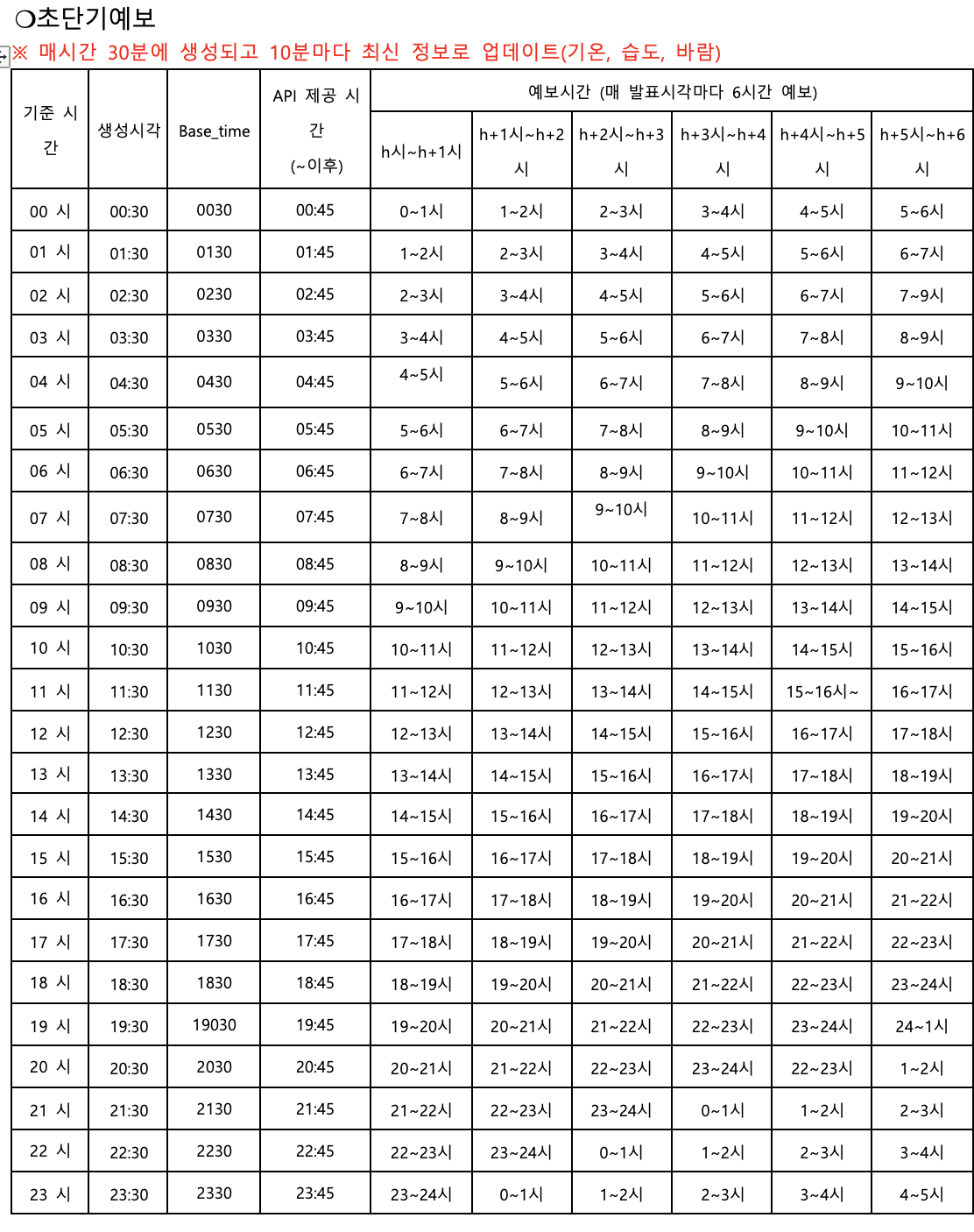
**1. 데이터 수집**

i. API 데이터



1. 서울 음식점 데이터

: 총 473,879개 => 이 중 영업상태가 ‘정상'인 데이터 124,500개

1. 날씨 데이터 => 기상청 (초단기예보) & 에어플로우(DE)



ii. 크롤링 데이터 : 총 4개의 사이트에서 진행

a. **네이버맵** - 26,121개 (리뷰 데이터 : 5,717개 )

b. **카카오맵** - 124,500개 (리뷰 데이터 : 63,951개)

c. **망고플레이트** - 74,328개 (리뷰 데이터 : 14,417개)

d. **다이닝** **코드** - 124,500개 (리뷰 데이터 : 16,760개 )

**2. 데이터 처리**

i. API 데이터 처리

* 영업상태가 ‘정상'인 데이터 124,500개 + 위도,경도 추가 + 도로명 주소 없는 가게 제거

=> 122,000개

ii. 크롤링 데이터 처리

* 상세정보 데이터 : 메뉴 정보, 태그, 영업 시간, 기타 정보(국민지원금, 주차장 등등)
* 리뷰 데이터

1. 카카오, 다이닝, 네이버, 망고에 리뷰가 없는 식당 삭제
2. []나 ‘’와 같이 리뷰가 없지만 null값으로 잡히지 않는 리뷰들 제거
3. 특수문자, 영어 제거
4. 가게별 리뷰 합치기

**3. 데이터 분석**

i. **데이터 전처리**

1. 토큰화 : koNLPy의 okt와 mecab 사용 (명사와 형용사 중심으로)
2. 불용어 제거 : 일반적인 리뷰 불용어 리스트 수집하여 확인 후 통합

군집화된 결과를 보고 불용어 추가 제거

ii. 데이터 분석 모델링 - **Sample Data** (Kakao 리뷰 1만개)

1. LDA 모델링 (LDA로 군집 형성 확인 후 리뷰데이터 전처리 추가 진행)

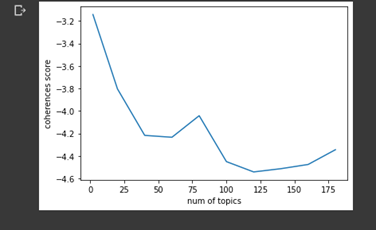
(kakao 음식점 데이터 10000개로 진행)

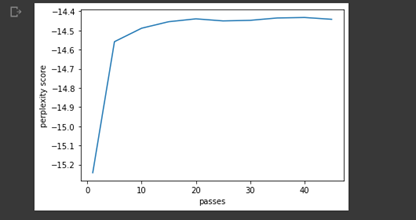
* 1. tokenize 결과가 공백인 것, 단어들이 하나로 묶여있는 것 제거
  2. pyLDAvis를 보며 단어 불용어 추가 (700여개)
  3. 리뷰별 토픽요소 정보 확인

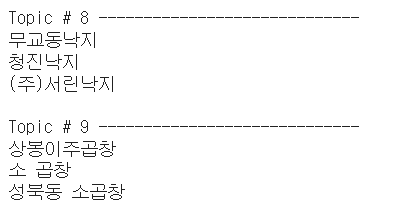
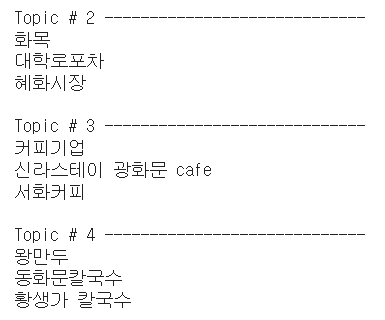


* 1. perplexity와 coherence를 확인하며 LDA 모델 수정

(perplexity가 낮은 것으로 passes 수정)

(coherence가 높은 것으로 topic 개수 수정)

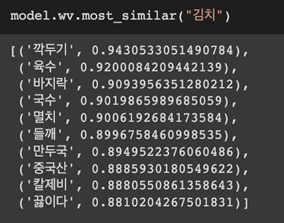
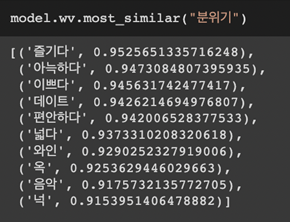
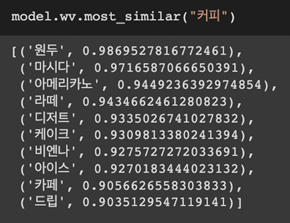


* 1. 음식점 별 가장 가까운 topic으로 할당
  2. 할당된 토픽별 음식점 그룹 확인

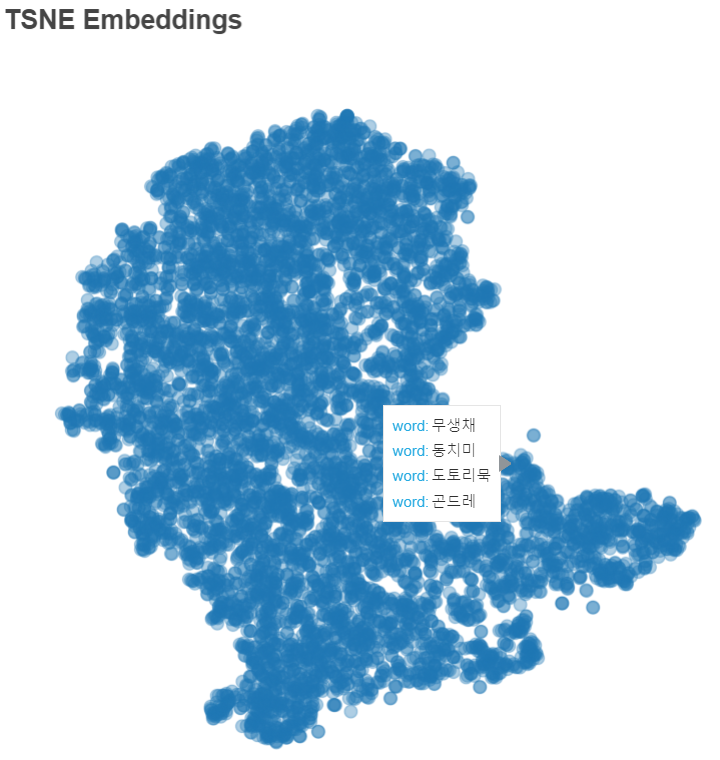
1. LDA & W2V 모델링(kakao 음식점 데이터 10000개로 진행)

(중심 단어를 LDA로 추출 후 주변 단어와 묶기 위해 W2V 진행)

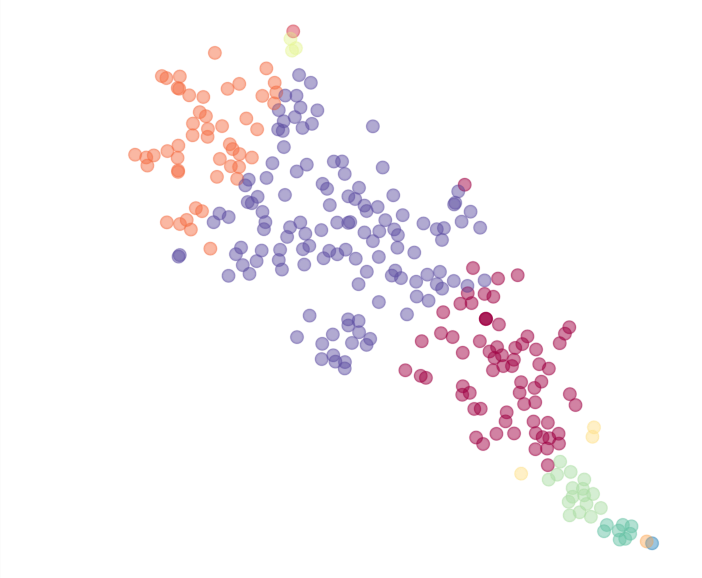
* 1. tokenize, 불용어 제거 등 전처리
  2. 몇개의 단어들로 W2V 확인



* 1. 분포를 확인하기 위해 T-SNE로 차원 축소 후 Bokeh와 plot으로 시각화



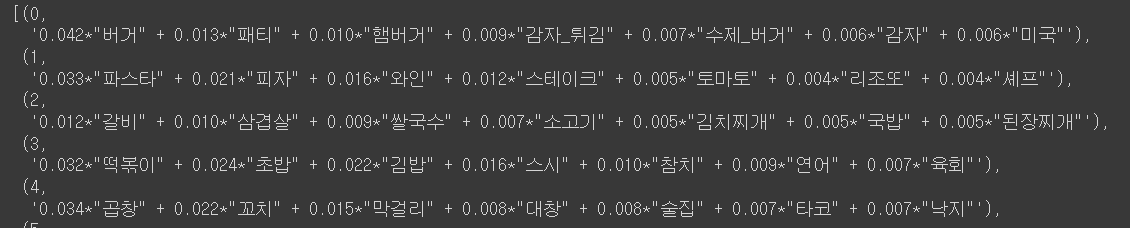
**단어 분포시각화**



**단어 클러스터링 - 음식점 분포 시각화**

ii. 데이터 분석 모델링 - Total Data (4개 사이트 리뷰 통합본)

1. LDA 모델링 1차 - iteration 400
   1. Sample Data로 진행한 전처리 적용
   2. pyLDAvis를 보며 단어 불용어 추가
   3. 리뷰별 토픽요소 정보 확인

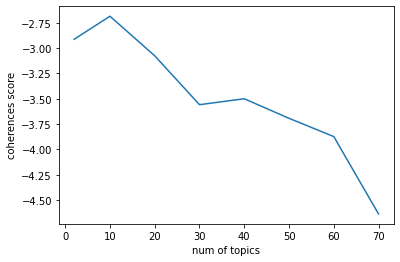
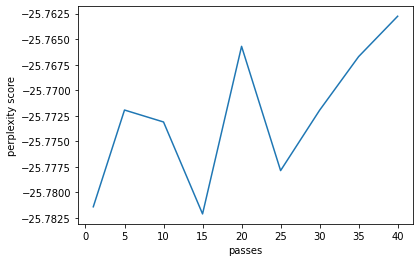


* 1. **perplexity(혼란도)**와 **coherence(일관성)**를 확인하며 LDA 모델 수정

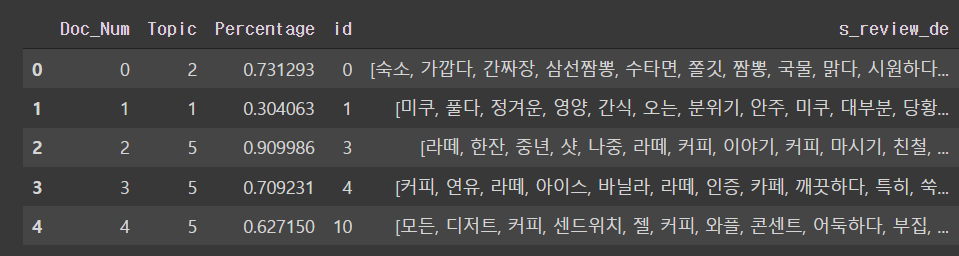
(perplexity가 낮은 것으로 passes 수정)

(coherence가 높은 것으로 topic 개수 수정)

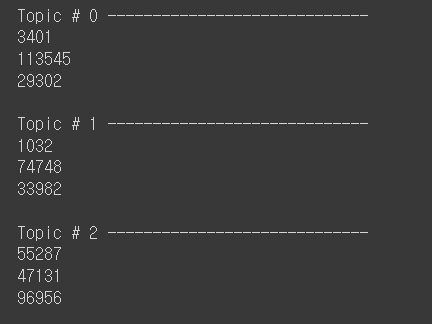
perplexity coherence



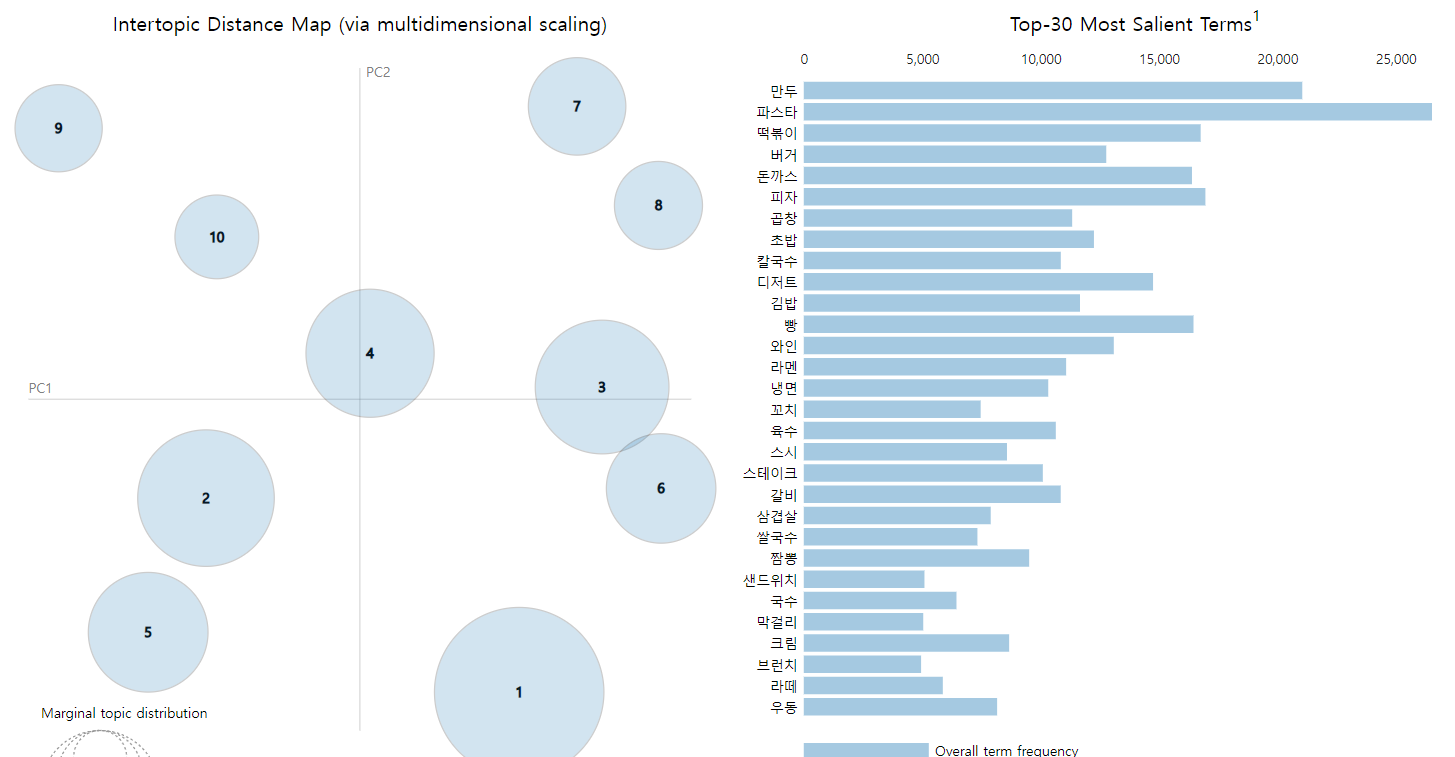
* 1. 음식점 별 가장 가까운 topic으로 할당



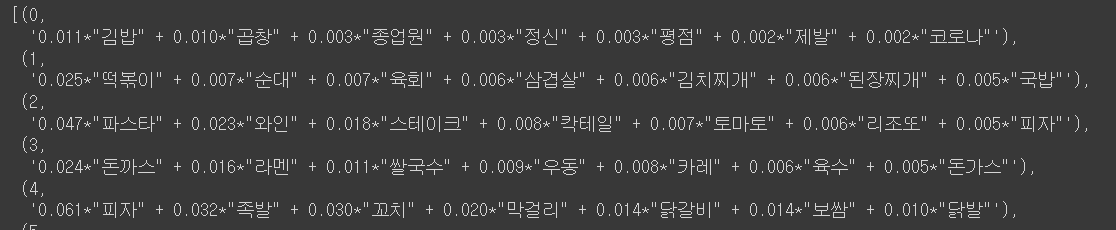
* 1. 할당된 토픽별 음식점 그룹 확인



* 1. pyLDAvis 군집화 확인

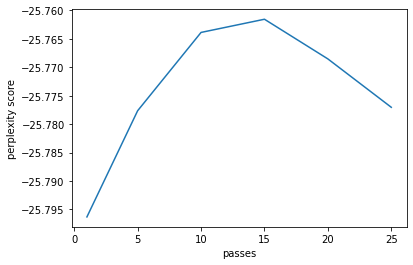
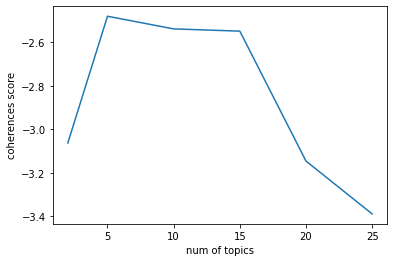


1. LDA 모델링 2차 - iteration 500
   1. Sample Data로 진행한 전처리 적용
   2. pyLDAvis를 보며 단어 불용어 추가
   3. 리뷰별 토픽요소 정보 확인

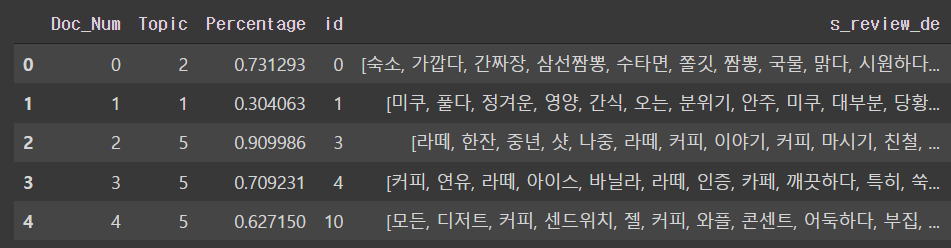


* 1. perplexity와 coherence를 확인하며 LDA 모델 수정

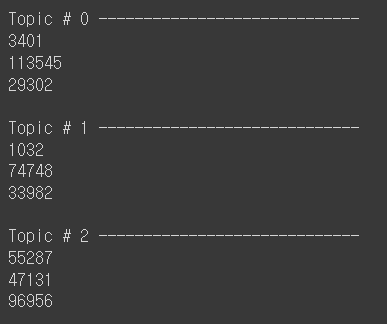
(perplexity가 낮은 것으로 passes 수정)

(coherence가 높은 것으로 topic 개수 수정)

* 1. 음식점 별 가장 가까운 topic으로 할당

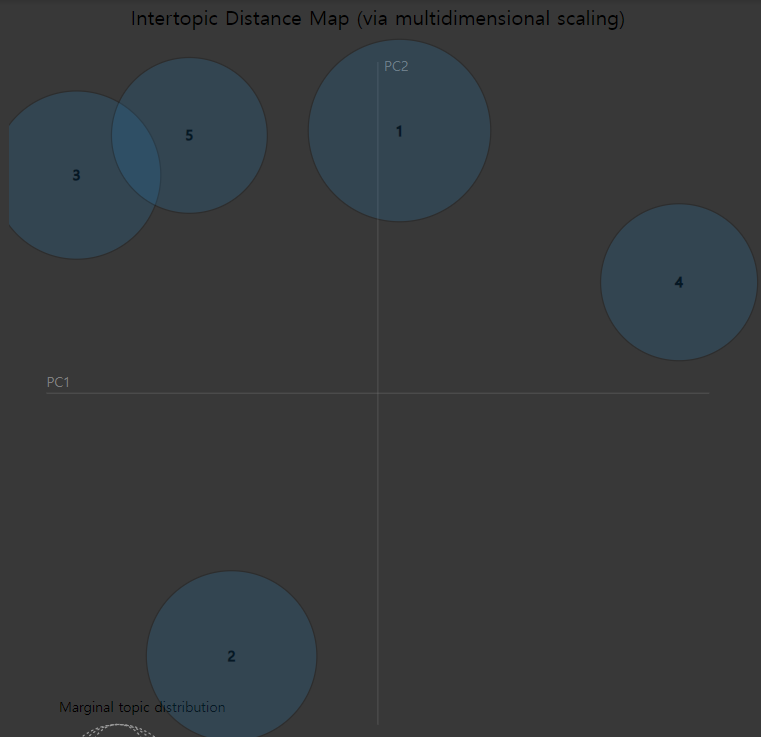
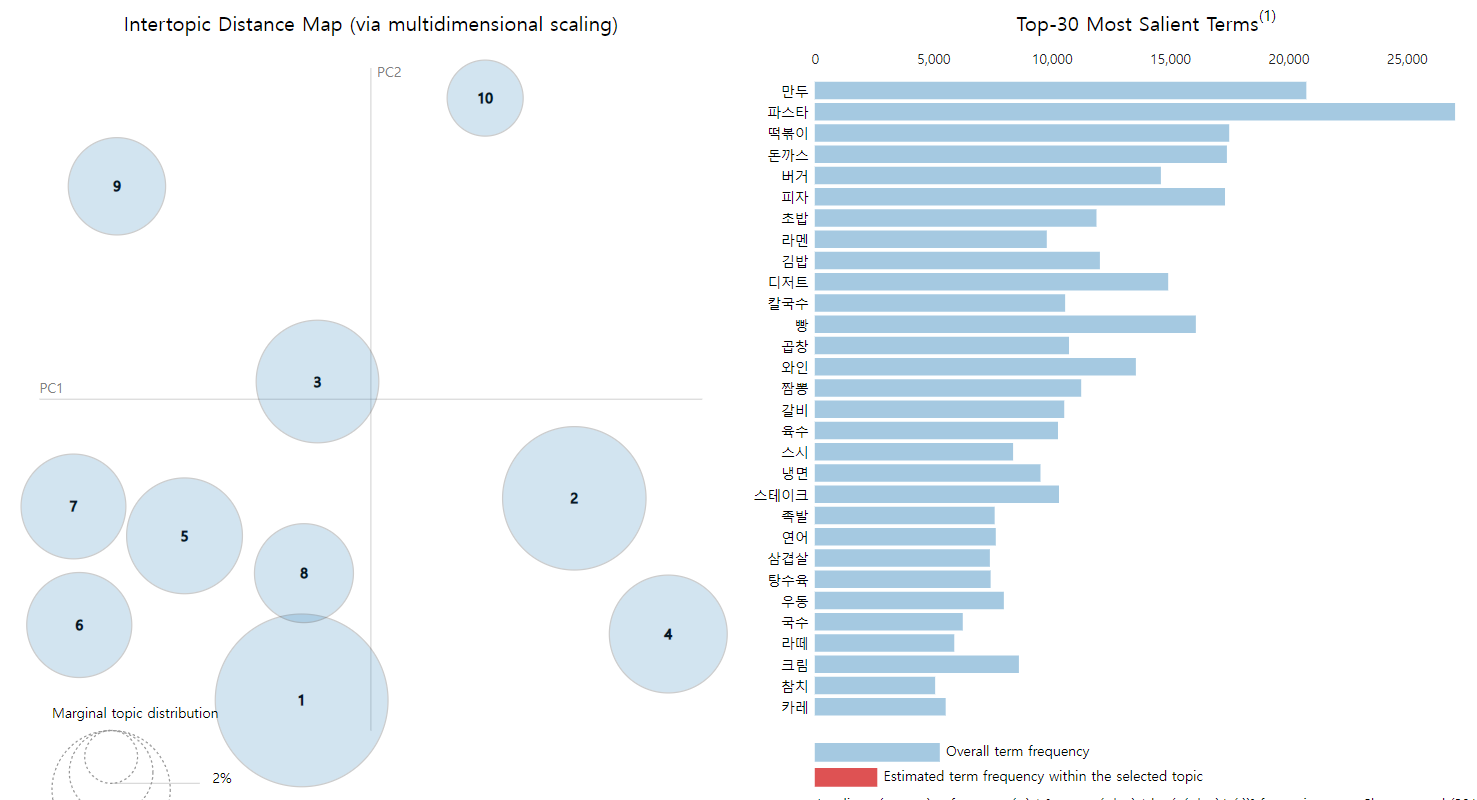


* 1. 할당된 토픽별 음식점 그룹 확인

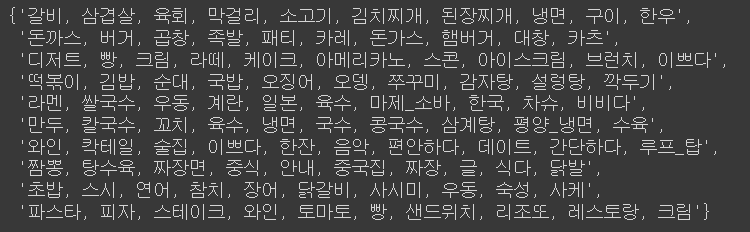


* 1. pyLDAvis 군집화 확인

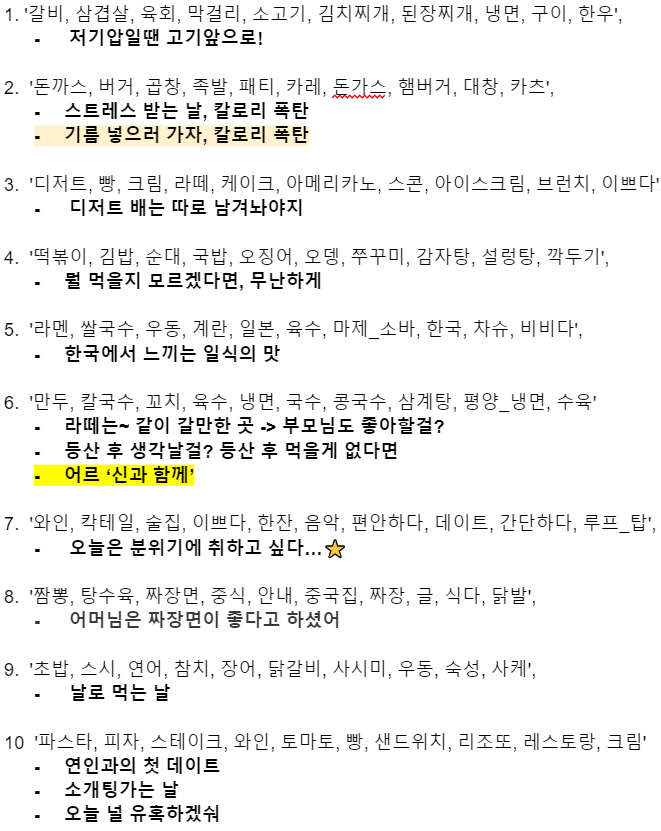
1. topic\_n = 5 2) topic\_n = 10

1. LDA 모델링 설정 완료 - iteration : 400, topic\_n : 10, passes : 15
   1. 키워드 확인



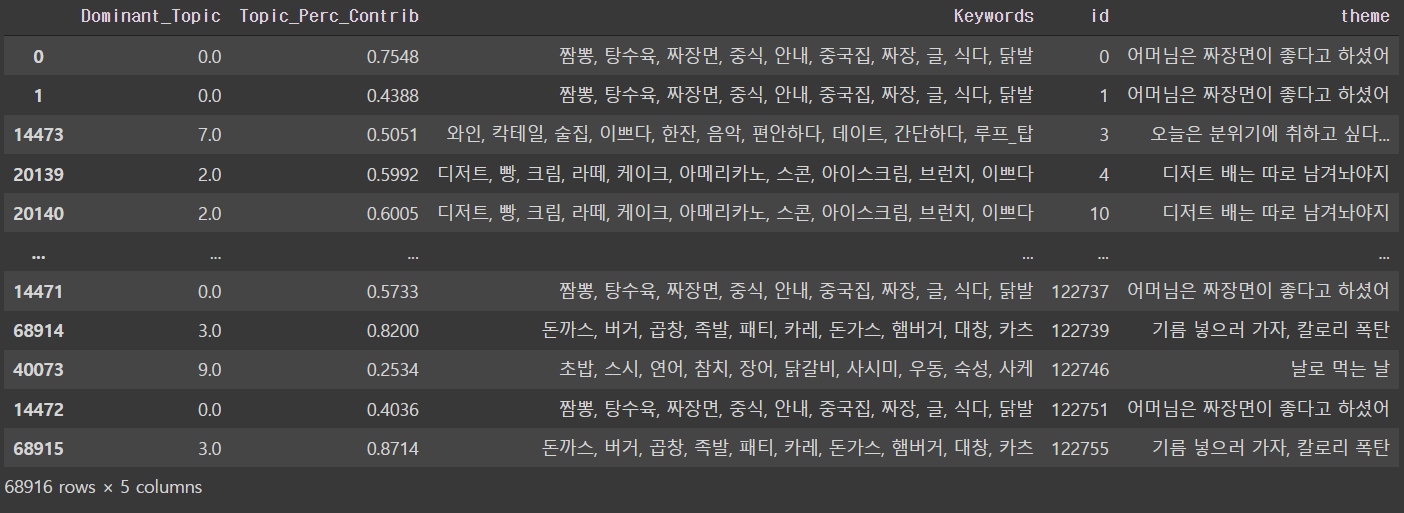
* 1. 토픽-키워드 별 테마 선정



* 1. 토픽에 테마 연결



* 1. 테마 및 토픽 dataframe

x

(별첨) 사용한 불용어 - stopwords.txt (1086개)

임에 영 종업원 사 좋아하다 만족 돼다 중 어느해 비추어 보아 정도의 아 뿐만 아니라 즉시 무엇 남들 동안 마르다 어느때 가보다 정신 어떻게 로우 그에 따르는 아니었다면 자기집 그러 가격 전부 봉투 서빙 아주 느끼하다 지 더구나 필링 메쓰겁다 진료 아홉 먹다 하기 거의 바꾸어서 옥 휴 그위에 어느쪽 더라도 가지 세 헐떡헐떡 전화하다 부터 제일 그런데 어느것 얘기 과 넓다 대로 하다 개의치않고 이럴정도로 엄청 그중에서 하는 김에 만들다 날 어떻 모두 내다 안나 귀엽다 위해서 이천칠 지경이다 이외에도 하도다 바뀌다 잔 당신 조선족 못하다 따름이다 어떠한 음식 설명 대하여 만일 하는것만 주 등등 이곳 죽다 난 통하여 그치지 남기다 여자 매장 찍다 할 수 있어 있어 가성 아이고 통하 일지라도 추천 때가 되어 플 무조건 와아 하곤하였다 들다 도달하다 것들 결제 줄은 몰랏다 줄은모른다 오다 평점 만 마음대로 만은 뿐만아니라 즉 한 후 정말 할 무난 고양이 낫다 건물 비로소 하마터면 같이 이유는 못하 의해되다 어기여차 얼마나 누구 알 이어서 무슨 물어보다 항상 웨이 기름 취소 향하여 따르 현금 기분 잠깐 화장실 하도록하다 아이 얼마든지 하고 하겠는가 운운 영업 옆사람 결론을 끝장 오히려 본인 보다더 하자마자 하면 할수록 년도 칠 계산 무릎쓰고 비록 말 너희 식당 했어요 요컨대 완전 이 전후 하도록 로 인하여 보면 향하다 중에서 보드득 밥 좋다해도 여 방문 사과 굳다 뭐 가끔 이만큼 사다 길음 재밌다 이 때문에 확인 하면된다 고로 매번 집 대로 존경 방역 으로써 바로 로써 앞의것 으로 그리하여 꾸다 생각한대로 더욱더 점 그들 조금 많다 퉤 와르르 점에서 보아 들자면 다니다 허허 처음 너희들 가 둘 주저하지 관련이 때 이 정도의 어쨋든 하면 솨 탕탕 에 달려 있다 입각하여 이렇게 직원 ㆍ 어떻다 말투 좋아 할 줄 안다 대하면 마치 이렇 얼마큼 정도에 이르다 더군다나 이로 모르다 반점 습니까 않으면 안 그러면 하물며 여섯 육 비걱거리다 하기보다는 해서는 안된다 좀 이렇구나 모른다 좋 결론을 낼 수 있다 그러니 일때 아쉽다 륙 반드시 과연 언제 근거로 말하 하고있었다 같은 따지지 않다 다양하다 그녀 훨씬 얼마간 번 응대 저희 ㅋㅋㅋ 공간 절대 홀 비교적 한적이 제발 받다 관계가 취지 의사 함으로써 예쁘다 넣다 아무도 그렇지만 할때 않는다면 하나 한 향해서 다른 방면으로 생각이다 해도 달려 같다면 봉\_탁주 쓰여 할 지경이다 헉헉 저것만큼 여전히 흥 그럼 해야한다 힘입어 등 개 살 와 방면으로 하다 아깝다 조차도 관해서는 알바생 울다 사먹다 ㅋㅋ ㅠㅠㅠ ㅜㅜ ㅠㅠ ㅎㅎ 맛있다 사장 님 사장 사장님 이천팔 만들 의 시초에 이쪽 마저 어떻해 바꿔말하면 물론 앉다 덮다 이와 같다 할지언정 화이팅 저것 토하다 설마 한하다 이렇게되면 요만큼 지든지 주방 으로서 저 어째서 줄은 이와 가령 그때 비슷하다 혹은 할 힘이 있다 위하 지금 로부터 돈 만약 갈다 바꾸어서 한다면 그냥 일곱 마시다 생각 겨우 하지마 오호 인 듯하다 몰라도 이것 겸사겸사 다시 자다 대행 않 삼 년 할줄알다 ㅇㅇ 한마디 봐라 대국 한다면 몰라도 지말고 조차 의해서 일 것이다 된바에야 따라서 받 해요 와 같은 사람들 아니면 논하지 시 포장 라 해도 이러이러하다 팅 손님 어찌하여 한다면 쉿 이지만 맛없다 크다 그렇다 시키다 바와같이 그에 안다 아니라면 든간에 더럽다 보아 함께 하기는한데 그런즉 버금 뿐만 퀄리티 흐흐 것과 또한 그리고 시키 주룩주룩 없다 올라가다 일단 할 따름이다 병원 저기 다음으로 고려하면 다만 쪽으로 수 주문 이와같다면 말할것도 없고 써다 할만하다 위생 얼마만큼 전 총적으로 친절하다 느낌 다시 말하자면 하는것이 낫다 종류 및 을 다를가 허 많은 부드럽다 이러한 무엇때문에 적 순 틈타 말다 되다 뚝뚝 인테리어 된다 비싸다 따로 이 외에 어찌됏든 그만이다 팍 우리들 이천육 그렇게 배민 예하면 에게 아니나 야 해서는 무료 재 있 너무 요만한걸 그러한즉 만큼 주다 놓다 할 생각이다 한켠으로는 하지 꼭 기다리다 바꾸다 참나 그럼에도 테이블 찍히다 쳇 셋 까닭에 아니다 아니라 주시 맵다 차라리 어울리다 다시다 하하 까닭으로 하기 위하여 알겠는가 설령 없고 크 안 모르 그렇 음료 굳이 짱 어느곳 어디 논하지 않다 제 그렇지않으면 뒤따라 대해 가다 딱 관하여 하게될 자주 설사 총적으로 보면 두 아이구 한번 총적으로 말하면 걸리다 다수 리뷰 첫번째로 이때 툭 그래 해도된다 구토하다 달라 그저 나 이로 인하여 않다 메뉴 댕그 타인 강아지 말씀 점에서 하는바 보이 냄새 잇따라 여보시오 사회 분 별로 안좋다 놓 이와 같은 놀라다 좍좍 공동으로 윙윙 연관되다 봉 다섯 밖에 안된다 꽈당 이젠 하지만 로 어찌하든지 하지 않도록 기대여 졸졸 구성 에 대해 가세 사람 소인 나오 우리 싫다 예 하구나 의지하여 각 특 습니다 하여금 반찬 불구하고 까지도 자 최고 마스크 되어 정착 예를 장소 할 수 있다 없 여차 생선 인젠 굽다 대다 것과 같이 바꾸어 자기 비추어 소리 근거하여 되는 만원 서서 하기만 ~\_인분 더불어 바꿔 말하면 더 이용하여 만약에 따위 남자 일반적으로 기침 세상 실로 이 되다 다시말하면 알바 엉엉 휘익 때문에 즐기다 가게 이와 반대로 의해 각각 수교 갖고말하자면 변하다 여러분 대하 요만한 것 주인 팩 진짜로 싶 내심 값 어때 모 괜찮다 영차 웨이\_팅 아야 어느 년도 첫 번째로 우에 관한 찾다 이런 앗 왜 계속 속 기본 하기만 하면 팔 이렇다 한 까닭에 이르다 위하여 견지에서 퍽 아래윗 말하자면 것이다 쾅쾅 언젠가 시각 다음 이천구 알았어 짜다 하여야 웨이팅 오르다 많이 자리 안된다 다를까 전자 한데 후 기준으로 가까스로 생각하 아줌마 규 따라 싸우다 쿵 오로지 쓰다 하지마라 말할것도 데 김에 여덟 카드 둥둥 고객 심지어 하게하다 저쪽 오 에 움 보이다 본대로 혹시 때문 사람들 얼마 안 되는 것 요만한 가서 먼저 하는것도 이르기까지 몰랏다 어투 죄송하다 할망정 내 알 수 있다 어쩔수 없다 부류의 붕붕 경우 틀림없다 동시에 합리 땡기다 아무거나 근처 어찌 차다 하도록시키다 ㅡㅡ 우선 낼 하든지 층 들면 다음에 마저도 인 기점으로 위에서 이래 후기 미치다 그러면 우르르 불친절하다 세트 많은 것 이리하여 다른 하게될것이다 않다면 에 가서 시간 되 몇 찌다 두 번째로 자마자 비길수 없다 단지 하려고하다 기타 앞 한항목 만 못하다 바꾸어서 말하면 편이 밖에 반대로 어쩔수 넘다 양도 마리 얼마 즈음하여 마카 맘 비다 비하면 할수록 하기 때문에 에서 하는 것이 그래도 드리다 그렇지 않다면 펄렁 또 들어가다 까악 콸콸 대박 주차 바꾸어말하자면 하였다 예약 불 힘들다 아니 일 까지 미치다 제외하고 나오다 버리다 미안하다 아이쿠 신선하다 이라면 신부 들 따지지 같 임에 틀림없다 지만 이봐 않기 위해서 인하여 공부 정갈하다 그치지 않다 여기 천원 까지 봐 발렛 존맛탱 되어다 할지라도 나가다 형식으로 아주머니 대해서 응 이유만으로 조심하다 만이 하고 있다 예를 들면 하는 비길수 서비스 코스 넷 어이 뿐이다 당장 참 하는것만 못하다 무렵 그래서 게우다 어 연이서 보는데서 서술한바와 막론하고 그러나 남짓 할뿐 힘이 많 바꾸어말하면 한적이있다 하 한 이유는 태도 곡 뒤이어 구 외에도 이번 외에 월 안되다 앞에서 그런 까닭에 씨 굿 아울러 끙끙 때가 그러므로 푸다 관리 그렇지 않도록 싸가지 를 정도에 잠시 결과에 이르다 듯하다 불문하고 예컨대 원 정도 어떤것들 이 밖에 하는 편이 낫다 자신 따르는 쿠폰 해도좋다 분들 유명하다 매 이었다 어떤 추가 바라다 그렇지 않으면 다소 된이상 멋지다 입장에서 하면서 너 혼자 그러니까 사실 하기에 하더라도 그 된 건지다 각자 종합한것과 답다 보 . 의거하여 해주다 지점 담배 근데 관련이 있다 어느 조용하다 나머지는 생각나다 명 나쁘다 네 소생 오직 삐걱 양은 따뜻하다 결국 않기 바와 시작하여 것 책 맞다 애기 거바 타다 줄 곧 문제 거니와 중의 아이야 삐걱거리다 이상 에 한하다 관계없이 깔끔하다 제각기 같다 오자마자 왜냐하면 최악 구체적으로 하곤 각종 아하 치료 게다가 여부 령 응당 양자 그런 서술한 맛집 예를 들자면 싶다 않고 더욱이는 적다 여기다 잘 하지 않는다면 부류의 사람들 튀다 좁다 인분 도착하다 허걱 약간 안암 보고 해봐요 있다 누가 헉 만은 아니다 어찌됏어 결과에 끼익 어떤것 주저하지 않고 하느니 딩동 말하면 상대적으로 그것 라

**참고 문헌**

**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, 최혜림(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 클래스간품사 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. Junseokk, <BTS의 성공요인 분석 - 토픽모델링과 소셜 네트워크 분석 활용>, <<Data공부>>, https://junseokk.tistory.com/5, (2021.08.26)