* + 안녕하세요. 저희는 카페추천 웹사이트 제작을 한 카페레코 팀입니다.
  + 먼저 목차를 간단한게 설명해드리겠습니다. 우선 문제 정의에서 주제 선정 동기를 설명해드리고 해결 방안 설명 및 제시 후 계획 및 사용 방법에서 사용한 툴과 알고리즘에 대해 설명 드리고 그리고 실제 완성본을 보여드리는 것으로 발표를 마무리 할 예정입니다.
  + 우선 문제 정의 및 주제 선정 동기입니다.
  + 저희가 생각 했을 때 카페는 남녀노소 가리지 않고 누구에게나 친숙한 장소이며 높은 수요를 가지고 있다고 생각했습니다. 또한 현재 전국에는 카페가 10만개 이상 존재하는 상황입니다. 그만큼 다양한 개성과 특징을 가진 카페가 많기 때문에 내가 원하는 취향의 카페를 찾기 쉽게 돕는 사이트가 있다면 편리하고 실용성이 있을 것이라고 생각했습니다. 저희 팀은 이번 캡스톤을 계기로 해결하고 가고자 하는 카페를 쉽게 추천해줄 수 있는 사이트를 만들어보고자 하였습니다.
  + 그렇다면 카페를 추천하기 위해선 어떤 방식을 사용해 문제를 해결해야 할지 알아보겠습니다.
  + 혹시 어디를 방문하기 전에 리뷰를 참고해 갈지 말지 결정해보신 적이 있으신가요? 방문자들이 남긴 리뷰를 읽어보는 것만으로 그 카페에 관한 정보들을 쉽게 유추할 수 있기 때문에 저희는 상당히 유의미하게 이용할 수 있을 것이라고 판단했습니다.
  + 저희 팀은 카페 리뷰와 별점을 수집해서 빅데이터를 만들고자 하였습니다. 그 중에서도 구글에서 제공하는 리뷰 기능이 가장 양이 다양했다 판단해 구글 리뷰를 사용하여 진행하기로 하였습니다. 구글 리뷰의 크롤링을 통해 리뷰와 별점 데이터를 자동화하여 수집하고 그 데이터들을 필터링해 내용을 가공하고 유의미한 내용들만 남긴 후에 그 의미 있는 데이터들을 기반으로 제작을 진행할 예정입니다.
  + 저희는 문장에서 명사만을 추출하고 이를 통해 내용을 분석하는 것을 목표로 잡고 진행하였습니다. 리뷰의 문장들을 말뭉치로 만들고 그 말뭉치에서 명사들만 분리해 키워드로 나온 이상 그 단어는 카페와 관련된 이슈라고 파악할 수 있습니다. 앞에서 보여드린 것처럼 상당히 유사한 카페들도 비슷한 단어들이 출력되지만 한 쪽에선 빵과 관련된, 버터, 베이커리 등의 단어를 확인할 수 있습니다. 긍정 부정의 구분 없이 명사만 사용하는 것으로도 충분히 활용이 가능하다고 판단하였습니다. 예를 들어 짜다는 리뷰를 보고 누군가는 짠 음식을 선호하기 때문에 긍정적으로, 짠 음식을 싫어하기 때문에 부정적으로 받아들일 수 있습니다. 저희는 키워드를 제시하고 이에 대한 판단에 대한 근거는 개인에 맡기고자 하였습니다. 맛있다 내용이 많은 카페라고 모두에게 맛있을 수는 없지만 호평이 많다는 것은 참고할 수 있습니다. 키워드 자체의 자의적인 해석을 하는 것이 아닌 워드투벡터로 단어끼리의 연관성을 제공하는 것이 저희가 지향하던 방향이고 이에 집중하였습니다.
  + 다음은 저희 팀이 계획한 내용에 따라 실제 사용한 방식들을 설명드리겠습니다.
  + 데이터 크롤링 단계에서는 파이썬을 기반으로 제작한 웹 어플리케이션 자동화와 테스트 포터블 프레임워크 셀레니움과 웹드라이버를 사용해 자동화를 진행했습니다. 소모되는 시간을 줄이기 위해 각종 명령어를 사용하여 속도를 높여 최적화도 진행합니다.
  + 크롤링에 사용할 데이터는 소상공인진흥공단 상권 정보의 데이터를 사용해 불러와서 사용합니다. 그대로 사용하는 것이 아니라 이 중에서 커피/카페전문점만 남겨 데이터를 수집합니다.
  + 구글 리뷰를 수집하며 생기는 다양한 문제가 있는데, 먼저 스크롤을 추가로 내리지 않으면 밑의 리뷰들은 수집되지 않는다는 문제는 리뷰의 길이와 높이를 측정해 이 높이만큼 추가로 내리는 방식을 사용하여 수집하였습니다. 이 방식을 사용하여 처음 목표인 200개의 리뷰를 확인하고 정보를 모을 수 있었습니다.
  + 또, 수집한 정보는 append명령어로 합친 후 csv 형태로 내보냅니다.
  + 파이썬에서 실행해 크롤링을 진행하는 모습입니다.
  + 다음은 파이썬에서 실행 후 csv 형태로 저장된 크롤링 결과 자료입니다.
  + 다음은 자연어 처리 단계입니다. 이 과정은 colab으로 진행합니다. 앞서 수집한 크롤링한 데이터들의 csv 파일을 불러옵니다.
  + 추출한 리뷰 텍스트들을 konply의 한국어 처리 패키지 okt의 기능을 사용해 내용을 정제하고 문장에서 한글 단어들만 남기는 과정을 거칩니다. 패키지만으론 100% 온전하게 걸러내지 못하기 때문에 그 외의 사용하기 어려운 단어들, 모든 카페에서 등장하기 때문에 큰 의미가 없는 단어들을 사용하기 어렵다고 판단, 불용어로 선정해 추가로 제거하는 과정을 거칩니다.
  + 남아있는 단어들에 벡터값을 부여하기 전에 추천받을 때에 부가적으로 보여줄 테마를 분류하기 위해 추가로 필터링해줍니다. 테마는 크게 인원수, 시간대, 디저트, 분위기, 커피 산미도로 분류 했습니다. 분류의 기준이 되는 단어들은 카페 키워드 중 불용어를 제외하고 수집되는 단어들로 여러 단어 선정하였습니다. 여기서도 벡터값을 기준으로 테마를 분류하려고 했으나 벡터값을 적용하는 것보다 이렇게 진행하는 쪽이 결과가 정확하기 때문에 이 방식을 채택했습니다.
  + 위에서 분류한 테마의 데이터를 따로 보관하고 카페들의 키워드들에 벡터값을 부여합니다. 부여한 벡터값을 전부 더한 후 단어 수만큼 나눠 키워드의 평균 값을 구합니다. 나중에 사용할 ko.bin 파일의 벡터 크기와 맞춰 200으로 설정합니다. 그리고 아주 적은 경우지만, 만약에 카페 키워드에 유의미한 단어가 선정되지 않아 벡터에 아무 값도 들어가지 않을 경우 임의로 '평범'이라는 키워드를 넣어 평범한 카페라는 테마를 부여해줍니다.
  + html에서 파이썬 파일로 돌려볼 유사도를 확인해보겠습니다. 부여한 벡터값을 카페와 카페끼리 비교하고, 카페와 키워드끼리 비교해서 유사도가 출력이 되는지 확인합니다. 그리고 그 유사도가 얼마나 정확도가 높은지 예상과 비교하여 판단합니다. 오른쪽의 카페와 카페간의 비교 결과는 뒤에 자세히 설명드릴 예정입니다.
  + 여기에서도 같은 방식으로 카페와 키워든의 유사도를 비교해 진행해봤습니다. 유사도에 관한 자세한 부분은 뒤에 자세히 설명 드리겠습니다.
  + 간단한 테스트가 끝났다면 마지막으로 필터링을 거친 데이터 정보들을 csv 파일 형식으로 내보냅니다. 이 과정에서 등록된 정보는 이름, 카페의 테마, 벡터값입니다.
  + 다음은 앞서 만든 데이터들을 DB에 등재하는 과정입니다. 마이에스큐엘을 이용해 위에서 내보내기한 csv 파일을 업로드 합니다. 크로스 플랫폼 웹서버 xampp를 이용해 마이에스큐엘과 아파치를 연동합니다.
  + phpmyinfo를 사용해 db를 관리합니다. csv로 저장한 내용들이기 때문에 파일을 import 하는것은 간단하게 진행할 수 있었습니다.
  + 다음은 검색 페이지입니다. 메인 페이지에서 이동할 수 있습니다.
  + HTML에서 등재한 DB를 이용하기 위해선 PHP에서 서버 이름, 사용자 이름, 패스워드, DB 이름 등을 입력해 연동하였습니다.
  + 테이블 형태로 저장돼있는 DB를 불러오기 위해 해당하는 이름에 맞는 열을 불러와서 html을 호출하고 테이블 형태로 만듭니다.
  + 검색에 사용한 키워드는 caption에서 확인이 가능한데, 이 예시에서는 감성을 기준으로 검색해 나온 것입니다. 수집했을 때 크롤링한 별점들의 평균을 낸 값을 카페에 할당하고 이게 높은 순서대로 상위권에 나올 수 있게 내림차순하였습니다.
  + 다음은 프론트엔드의 제작 과정입니다.  테이블 형태로 저장돼있는 DB를 불러오기 위해 해당하는 이름에 맞는 열을 불러와서 html을 호출하고 테이블 형태로 만듭니다 0
  + 파이스크립트를 이용해 사용자가 입력한 키워드를 벡터값으로 변환시킵니다.
  + 사용자의 위치를 받아올 수 있게 하기 위해 구글 맵 API를 띄워줍니다.
  + 지금부터는 실제 완성본을 보여드릴 예정입니다.
  + 사용한 알고리즘 워드투벡터의 정확도와 이해를 돕기 위해 구체적인 예시를 사용하여 함께 설명 드리고자 합니다. 현재 보이는 두 사진은 왼쪽부터 조양방직, 카페 어니언이란 이름의 일반적인 카페가 아닌 예전에 운영하다 중단된 폐공장을 리폼해 카페로 이용하고 있는 곳으로 특이한 컨셉으로 많은 인기를 끌고 있습니다. 특징이 뚜렷해 그만큼 추출되는 단어들이 명확했고 구글 리뷰의 종류가 다양했기 때문에 예시를 드는데 사용하였습니다.
  + 왼쪽은 구글에서 캡쳐해온 별점과 구글에서 자체적으로 운영 중인 자주 나온 키워드들의 목록이고 오른쪽은 저희가 수집하고 가공한 결과입니다. 우선 조양방직의 검색 결과 볼거리, 공장, 소품, 감성 등의 키워드를 확인할 수 있습니다. 실제 저희 크롤링 결과에도 마찬가지로 볼거리, 공장, 소품, 감성의 키워드를 포함해 사진, 분위기, 추억, 레트로 등의 키워드들을 추가로 확인할 수 있습니다.
  + 다음은 카페 어니언의 검색 결과입니다. 앞서 말씀드린 카페와 공장이라는 교집합이 있는 만큼 키워드에서도 유사하게 공장, 분위기, 사진, 감성, 인테리어 등의 키워드가 비슷하게 도출되었습니다. 가장 눈에 띄는 차이점은 카페 어니언에선 가장 많이 출력된 키워드가 빵, 베이커리, 버터인데 이러한 키워드들은 카페 어니언에서 제과제빵 또한 취급한다는 점을 확인할 수 있습니다.
  + 다음 사진들은 페이드인, 봄날, 풍세커피의 실제 사진입니다. 해당 카페들은 별다른 기준 없이 선정했고 앞서 말씀드린 카페들과 유사도를 비교하는데 사용합니다.
  + 페이드인은 뷰, 책, 감성, 사진, 경치와 같은 단어 빈도수가 높았고 봄날은 사진, 뷰, 힐링, 케이크, 감성 풍세 커피는 힐링, 가을, 뷰, 당근 케이크 등을 확인할 수 있습니다. 이러한 키워드들로 카페들의 대략적인 느낌은 충분히 예측이 가능합니다.
  + 가장 먼저 소개시켜드린 조양방직을 기준으로 잡고 유사도를 비교해보면 비슷한 테마를 갖고 있고 실제로 비슷한 키워드들이 많이 도출된 카페 어니언과 유사도가 가장 높게 나왔고 그 외의 카페들은 상대적으로 낮은 것을 확인할 수 있습니다.
  + 금부터는 실제 사용 화면입니다. 우선 인천 소재의 카페에 아까 예시로 보여드린 조양방직 카페에 가장 많이 집계된 사진이란 단어를 검색하고 잘 있는진 찾아보았습니다. 제대로 검색이 되는 것을 확인할 수 있습니다.
  + 마찬가지로 많이 집계되고 다른 카페에서도 자주 언급되는 분위기라는 단어 또한 검색하고 조양방직을 찾아보았습니다. 마찬가지로 제대로 검색되고 있습니다.
  + 다음은 맛이라고 검색하고 동시에 평균 별점이 높은 카페들의 실제 리뷰입니다. 서울 소재의 카페에서 맛이라는 키워드를 검색했을 때 상위권에 위치한 카페들인만큼 실제로 구글 리뷰에서 검색해보면 맛이 좋다는 내용들의 리뷰를 쉽게 확인할 수 있습니다.
  + 마찬가지로 분위기를 검색했을 때 상위에 위치한 카페들의 실제 리뷰입니다. 분위기라는 키워드들이 많이 나왔던 카페들인만큼 실제 리뷰에서도 분위기라는 단어들이 즐비해있음을 확인할 수 있습니다.
  + 메인 화면에서는 로고 아래로 리뷰게시판, 로그인, 회원가입 창을 따로 빼두었습니다. 리뷰 게시판에서는 게시된 글들을 한꺼번에 모아볼 수 있으며 게시글 작성하기를 누르면 게시글을 작성할 수 있습니다. 로그인 페이지에서는 아이디와 비밀번호를 입력받을 수 있고 만약 정보가 없으면 회원가입을 진행할 수 있습니다.
  + 다음은 회원 정보와 추가로 지도를 삽입한 웹페이지의 구현 모습입니다. 아래로는 카페들이 테마 키워드와 함께 추천됩니다. 추천 받기를 누르면 검색창으로 넘어갑니다.
  + 실제로 이용하는 상황을 가정해 검색하는 영상을 보여드리고 저희 팀의 발표를 마무리하겠습니다. 우선 인천 소재의 카페입니다. 빵, 맛있는, 감성 순으로 검색해보았습니다. 입력한 키워드와 관련된 리뷰가 많았던 카페들을 확인할 수 있습니다.
  + 다음은 서울 소재의 카페들에 인테리어, 분위기, 볼거리 순으로 검색해보았습니다. 마찬가지로 다양한 카페들의 정보를 확인할 수 있습니다.
  + 이상으로 발표를 마치겠습니다. 감사합니다.

**워드 투 벡터를 이용해 제작한 추천 알고리즘 자연어 처리 프로젝트**

* 최근 카페의 입지는 세대를 가리지 않고 즐겨 찾는 정도의 접근성을 가지고 있으며 그 수요만큼 수도 10만 개가 넘어가고 있고, 그만큼 다양한 개성의 카페들이 많다. 단순 별점과 소개 블로그에 의존해 카페를 찾아가지 않고 내가 원하는 취향의 카페만을 추천해주는 사이트가 있다면 편리하고 실용성이 있을 것이라고 생각했기 때문에 사용자의 위치, 그에 걸맞는 키워드를 가진 카페를 추천하는 서비스를 구현하고자 한다.
* 구글 지도에서 사용자들이 남긴 리뷰들과 별점들의 빅데이터들을 좀 더 유의미하게 사용할 수 있도록 자연어 처리를 통해서 가공하여 이를 기반으로 이용하는 웹페이지를 제작하고자 하였다.
* HTML과 CSS를 이용해 웹페이지의 기본적인 틀과 디자인을 구성했고 Python과 Selenuim을 이용해 DB에 넣을 데이터들을 수집하고 내용들을 정제해 문장들에서 유의미한 단어들을 추출하고 wordcloud를 이용해 벡터화시켜 각 단어별의 연관성을 학습시켰다. mysql를 이용해 테이블에 정제한 데이터들을 집어 넣고 xampp를 이용해 사전에 제작해둔 웹페이지와 연결해서 홈페이지를 완성시켰다. 사용자가 이용하기 쉽게 원하는 키워드를 아무렇게나 검색해도 벡터화 된 값을 출력시킬 수 있게 하기 위해 wordcloud 알고리즘을 통해 입력한 내용과 최대한 유사한 키워드들을 갖고 있는 카페들을 출력할 수 있게 알고리즘을 완성시켰다.