안녕하십니다. 탕탕특공대 발표 시작하겠습니다. (넘김)

발표 순서는 앞에 보이시는 것처럼 ~입니다.(넘김)

1.백종원?  
저희는 텍스트 마이닝을 활용한 백종원 성공요인 분석을 진행하였습니다.  
백종원은 한국의 유명한 셰프이자 요리 연구가로, 대중들에게는 '백선생'으로 더욱 잘 알려져 있습니다.

백종원은 다양한 요리 프로그램과 SNS를 통해 많은 이들에게 그의 요리 레시피와 요리 방법을 알리며 인기를 끌고 있습니다. 대중에게 사랑 받고 있는 셰프이자 인플루언서 백종원의 성공요인을 텍스트 마이닝을 통해 분석하고자 합니다.

(ppt를 보며 세부적으로 설명....각자 맡은 음식 카테고리 설명..)

2. 크롤링

저희는 총 2가지 방식으로 데이터 크롤링을 수행하였습니다.

1) 유튜브 API 입니다. 유튜브 API를 사용하면, 프로그래밍을 통해 유튜브 데이터를 쉽게 접근하고 수집할 수 있었습니다. 이 방법을 통해 우리는 채널 정보, 동영상 메타데이터, 댓글과 같은 다양한 정보를 손쉽게 얻을 수 있었습니다. 이 과정에서의 큰 장점 중 하나는 데이터 수집 자동화입니다. API를 통한 데이터 수집은 웹 크롤링에 비해 훨씬 안정적이며 효율적입니다. (넘김) 특히, API 호출을 자동화함으로써 대량의 데이터를 신속하게 수집할 수 있게 됩니다. (넘김)

2) 셀레니움을 사용한 크롤링입니다. 셀레니움은 복잡한 JavaScript 기반 웹사이트 크롤링에 최적화된 도구라 생각하였고 가장 큰 장점 중 하나는 바로 동적 콘텐츠 크롤링 능력입니다. 많은 현대 웹사이트가 동적으로 콘텐츠를 로드하기 때문에, 셀레니움 같은 도구 없이는 이러한 데이터에 접근하기 어렵다고 생각하였습니다. (넘김) 따라서, 셀레니움은 웹 크롤링에서 필수적인 도구 중 하나로 자리 잡고 있으며 필요한 데이터를 효과적으로 수집하고 분석을 가능케 도와줍니다. (넘김)

3. 토큰아이징 (넘김)

저희는 많은 데이터처리에 있어 속도와 정확도에 적절한 균형을 이루는 형태소 분석기가 필요하다고 느껴 카카오에서 개발한 형태소 분석기로 기존 사전과 규칙 기반 대신, 데이터 기반의 알고리즘으로 형태소를 분석하며, 국립국어원의 21세기 세종계획 코퍼스의 약 85만 문장, 천만 어절의 코퍼스를 사용하여 학습한 khaiii를 사용하였습니다. (넘김)

앞에 보이시는 것처럼 khaiii를 설치하는 과정은 이러합니다(넘김)

저희는 데이터 전처리 과정을 총 4단계로 진행하였습니다.(넘김)

1) 정제 과정입니다.

크롤링해 온 데이터를 보니 한국어 이외에 영어,이모티콘,특수 문자 등 불필요한 정보들이 포함되 있었습니다. 그래서 저희는 replace(),strip() 함수를 사용하여 제거하였습니다.

(넘김)

2) 정규화 과정입니다.

텍스트 데이터 내의 단어들을 일관된 형태로 통일하는 과정으로 리뷰 내의 다양한 표현을 통일하여 분석의 일관성을 보장하기 위해 필요하다고 생각하였습니다. 먼저 가장 많이 나오는 토큰을 나열후 수동적으로 동의어 처리 사전을 생성후 적용하였습니다.(넘김)

3) 어간/표제어 추출

자연어 처리(NLP)에서 어간 추출과 표제어 추출은 텍스트 데이터의 단어들을 기본 형태로 환원시키는 과정이며 데이터의 일관성을 높이고, 분석의 정확도를 개선 가능케 해줍니다. Khaiii를 사용하여 텍스트의 토큰화 및 형태소 분석을 수행하였으며. Khaiii는 한국어 텍스트의 형태소 분석을 통해 각 단어의 기본 형태를 식별하므로, 토큰화 과정에서 이미 어간 추출과 표제어 추출이 자연스럽게 이루어졌습니다. 이는 분석 과정에서 단어들을 그들의 어근 형태로 통일시키는 효과를 가지며, 별도의 어간 추출이나 표제어 추출 과정을 요구하지 않았습니다.(넘김)

4) 불용어 처리

중요한 키워드를 보다 명확하게 도출하기 위해 필요한 과정이고 의미없는 분석 결과를 방지해주기 위해 수행하였습니다. 사전에 정의된 불용어 사전을 만들어 텍스트 파일로 저장후 한글 깨짐 방짐을 위해 utf-8-sig 인코딩으로 읽어들여 토큰화 된 리뷰 텍스트에서 불용어를 제거하였습니다. 이를 통해 분석의 집중도를 높이고, 중요한 정보 추출에 집중할 수 있었습니다. (넘김)

4. 다음으로 전처리된 데이터를 가지고 수행한 LDA 토픽모델링을 설명드리겠습니다.

저희는 다양한 리뷰들 속에서 백종원의 프랜차이즈에 대한 공통된 토픽이나 주제를 식별하기 위해 LDA 모델링을 사용하였습니다. (넘김) LDA 모델을 활용해 앞에서 말씀드린 한식,양중,중식,일식,카페 총 5가지 카테고리에 맞는 키워드들이 무엇인지 한번에 알 수 있었습니다. 앞에 보이시는 것이 결과입니다. (넘김)

다음으로는 앞에서 뽑은 키워드들로 만든 워드클라우드입니다. WordCloud는 텍스트 데이터에서 주요 키워드를 시각적으로 돋보이게 표현하는 데 효과적인 도구로, 이를 통해 각 토픽의 핵심 단어와 그 중요도를 한눈에 파악할 수 있다고 생각하였습니다. (넘김)

그리고 plotly의 Network Graph는 토픽 간의 관계나 단어 간의 연관성을 시각화하는 데 유용하며. 복잡한 관계를 직관적으로 이해하고, 각 토픽 또는 단어가 어떻게 연결되어 있는지 탐색할 수 있다고 생각함. (넘김) LDA 모델 결과를 기반으로 토픽 간 또는 토픽 내 단어 간의 관계 구현. (넘김)

5. 결과 해석

마지막으로 요식업 분야의 전반적인 분석과 앞에서 보신 5개의 카테고리 중 중식 카테고리에 대해서 구체적으로 결과 해석을 설명드리겠습니다.

1) 먼저, 2020년부터 음식과 요식업에 대한 관심이 급증하고 있다는 점을 강조하고 싶습니다. 시계열 회귀식 분석에 따르면, 이러한 추세는 앞으로도 계속 증가할 것으로 예상되고 있습니다. 이는 요식업에 대한 전망이 비교적 밝다는 것을 의미합니다. 하지만 그래프에 보이시는 것처럼 2020과 2021사이에는 발전이 더딘 것을 알 수 있습니다. 이때는 코로나19 팬데믹이 전 세계를 강타하면서 요식업이 크게 위축되었습니다. 이 시기의 낮은 요식업 활동 빈도는 팬데믹의 직접적인 영향에서 비롯된 것으로 보입니다. 이는 요식업이 내부적인 성공 요인뿐만 아니라 외부적인 방해 요인에도 상당히 민감하다는 것을 시사하며 예측하기는 어렵다고 판단합니다.(넘김)

2) 5개의 카테고리의 데이터를 분석한 결과 백종원 하면 떠오르는 ‘가성비’ ‘양’ ‘맛’ 등의 단어들이 중복되어 도출됨을 알 수 있었습니다. 그래서 대표적으로 중식 카테고리에 대해 알아보겠습니다.

중식 요식업에서의 성공 요인과 그 변화를 살펴보면 중식 업계의 트렌드를 이해하는 데 도움이 됩니다. '맛'이 가장 중요한 요소로 부각되고 있으며, 특히 국물과 소스에 대한 관심이 높은 것으로 나타났습니다. 이는 중식 업계에서 고품질의 맛있는 국물과 소스를 제공하는 것이 성공에 중요하다는 것을 시사합니다. 또한, 짜장면, 짬뽕, 탕수육이 중식 메뉴 중 가장 중요한 것으로 나타났는데, 이는 이 세 가지 메뉴가 중식 시장에서 지배적인 위치를 차지하고 있음을 보여줍니다. 이러한 인사이트를 통해 중식 요식업 경영자들은 메뉴 및 맛의 품질을 높이고 이를 강조하는 마케팅 전략을 고려할 수 있을 것입니다. 가성비 또한 중요한 요인으로 드러났지만, 맛과 음식 종류에 비해 상대적으로 중요성이 낮은 것으로 나타났습니다. 이를 고려하여, 중식 요식업은 경쟁사보다 높은 품질의 음식을 제공하고 동시에 가격 대비 만족스러운 서비스를 유지하는 것이 중요할 것입니다. 이 아래로는 서비스와 접근성이 순위에 들어가있음을 알 수 있습니다. (넘김)

6. 마무리  
직접 크롤링을 통해 데이터를 수집하고 분석한 결과, 텍스트마이닝은 비지도 학습으로 주관적인 해석이 불가피하다고 느꼈습니다. 이를 해결하기 위해 과학적인 방법을 사용하여 타당성을 확보할 수 있다고 판단했습니다. 저희가 고려한 과학적인 방법은 도메인 지식, 양적 분석, 질적 분석이며 삼각 검증을 통해 상호보완적인 분석을 시행하였을 때 타당성이 확보된다고 생각하였습니다. 이러한 생각을 바탕으로 저희의 프로젝트는 요식업에 대한 도메인 지식을 적극 활용하였으며, 수치 기반의 빈도 분석과 비율 분석을 수행하여 양적 분석을 진행했습니다. 또한, 사람에 의한 문서 분석과 문서 관찰에 해당하는 질적 분석을 통해 부족한 부분을 보완하였습니다. (넘김)