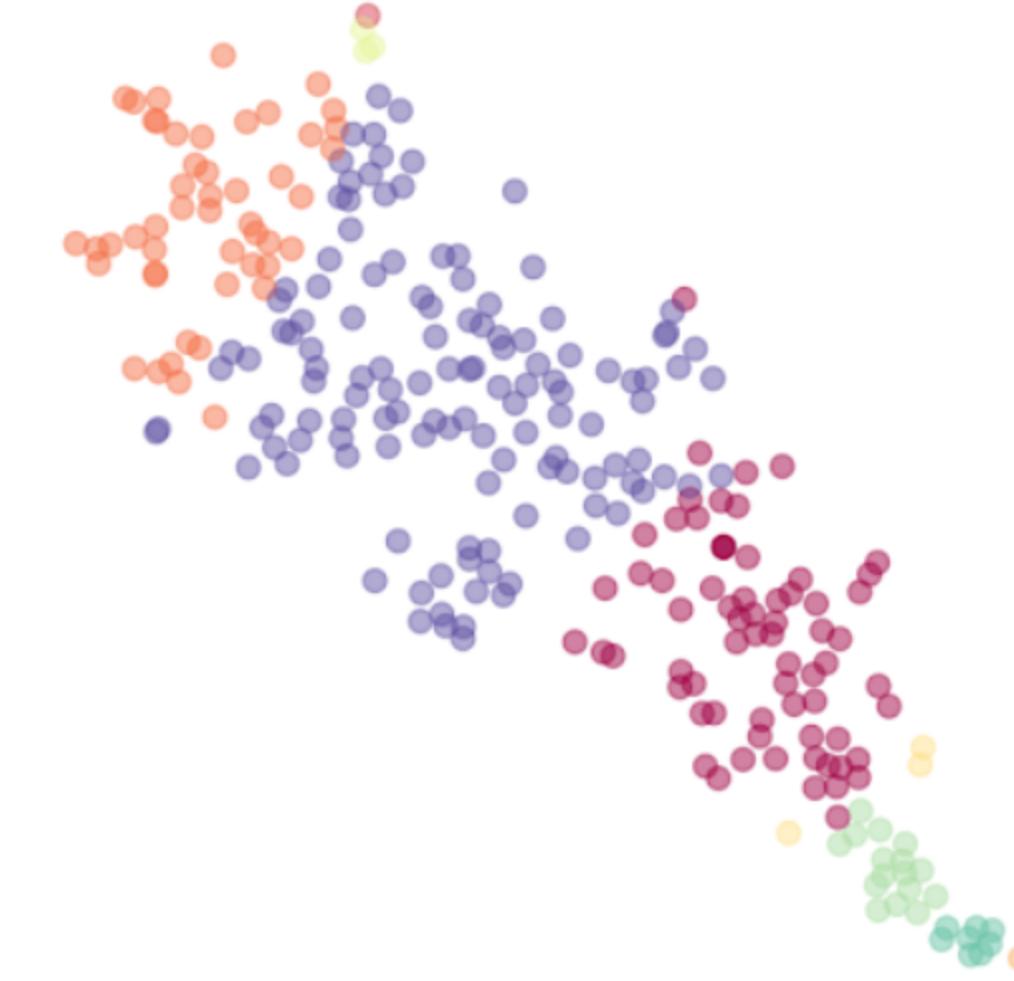
단어 분포

TSNE Embeddings



단어 클러스팅 & 음식점 분포

TSNE Embeddings





3-3 Total Data

사이트명	리뷰가 달린 음식점 개수
카카오 맵	63,951개
네이버 맵	5,717개
다이닝코드	16,760개
망고플레이트	14,417개
총 합계	음식점 124,500개 중 70,558개

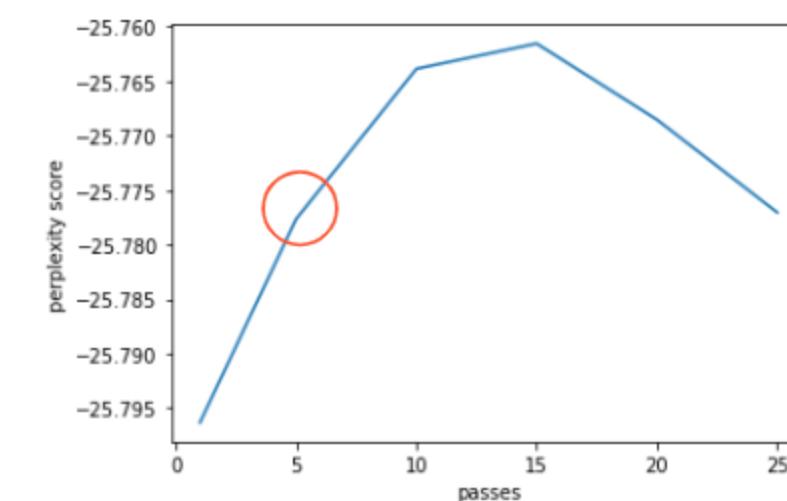


3-3 Total Data iteration:500

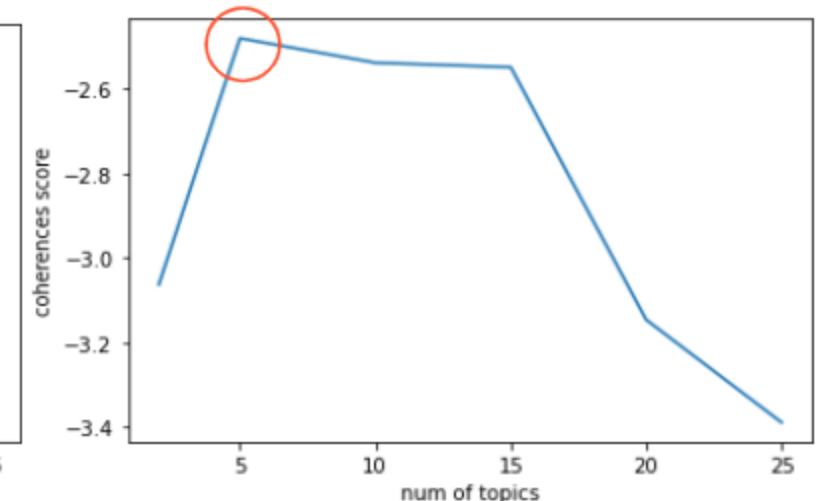
1. Tokenize

```
[0,
  '0.011*"김밥" + 0.010*"곱창" + 0.003*"종업원" + 0.003*"정신" + 0.003*"평점"
(1,
  '0.025*"떡볶이" + 0.007*"순대" + 0.007*"육회" + 0.006*"삼겹살" + 0.006*"김
(2,
  '0.047*"파스타" + 0.023*"와인" + 0.018*"스테이크" + 0.008*"칵테일" + 0.007*
(3,
  '0.024*"돈까스" + 0.016*"라면" + 0.011*"쌀국수" + 0.009*"우동" + 0.008*"카
(4,
  '0.061*"피자" + 0.032*"족발" + 0.030*"꼬치" + 0.020*"막걸리" + 0.014*"닭갈
(5,
```

2. Tuning



Perplexity-pass:5



Coherence-topic:5

3. Assign_Topic

4. Check_Group

```
Topic # 0 -----
90153
69252
27592

Topic # 1 -----
106725
983
98587

Topic # 2 -----
44357
8169
553
```

```
Topic # 3 -----
85994
98252
87681

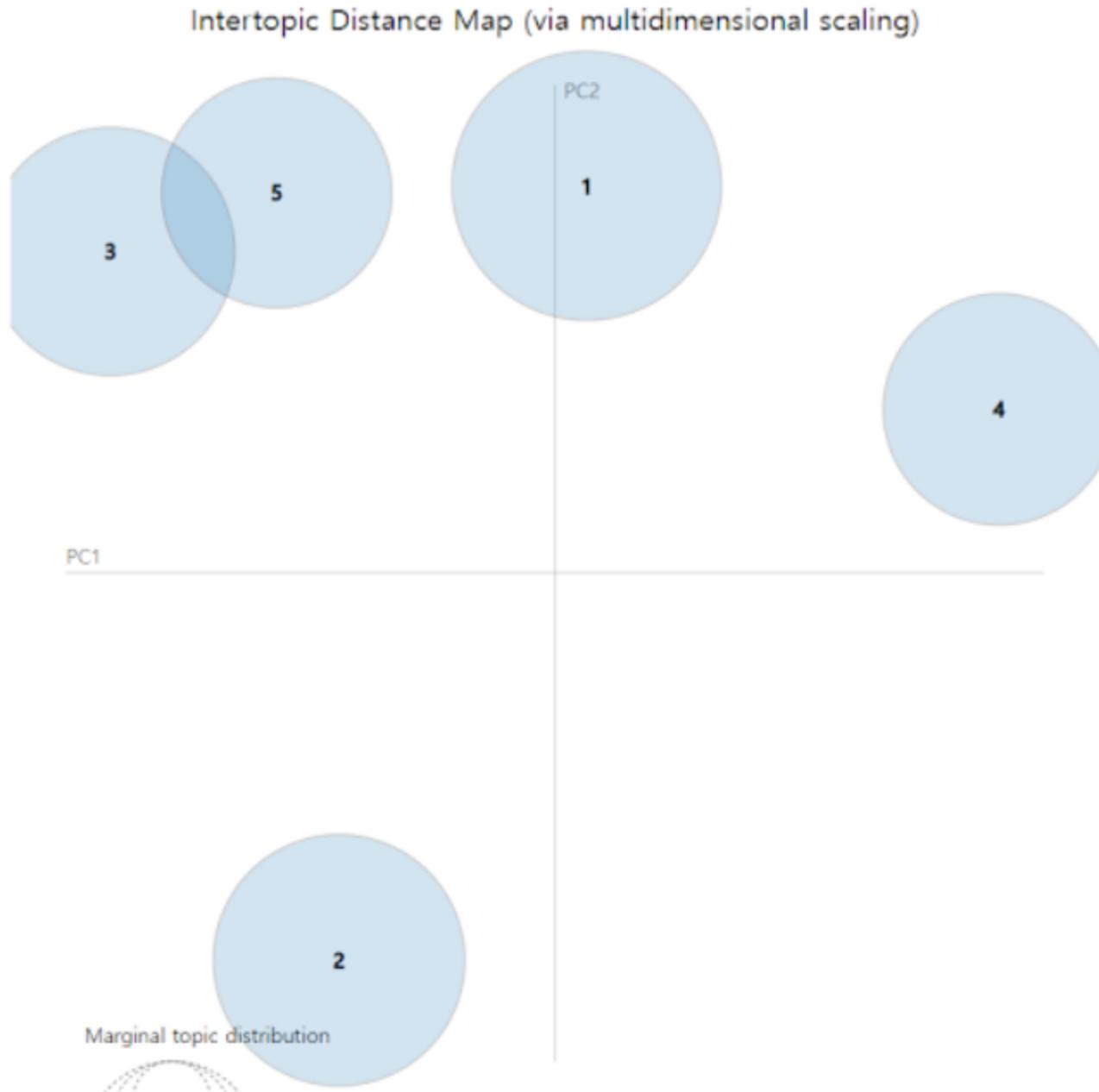
Topic # 4 -----
65024
105077
65502

Topic # 5 -----
82017
109158
117275
```



3-3 Total Data_iteration:500

Tuning Value - iteration: 500 / passes: 5



Topic_n: 5



Topic_n: 10

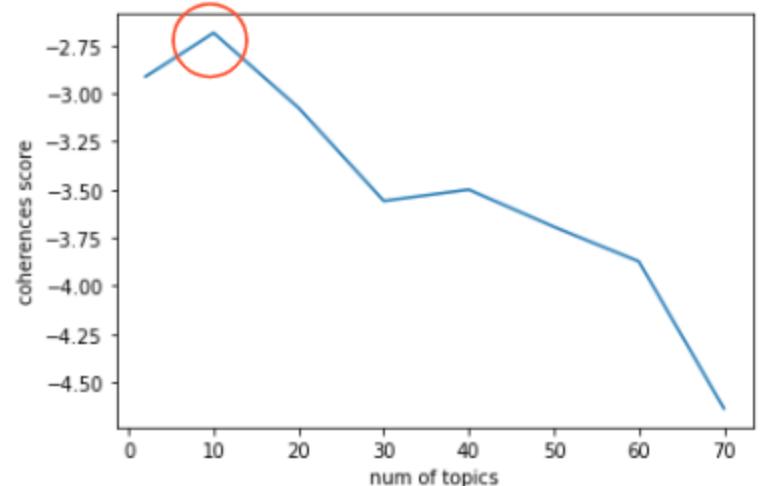
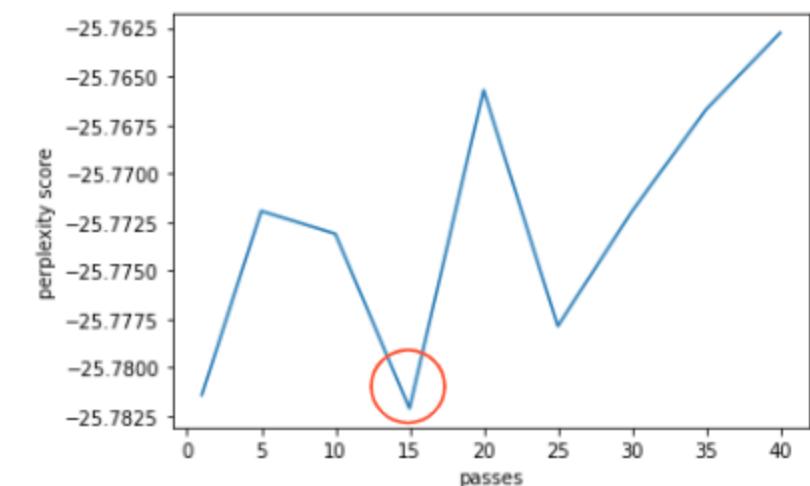


3-3 Total Data_iteration:400

1. Tokenize

```
[ (0,
  '0.042*"버거" + 0.013*"패티" + 0.010*"햄버거" + 0.009*"감자_튀김" + 0.0
(1,
  '0.033*"파스타" + 0.021*"피자" + 0.016*"와인" + 0.012*"스테이크" + 0.0
(2,
  '0.012*"갈비" + 0.010*"삼겹살" + 0.009*"쌀국수" + 0.007*"소고기" + 0.0
(3,
  '0.032*"떡볶이" + 0.024*"초밥" + 0.022*"김밥" + 0.016*"스시" + 0.010*"
(4,
  '0.034*"곡초" + 0.022*"고지" + 0.015*"만건리" + 0.008*"대초" + 0.008*"
```

2. Tuning



3. Assign_Topic

Doc_Num	Topic	Percentage	id	
0	0	2	0.731293	0 [숙소, 가깝다, 간짜장, 삼선짬뽕, 수타면, 쫄깃, 짬뽕, 국물]
1	1	1	0.304063	1 [미쿠, 풀다, 정겨운, 영양, 간식, 오는, 분위기, 안주, 미]
2	2	5	0.909986	3 [라떼, 한잔, 중년, 샷, 나중, 라떼, 커피, 이야기, 커]
3	3	5	0.709231	4 [커피, 연유, 라떼, 아이스, 바닐라, 라떼, 인증, 카페, 커]
4	4	5	0.627150	10 [모든, 디저트, 커피, 샌드위치, 젤, 커피, 와플, 콘센트,

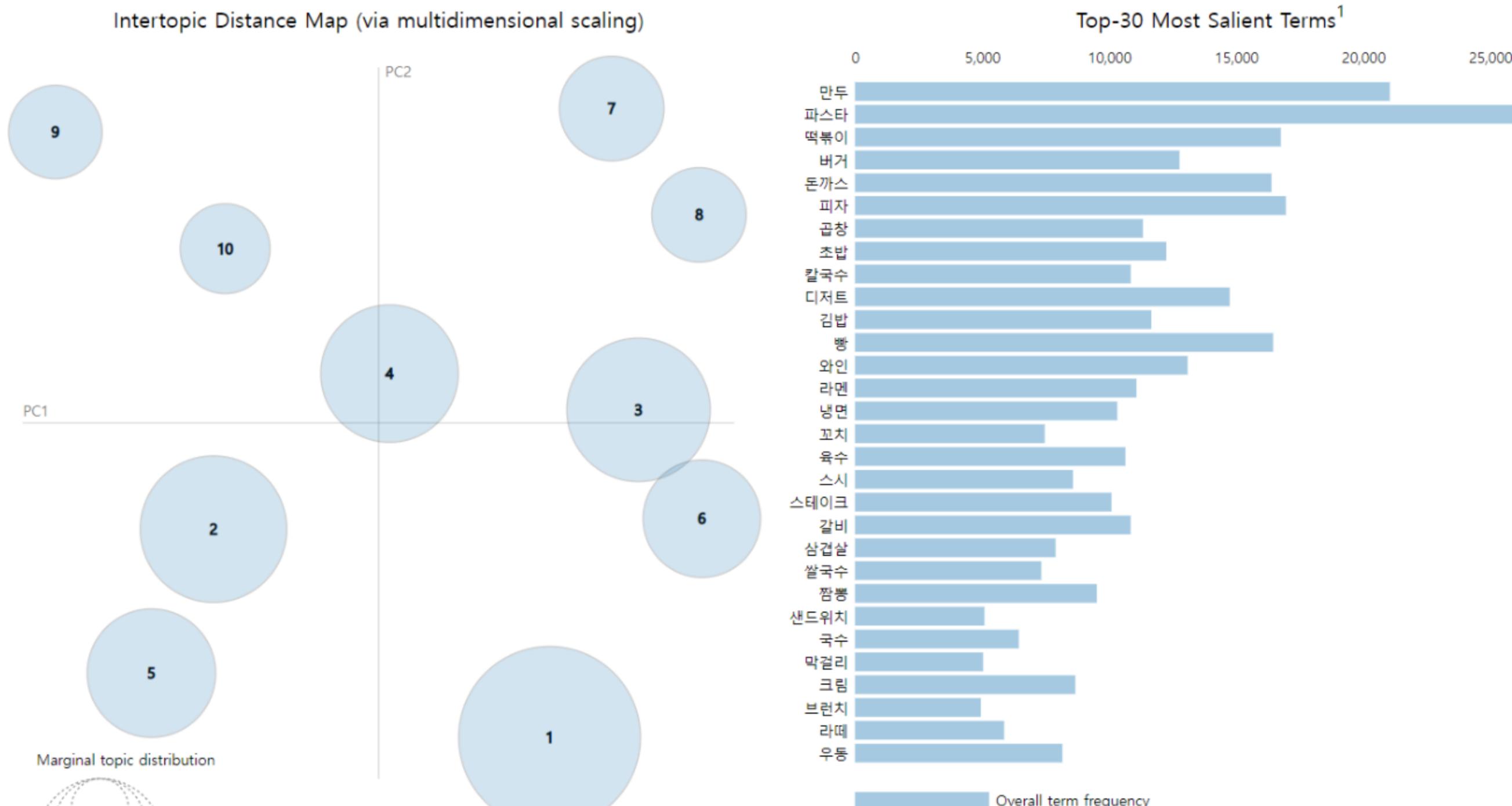
4. Check_Group

Topic # 0 -----	Topic # 3 -----
90153	85994
69252	98252
27592	87681
Topic # 1 -----	Topic # 4 -----
106725	65024
983	105077
98587	65502
Topic # 2 -----	Topic # 5 -----
44357	82017
8169	109158
553	117275



3-3 Total Data_iteration:400

Tuning Value - iteration: 400 / topic_n: 10 / passes: 15





3-4 Theme 선정

1. '갈비, 삼겹살, 육회, 막걸리, 소고기, 김치찌개, 된장찌개, 냉면, 구이, 한우',
 - 저기압일땐 고기앞으로!

2. '돈까스, 버거, 곱창, 족발, 패티, 카레, 돈가스, 햄버거, 대창, 카츠',
 - 스트레스 받는 날, 칼로리 폭탄
 - 기름 넣으러 가자, 칼로리 폭탄

3. '디저트, 빵, 크림, 라떼, 케이크, 아메리카노, 스콘, 아이스크림, 브런치, 이쁘다'
 - 디저트 배는 따로 남겨놔야지

4. '떡볶이, 김밥, 순대, 국밥, 오징어, 오뎅, 쭈꾸미, 감자탕, 설렁탕, 깍두기',
 - 뭘 먹을지 모르겠다면, 무난하게

5. '라면, 쌀국수, 우동, 계란, 일본, 육수, 마제_소바, 한국, 차슈, 비비다',
 - 한국에서 느끼는 일식의 맛

6. '만두, 칼국수, 꼬치, 육수, 냉면, 국수, 콩국수, 삼계탕, 평양_냉면, 수육'
 - 라떼는~ 같이 갈만한 곳 -> 부모님도 좋아할걸?
 - 등산 후 생각날걸? 등산 후 먹을게 없다면
 - 어르 '신과 함께'

7. '와인, 칵테일, 술집, 이쁘다, 한잔, 음악, 편안하다, 데이트, 간단하다, 루프_탑',
 - 오늘은 분위기에 취하고 싶다...★

8. '짬뽕, 탕수육, 짜장면, 중식, 안내, 중국집, 짜장, 글, 식다, 닭발',
 - 어머님은 짜장면이 좋다고 하셨어

9. '초밥, 스시, 연어, 참치, 장어, 닭갈비, 사시미, 우동, 숙성, 사케',
 - 날로 먹는 날

10. '파스타, 피자, 스테이크, 와인, 토마토, 빵, 샌드위치, 리조또, 레스토랑, 크림'
 - 연인과의 첫 데이트
 - 소개팅가는 날
 - 오늘 널 유혹하겠쉬

Topic-Theme DataFrame

Dominant_Topic	theme
0	0 어머님은 짜장면이 좋다고 하셨어
1	1 한국에서 느끼는 일식의 맛
2	2 디저트 배는 따로 남겨놔야지
3	3 기름 넣으러 가자, 칼로리 폭탄
4	4 소개팅가는 날
5	5 어르 '신과 함께'
6	6 뭘 먹을지 모르겠다면, 무난하게
7	7 오늘은 분위기에 취하고 싶다...★
8	8 저기압일땐 고기앞으로!
9	9 날로 먹는 날



Theme 선정

테마	키워드
저기압일땐 고기 앞으로!	갈비, 삼겹살, 육회, 막걸리, 소고기, 김치찌개, 된장찌개, 냉면, 구이, 한우
기름 넣으러 가자, 칼로리 폭탄	돈까스, 버거, 곱창, 족발, 패티, 카레, 돈가스, 햄버거, 대창, 카츠
디저트 배는 따로 남겨놔야지	디저트, 빵, 크림, 라떼, 케이크, 아메리카노, 스콘, 아이스크림, 브런치, 이쁘다
뭘 먹을지 모르겠다면 무난하게	떡볶이, 김밥, 순대, 국밥, 오징어, 오뎅, 주꾸미, 감자탕, 설렁탕, 깍두기
한국에서 느끼는 일식의 맛	라멘, 쌀국수, 우동, 계란, 일본, 육수, 마제, 소바, 한국, 차슈, 비비다
어르 '신과 함께'	만두, 칼국수, 꼬치, 육수, 냉면, 국수, 콩국수, 삼계탕, 평양냉면, 수육
오늘은 분위기에 취하고 싶다	와인, 칵테일, 술집, 이쁘다, 한잔, 음악, 편안하다, 데이트, 간단하다, 루프탑
어머님은 짜장면이 좋다고 하셨어	짬뽕, 짜장면, 탕수육, 중식, 안내, 중국집, 짜장, 글, 식다, 닭발
날로 먹는 날	초밥, 스시, 연어, 참치, 장어, 닭갈비, 사시미, 우동, 숙성, 사케
소개팅 가는 날	파스타, 피자, 스테이크, 와인, 토마토, 빵, 샌드위치, 리조또, 레스토랑, 크림

4. 서비스 시연



여기까지
다음





5. 마무리

5-1. 한계점 및 기대효과

5-2. 참고자료





한계점

1. 서울 지역 한정으로 서비스를 제공
2. API를 통해 음식점을 가져오다 보니 정보의 최신화가 빠르지 않음
3. 크롤링으로 데이터를 수집하는 시간이 오래 걸려서 더 많은 음식점의 정보를 제공하지 못함

기대효과

1. 테마별, 날씨별 음식점 추천을 통해 소비자에게 다양한 정보 제공 및 외식 문화 활성화
2. 서울 상권을 분석하고자 하는 사람들에게 정보를 제공
3. 서울에 음식점 창업을 하고자 하는 사람들에게 정보를 제공
4. sns마케팅을 하고자 하는 점주들에게 홍보효과 기대
5. 서울 음식점 상권 활성화



논문

1. 김재현(2019), 「단어의 중요도와 연관성을 고려한 키워드 추천 방법」, 조선대학교 산업기술융합대학원, 31p.
2. 박상현(2017), 「토픽모델링과 인공신경망에 기반한 온라인 쇼핑몰 리뷰 데이터 분류 및 응용」, 경희대학교 대학원, 51p.
3. 최환석, 팽전, 이우섭(2020), 「머신러닝 기반 음식점 추천시스템 설계 및 구현」, 『디지털콘텐츠학회논문지』 제 21권 제 2호, 259p.-268p.
4. 연종흠, 이동주, 심준호, 이상구(2011), 「상품 리뷰 데이터와 감성 분석 처리 모델링」, 『한국전자거래학회지』 제 16권 제 4호, 125p.-137p.
5. 윤상훈, 김근형(2021), 「Word2Vec를 이용한 토픽모델링의 확장 및 분석사례」, 『정보시스템연구』 제 30권 제 1호, 45p.-64p.
6. 조찬열, 정구임, 서양민, 최예림(2017), 「날씨 및 요일 특성이 음식점 메뉴 검색시스템 이용에 미치는 영향에 관한 실증 연구」, 『스마트미디어저널』 제 6권 제 2호, 50p.-56p.

블로그

1. 이영민, <LDA - R을 이용한 토픽 분석>, <<공간 감성어 사전 구축 기법>>, <https://brunch.co.kr/@mapthecity/2>, (2016.07.06)
2. <[NLP] KoNLPy 이용하여 한국어 토큰화, 형태소 분석하기 및 클래스간 품사 태그 비교표 [한국어 자연어처리]>, <<고졸입니다만>>, <https://mr-doosun.tistory.com/22>, (2021.06.28)
3. hyk0425, <한국어 분석(형태소 분석)>, <<야금야금>>, <https://hyk0425.tistory.com/14>, (2020.07.25)
4. Junseokkk, <BTS의 성공요인 분석 - 토픽모델링과 소셜 네트워크 분석 활용>, <<Data공부>>, <https://junseokkk.tistory.com/5>, (2021.08.26)



블로그

5. broccoliindb, <mysql 데이터를 elasticsearch로 검색하기>, <<bloccoliindb.log>>, <https://velog.io/@broccoliindb/mysql-%EB%8D%B0%EC%9D%B4%ED%84% B0%EB%A5%BC-elasticsearch%EB%A1%9C-%EA%B2%80%EC%83%89%ED%95%98%EA%B8%B0>, (2021.05.20)
6. lulurara, <[mySQL]컬럼 추가삭제, 외래키 지정/삭제, 제약조건확인>, <<lulurara>>, <https://happylulurara.tistory.com/127>, (2020.11.26)
7. 효벨, <Mysql 특정 테이블 Dump>, <<개발자의 끄적끄적>>, <https://solbel.tistory.com/1321>, (2021.01.06)
8. 강명훈, <Logstash 윈도우 파이프라인>, <<Easy to analyze if you can cleaning data>>, <https://kangmyounghun.blogspot.com/2019/10/logstash.html>, (2019, 10,13)
9. 밤둘레, <Mysql ERD 추출>, <<Bamdule>>, <https://bamdule.tistory.com/44>, (2020.01.27)
10. 까만손오공, <[문제해결] You are using safe update mode>, <<까만손오공>>, <https://blog.naver.com/kkson50/221251167091>, (2018.04.12)
11. 관리자, <Mac brew ELK (Elasticsearch + logstash + kibana) 설치 + mysql 연동>, <<TRANDENT>>, <http://trandent.com/article/etc/detail/323366>, (2022.01.18)
12. <[django] Model IntergerField에 최소/최대값 지정하기>, <<June Dev Blog>>, <https://enfanthoon.tistory.com/180>, (2021.01.19)
13. 초보군붕이, <ubuntu 18.04에서 mysqlclient 설치 오류 해결>, <<개발자를 꿈꾸는 군인의 일기장>>, <https://iamiet.tistory.com/54>, (2021.07.18)
14. 마이쮸, <Django(장고)에서 form을 이용한 style 활용법>, <<IT, I Think>>, <https://cholol.tistory.com/510>, (2021.02.03)



참고 사이트

1. pyspark

<https://spark.apache.org/docs/latest/api/python/reference/index.html>

2. pyproj

<https://pyproj4.github.io/pyproj/stable/>

3. kakao map

<https://apis.map.kakao.com/web/documentation/>

4. airflow

<https://airflow.apache.org/docs/>

5. [django]starfield

<https://pypi.org/project/django-starfield/#description>



Q&A





감사합니다

