sw응용 프로젝트

최종 발표 보고서

**11조 2017920009 김연수**

**2017920016 김지산**

**2017920060 주형진**

**1.시나리오 및 주요기능**

* **주요기능**
  + **사용자 유형**
    - 1. 전국 인프라 지수가 궁금한 사용자
      * 1-1. 자기가 살고 있는 곳의 인프라가 궁금한 사람
      * 1-2. 지역별 인프라 순위가 궁금한 사람

* 2. 새로운 거주지를 찾는 사용자
  + 2-1. 현실적으로 집값을 고려하여 나에게 가장 이상적인 거주지를 알고 싶은 사람

* 2-2. 새로운 곳을 결정할 기준이 필요한 사람
  + 2-2-1. 자신이 직접 결정
  + 2-2-2. 자신도 기준을 잘 모르는 사람

* 시각화(1-1, 1-2)
* 원하는 카테고리를 선택하면 해당 지수순으로 지역들을 정렬하여 테이블 형태로 시각화합니다

* 집값을 고려한 지역 추천 기능(2-1)
  + 선택한 지역을 입력하면 해당 지역과 유사한 인프라 지수를 가지는 지역 10곳을 집값 순으로 정렬하여 추천합니다.
  + 지역의 유사도는 각 지수를 맨하튼 거리로 계산하여 유사도가 높은 순으로 지역을 추천합니다.

* 나만의 행복도시 추천 (2-2-1)
  + 좌측의 슬라이드 바를 통해 사용자가 원하는 가중치를 결정하고 검색합니다.
  + 해당 가중치를 통해 계산된 지수를 바탕으로 정렬된 테이블을 출력하고 해당 테이블의 행을 누르면 인프라 지수를 막대 그래프형태로 출력합니다.

* 지역 월드컵(2-2-2)
  + 이상형 월드컵을 통해 사용자 맞춤의 인프라 가중치를 결정하는 과정입니다.
  + 각 도시는 사전에 8개로 클러스터링 된 군집 안에서 랜덤으로 하나씩 선택되어 서로 맞붙게 됨.
  + 각 라운드에서 선택된 도시의 지수를 기반으로 가중치가 합산되는 방식입니다.
  + 월드컵을 완료하면 계산된 가중치를 통해 나만의 행복도시 추천 페이지로 바로 넘어갈 수 있습니다.

**2.데이터 셋**

 저희프로젝트는 국회미래연구원에서 발표한 행복역량지수의 하위 7개 영역을 사용해서 진행되었습니다. 행복역량지수란 건강, 환경, 경제, 교육, 관계 및 사회참여, 여가의 7개 영역 하위 35개 지표로 구성된 데이터입니다. 한 사람이 누릴 수 있는 인프라에 대한 지수를 산출하기 위해 인구 n명당 혹은 단위 면적당으로 나눈 값들을 사용하는 것이 특징입니다.  자살률, 미세먼지 같은 부정적인 하위 지표는 역코딩 (1-표준화 수치) 되어 있습니다.

**3.어플리케이션 사용예시**

 메인 페이지에서 왼쪽에 메뉴에 내 인프라 선호도 알아보기, 나만의 행복도시, 집값기반 지역추천, 인프라가 있습니다.

 먼저 인프라 메뉴는 7가지 각 인프라의 소테카고리 까지의 전국 순위를 보여줍니다 사용자는 자신이 사는 지역의 순위를 보거나 추후에 집값 기반 지역추천에서 사용할 자신의 이상적인 도시를 찾아 볼 수 있습니다.

 나만의 행복도시 메뉴에서는 7가지 인프라 지수의 가중치를 사용자가 직접 설정하여 자신에게 맞는 도시들을 볼 수 있고 그 지역들의 인프라 지수를  차트를 통해 보여줍니다.

 내 인프라 선호도 알아보기 메뉴에서는 이상형 월드컵을 통해 자신이 선호하는 지역들의 인프라 특징을 종합하여 각 인프라에 대한 선호도를 알려줍니다 그리고 나만의 행복도시 메뉴에서와 같이 그 가중치 값을 입력하여 결과를 사용자에게 보여줍니다 .

 집값기반 지역추천에서는 사용자가 자신이 살고싶어하는 이상도시를 검색하면 그와 가장 비슷한 인프라 특징을 띄는 지역들을 집값 순으로 정렬합니다. 사용자는 추천받은 지역들 중에 자신이 원하는 집값을 가진 지역을 알아낼 수 있습니다.

**4.분석 및 평가**

* 클러스터링
  + 클러스터링을 통하여 유사한 지역끼리 8개의 군집으로 클러스터링한 결과 군집별로 특정한 인프라가 높고 낮은 것을 볼 수 있으며 1군집은 군 , 2군집은 주요 광역시, 3군집은 경기도,  4군집은 전라도,  6군집은 충청도와 경상도 7군집은 서울과 주요 광역시의 형태로 지리적으로도 인접한 지역들이 모여있는 것이 확인 되었습니다

* 내 인프라 선호도 알기에서 사용되는 이상형 월드컵에서 상반된 지역끼리 비교를 하게 하기 위해 인프라지수를 기반으로 지역들을 클러스터링을 하였는데  처음에 7가지의 인프라 지수를 그대로 사용하여 k-means 클러스터링을 사용한 결과 평균 실루엣 지수가 0.2이하였고 2차원으로 시각화 할 수 없었기에 먼저 차원축소 방법 중 PCA기법을 사용하여 주성분 2개를 추출하여 k-means 클러스터링을 하였습니다 군집 8개 기준 실루엣 지수가 0.36~0.38로 차원축소 전보다 클러스터링이 훨씬 잘 된 것을 알 수 있습니다 이에 더하여 차원축소 방법 중 manifold learning중 tsne기법을 사용하여 k-means 클러스터링을 하였더니 실루엣 지수가 0.40~0.42로 PCA기법보다 클러스터링이 더 잘 된 것을 알 수 있습니다  
   일반적으로 실루엣 지수가 0.5이상이어야 클러스터링이 잘 됬다고 볼 수 있지만 저희는 지역들을 여러 군집으로 군집화 하고 싶어 군집갯수를 8개로 두었고 인프라 지수가 정규화를 통해 계산된 것이므로 중앙값에 몰리는 것을 염두에 두면 뛰어난 클러스터링이라고 할 수 있습니다

* 상관계수
  + 행복역량 지수 7개 영역의 관련성을 살펴보기위해 각 영역간의 상관계수를 계산해본 결과 인프라 간의 관계중 몇 가지 특이한 관계를 찾을 수 있었습니다.

* 관계 및 사회참여 : 최소 한 두개 이상의 인프라와 관계를 가지는 다른 인프라들과 달리 유일하게 다른 인프라와의 상관계수가 거의 0에 가깝습니다.

* 행복역량 지수 : 건강, 경제, 안전, 관계 및 사회참여, 교육, 여가, 환경 순으로 행복 역량지수와 양의관계를 가졌습니다. 건강은 0.76으로 매우 큰 상관 계수를 가지고, 환경의 경우 -0.02의 상관계수를 가져 환경과 행복역량지수는 경향성에 있어서 유의미한 관계를 갖지 않음을 알 수 있습니다.

* 또한  카카오 “같이가치”에서 축적된 1,756,142명 이용자의 4,110,081건의 응답을 활용한 삶의 만족도 지수와 저희 데이터의 각 인프라 지수와의 상관계수를 계산해본 결과 환경, 여가, 안전, 사회, 건강, 교육, 경제 순으로 결과가 나타났고 이는 독특한 결과라고 생각했습니다. 국회미래연구원에서 국민행복지수를 계산할때  전문가들은  건강, 경제, 안전, 환경, 교육, 관계 및 사회참여, 여가순으로 가중치를 부여했지만, 실제 국민들이 말하는 삶의 만족도와는 조금 다른 순서를 보입니다.  
   이는 국회미래연구원의 국민행복지수를 기반으로 한 행복지도가 개인이 느끼는 행복과는 조금 거리가 있다는 점을 보여줍니다. 그렇기에 각 인프라별 가중치를 사용자의 기호에 맞게 직접 설정하여 사용자에게 맞는 행복도시를 추천해주는 저희 어플리케이션의 우수성을 보여줍니다

* 유사 어플리케이션과 비교
  + 저희와 유사한 어플리케이션을 찾던 중 통계지리정보 서비스에서 제공하는 ‘살고싶은 우리동네’ 기능을 발견했습니다. 이 서비스는 저희가 만든 어플리케이션과 유사하게 자연, 주택, 인구, 안전등의 카테고리를 기반으로 사용자에게 상/중/하 3가지로 이루어진 가중치를 입력받아서 추천을 해줍니다. 저희와 해당 서비스는 몇가지 차이점이 있습니다. 먼저 추천시 사용하는 카테고리의 수 입니다. 해당 서비스는 5가지의 인프라를 사용하지만 저희는 주택, 인구를 제외하고 7가지 카테고리를 제공하여 조금더 세부적인 것을 반영할 수 있습니다. 다음으로는 가중치입니다. 사용자는 자신이 어떤 인프라를 선호하지는지 모르는 경우가 많습니다. 선호한다 해도 그 정도를 나타내기가 쉽지 않기에 ‘살고싶은 우리동네’를 이용할 땐 처음에 정확한 추천이 이뤄지기 쉽지 않습니다. 반면 저희의 어플리케이션은 사용자의 맞춤 지수를 찾아주는 과정도 들어있고, 비중치를 조금 더 세부적으로 설정할 수 있는 장점이 있습니다.

* 추천 개인화 지수
  + 행이 케이스, 열을 각 지역으로 설정한 행렬을 0으로 초기화 하고 7개 인프라 가중치를 랜덤으로 설정한 2000개의 케이스를 시뮬레이션 하여 각 케이스마다 추천된 지역 상위 10개를 숫자를 0에서 1로 설정한 뒤 각 케이스별 코사인 유사도를 계산하였고 해당 값을 1에서 빼주어 지역들이 개인별로 얼마나 다양하게 추천되었는지 계산해 보았습니다 결과는 0.5정도로 생각보다 높은 수치가 나왔고 그렇게 다양한 추천이 이루어졌다고 말하긴 어렵습니다 그 이유를 분석해 보자면 7개의 인프라 지수를 종합한 행복역량 지수 순위에서 서울시 종로구가 압도적인 지수들을 가지고 있어서 시뮬레이션의 다수의 케이스에서 10위 내에 오르는 것과 그 외에도 각 인프라 지수들이 정규화를 통해 계산되었기 때문에 중간값에 몰리고 양 극값에서는 큰 차이를 보여 행복역량 지수가 높은 다수의 지역들이 결국 모든 케이스에서 순위가 높을 확률이 높아져버리는 점을 둘 수 있습니다