4차 5기 실무 Project

|  |
| --- |
| Issue What Show(이슈왓슈) |

**2019년 11월 4일**

**Open API를 활용한 빅데이터 전처리 시각화**

라떼는 말이야

박종선

김재현

이희수

문진한

**목 차**

**1. 프로젝트 개요**  **3**

1.1 프로젝트 기획 배경 및 목표 3

1.2 구성원 및 역할 3

1.3 프로젝트 추진 일정 4

**2. 프로젝트 현황**  **5**

2.1 시장 분석 5

2.2 경쟁 제품 장단점 분석 5

2.3 차별화 핵심 전략 기술 5

**3. 프로젝트 결과**  **6**

3.1 주요 기능 및 상세 동작 6

3.2 사용 기술 및 인프라 구축 8

3.3 핵심 알고리즘 8

3.4 회차별 멘토링 결과 10

**4. 기대 효과**  **10**

4.1 향후 개선 사항 10

4.2 기대 효과 11

**5. 개발 후기**  **11**

**6. 강사 및 멘토 의견**  **12**

**1. 프로젝트 개요**

1.1 프로젝트 기획 배경 및 목표

Open API 다양한 검색어, 키워드가 급상승하고, 급락 한다. 검색어를 보는 입장에서 알 수 있는 정보는 이 정보가 핫 검색어라는 사실 하나이다. 여기에 이제까진 교육과정에서 배운 내용을 토대로 검색어의 중요성, 맥락, 연관성을 분석하고 이를 시각화 한다.

목표는 첫번째 , 급상승 검색어를 가장 대표적인 4개의 웹사이트의 실시간 검색어를 실시간 현황을 통해 가져오고, 두번째 , 실시간 검색어 키워드의 중요성과 키워들 간의 연관성을 어떻게 분석 할 것인가에 대한 분석 방법 토론 , 세번째 , 이를 어떻게 시각화 할 것인가 , 네번째 이를 구현하는 것이다.

1.2 구성원 및 역할

|  |  |  |  |
| --- | --- | --- | --- |
| 이름 | 전공 | 역할 | 구현 부분 |
| 박종선 | 건축학과 | 팀장 | 프로젝트 관리  알고리즘 개발  UI 기획 및 개발  Display 개발 |
| 김재현 | 물리학과 | 팀원 | 데이터베이스 개발  데이터 분석  알고리즘 개발  서버개발  UI 기획 및 개발 |
| 이희수 | 환경공학과 | 팀원 | 알고리즘 개발  데이터 분석  UI 기획 및 개발 |
| 문진한 | 한방의료공학과 | 팀원 | 프로젝트 관리  데이터 분석  알고리즘 개발 |

1.3 프로젝트 추진 일정

|  |  |  |  |
| --- | --- | --- | --- |
| 구분 | 기간 | 활동 | 비고 |
| 사전  기획 | 9/23(월) ~  9/26(목) | 프로젝트 기획 및 팀 구성 |  |
| 9/27(금) | 프로젝트 멘토링 1차 | 현업멘토 참여 |
| 9/30(월) | 프로젝트 결정 | 잇슈왓슈 |
| PJT  수행  /  완료 | 9/30(월) ~  10/2(수) | 프로젝트 수행,  형태소 분석기 제작. |  |
| 10/4(금) | 깃 플로우 공부, 형태소 분석 |  |
| 10/7(월) ~  10/8(화) | 크롤러 제작 | 구글 크롤러 제외 |
| 10/10(목) ~  10/11(금) | 데이터 베이스 구조 확립 |  |
| 10/14(월) ~  10/17(목) | 김재현: 전처리 작업  박종선: 시각화  이희수: 통계적 분석  문진한: 통계적 분석 |  |
| 10/18(금) | 프로젝트 멘토링 2차 | 현업멘토 참여 |
| 10/21(월)  ~  10/22(화) | 김재현: RTF 알고리즘 개발  박종선: 3D-Force-Graph 개발  이희수,문진한: 통계 알고리즘 개발 |  |
| 10/23(수) ~  11/4(월) | 김재현, 박종선, 이희수: 시각화 고도화  ,문진한: PPT 제작,발표준비 |  |

**2. 프로젝트 현황**

2.1 시장분석 [소제목 : 12pt / 맑은 고딕]

현재 목표와 가장 비슷한 사이트의 이름은 'SOMETREND'로 해당 사이트의 경우 그래프와 워드 클라우드를 이용하여 시각화 하고 있다. 다른 이슈 검색 사이트의 경우 현 실시간 검색어를 서비스로 제공한다. 우리나라의 대표적인 두 사이트 네이버와 다음에서도 기본으로 실시간 검색어 기능이 포함되어 있다. 현재 시장에는 다양한 실시간 검색 현황 보여주는 사이트가 존재함을 확인 할수 있으며, 이에 대한 관심이 높음을 알 수 있다.

2.2 경쟁 제품 장단점 분석

네이버와 다음의 급상승 인기 검색어는 매우 강력한 트랜드를 보여주는 장치이다. 사람들이검색하는 검색어를 word count하여 순위를 정한다. 두 사이트는 한국의 최대 이용 사이트로써 검색어가 곧 현재의 트렌드라 해도 무방하다. 이를 통해 이용자는 현재의 이슈를 확인하는데, 중요한 점은 이용자의 검색어로 해당되는 것임으로 두 사이트의 성격에 따라 실시간 검색어가 다르다. SOMETREND의 경우 실시간 검색어 뿐만 아니라 검색어와 그에 대한 파생 연관어를 시각화(워드 클라우드, 그래프)를 통해 보여주지만, 검색어 사이의 연관도는 알 수가 없다.

2.3 차별화 핵심 전략 기술

현재 제공 되는 사이트는 주요 사이트의 핵심 키워드를 통해 검색어 끼리의 연관분석을 통해 통합적인 현 우리나라의 관심 현황을 3d-graph 시각화 라이브러리를 통해 보여주어 이용자가 트렌트 검색어 현황을 한 눈에 파악 하기 쉽다.

대형사이트 다음과 네이버의 실시간 검색어를 가져오고 이를 통해 뉴스 검색을 하여 현 우리나라의 전체적인 관심현황을 가져오고 이를 한눈에 볼 수 있도록 시각화 한다.

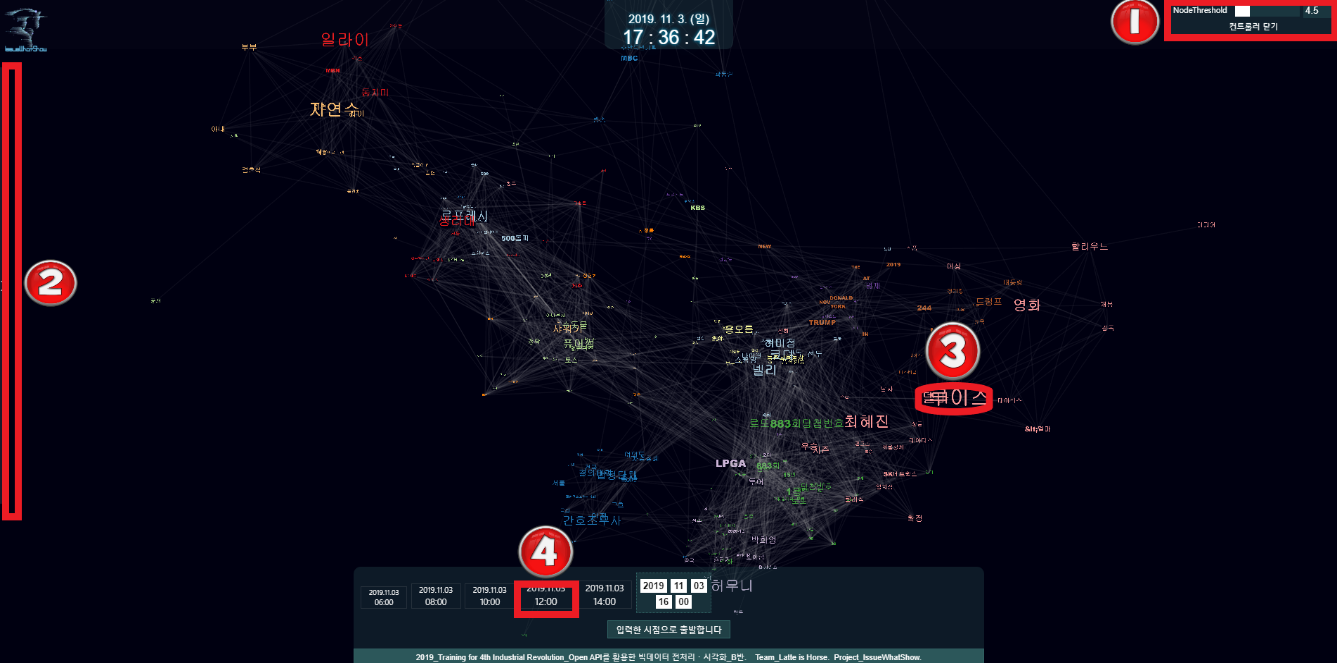
각 검색어를 클릭시 해당 검색어 관련 뉴스를 볼 수 있는 기능 추가를 하여 이슈 파악이 용이하다.

2시간의 텀을 두고 검색어를 가져오고, 분석하기 때문에 2시간 단위로 각 시간때의 핫 검색어를 파악 하기 쉽다.

이런 다양한 기능을 통해 종합적인 우리나라의 핫이슈 현황 파악을 시각화 한 것이 핵심 기술이다..

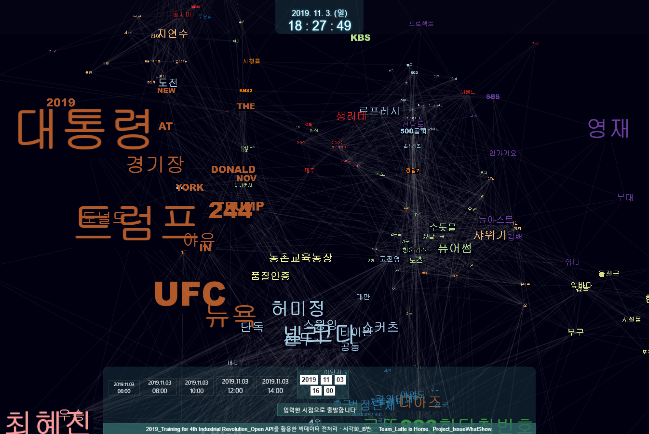
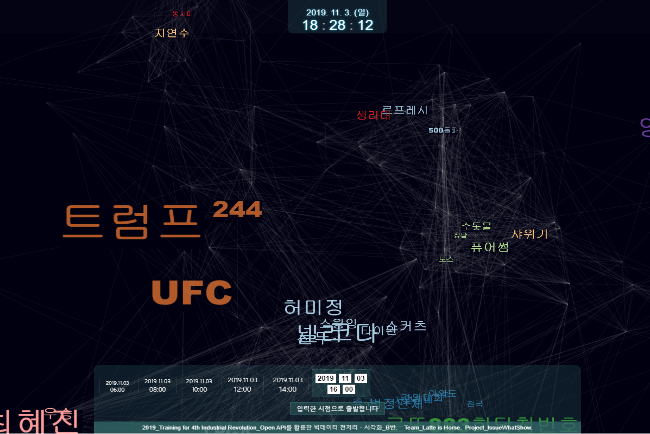
**3. 프로젝트 개발 결과**

3.1 주요 기능 및 상세 동작



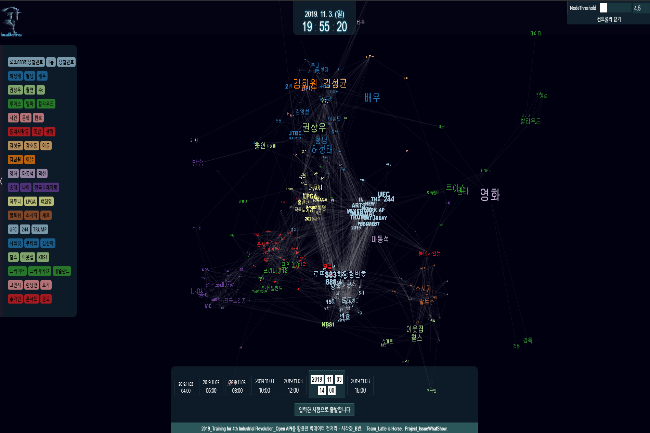
- 홈 화면은 전체적인 검색어가 노드화 되어 링크로 연결된 모양이다.(검색어는 글자로, 각 검색어는 거미줄처럼 연관되어 링크로 연관도 확인 할수 있다.) 상단 왼쪽은 팀 마크, 상단 중앙 현 시간 표현, 상단 오른쪽 각 검색어 노드 개수 지정 , 중앙 다양한 검색어 를 연관도와 함께 표시, 중앙 왼쪽 검색어 표시란, 중앙 아래 시간 별 검색어를 볼 수 있는 기능이 있다

 노드 개수 조절

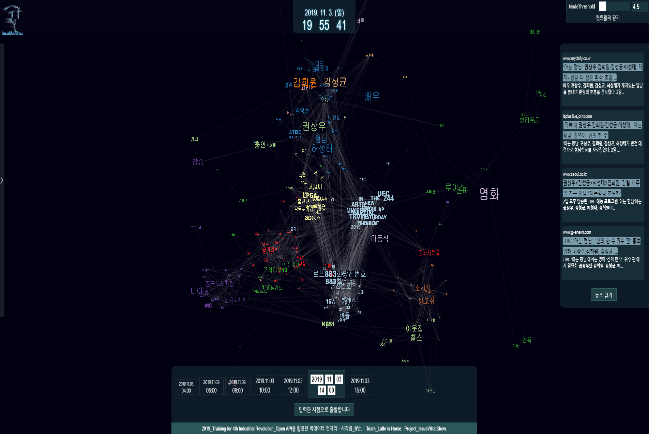
- 노드 컨트롤러 를 크게 할 수록 검색어 순위가 낮은(연관도가 낮은) 검색어 들이 사라지게 된다. 자유롭게 조절 가능하며 기본 값은 4.5로 이 값보다 커지면 검색어 순위가 높은 검색어만 남고4.5보다 낮으면 검색어 순위가 낮은 검색어도 보이게 된다.

 검색어를 한 눈에

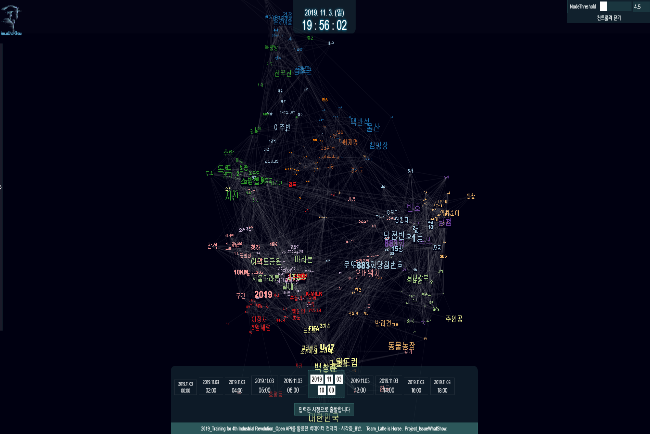
- 2. 왼쪽의 화살표 버튼 클릭시 숨어있던 테이블이 나타나며 실시간 대표 검색어가 첫번째 행에 표시가 되고 각 검색어의 연관 검색어 2가지가 연결되어 테이블의 형식으로 확인할 수 있다.

테이블에 표시된 검색어 클릭시 본문에 있는 해당 검색어를 가진 노드가 중앙으로 확대되어 나타난다.

 검색어 관련 뉴스 제공

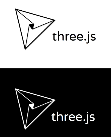
- 3. 기본 화면에 존재하는 다양한 노드클릭하면 해당 노드에 들어있는 검색어를 네이버 api를 통해 뉴스를 검색해 와 이를 4개까지 간략하게 뉴스 제목과 뉴스 내용을 확인 할 수 있다.

 시간에 따른 검색어 제공



- 4. 아랫쪽 하단에 시간 순서를 확인 할 수 있고 해당 시간으로 이동하면 시간여행을 하듯 그 시간때의 주요 검색어를 확인 할 수 있다. 하얀색 빈칸에 원하는 시간대를 작성 후 입력한 시점으로 이동 버튼을 클릭하면 해당 시간대와 가장 가까운 시간대로 이동하게 된다.

3.2 사용 기술 및 인프라 구축

  통해 web 서비스 구현. AWS 서버 구축



도에인 주소는 (<http://www.issuewhatshow.com>),

3.3 핵심 알고리즘

- 1. 형태소 분석기

Khaiii 사용 - khaiii는 카카오에서 개발한 세 번째 형태소 분석기로 CNN(Convolutional Neural Netword)기술을 적용해 학습한 형태소 분석기로 C++fh 구현하여 GPU없이도 비교적 빠르게 동작한다. 또한 총 10개의 분석기를 비교 분석 중 우리에 맞는 분석 능력 뿐만 아니라 속도 또한 2번째로 성능이 우수했다.

- 2. word2vec

= 중심단어(c)가 주어졌을 때 주변단어(o)가 나타날 확률

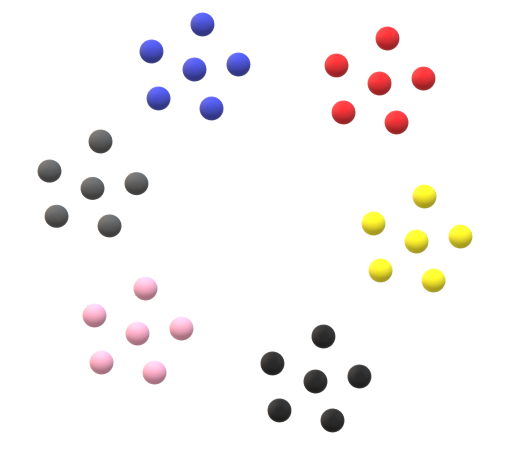
검색어의 연관도 분석을 위해 사용한 것으로 2013년 구글에서 개발한 방법론이다. 이름 그대로 단어(word)를 벡터(vector)로 바꿔주는 것으로 이로 인해 단어를 수치화 하여 계산 할 수 있게 된다. 단어를 넣기 위해서 기사를 형태소 분석기를 이용하여 전처리 과정을 거친 후 사용 하게 된다. 이를 통해 검색어 사이의 연관도를 분석하여 관련단어를 연관도 있게 시각화한다.

- 3.검색어 중요도 분석(TF-IDF)

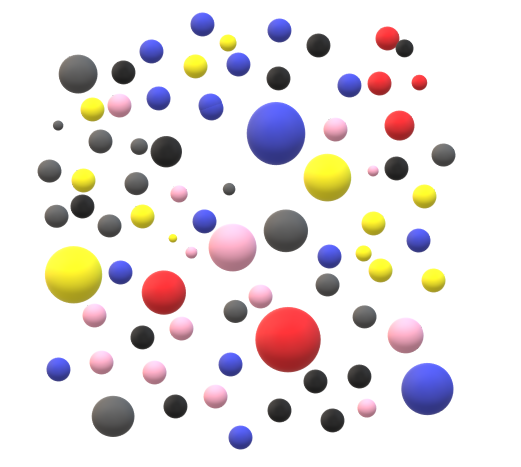
=> t

‘Term Frequency’ + ‘Inverse Document Frequency’가 결합된 단어 빈도-역문서 빈도 라 해석된다. TF-IDF는 문장에 사용 된 모든 단어에 점수를 부여하는 것으로 문서 내 비중있는 단어를 추출할 수 있게 한다. 기본적으로 ‘특정 단어의 중요도는 단어가 출현한 횟수에 비례하고, 그 단어가 언급된 모든 문서의 총수에 반비례한다’의 명제에 기초한다.(이말례’배환국,2002) 이를 이용하여 모든 뉴스의 단어의 중요도를 찾아 나열하여 현 시각 가장 의미있는 검색어를 찾는다.

- 4. RTF

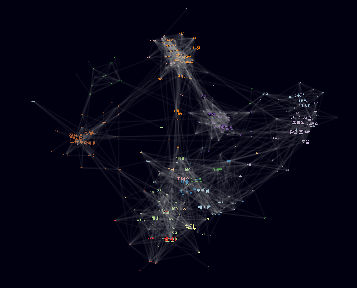
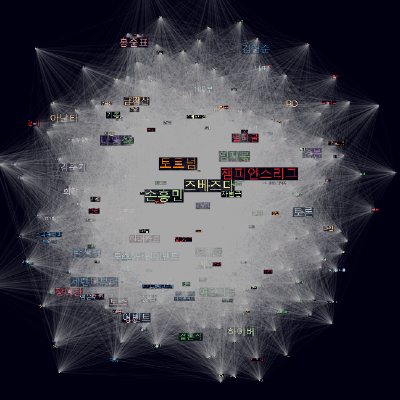
1번째 방법 (word2vec만 이용)

-Word2vec만 이용할 경우 각 키워드 별 뉴스 검색시 , 각 키워드 별로 연관 검색어 그림이 그려져 가운데를 중심으로 총 키워드(20개)별 연관 검색이 나타나게 된다. 오른쪽 예시는 6개의 키워드가 있다 가정하고 그린 것으로 키워드 별 연관관계는 표현되지 않게 된다. 나쁘지 않지만 키워드의 경우 연관되어 같이 실시간 검색이 되는 경우가 많아 이를 표현하지 못하며, 어느 것이 중요한 키워드인지 모른다.

2번째 방법(TF-IDF)

-어느 키워드가 전체 문서 내의 중요함을 나타내는지 표현하는데 아주 유용하다. 나타내면 오른쪽과 같이 나타나게 되는데 크기가 클수록 중요도가 높다. 하지만 이 방법은 같은 검색어 상의 연관검색어의 관계를 나타내지 못한다. 즉 그룹화 시키지 못한다.

3 번째 방법(word2vec + TF이용) 🡺 RTF



그리하여 분석을 통해 적절한 분석 방법을 발견해 냈는데, 이를 RTF라 명명하겠다. 먼저 각 키워드별 뉴스를 TF하는 것이 아닌 전체 키워드를 TF 을 통해 중요도 처리를 한다. 그럼 첫번째 그림과 같이 되는데, 이 상태에서 어떤 단어들 끼리 연관도가 높은지 알 수 없고 알 수 있는 것은 단순한 중요도(글씨 크기에 따라 중요도가 높다) 뿐이다. 이 상태에서 word2vec을 적용 다양한 검색어 중 연관도가 높은 검색어끼리 그룹화 하게 만들어 2번째 그림을 최종적으로 뽑아내었다. 실제로 우리가 원하는 키워들별 연관검색어 그룹과 연관검색어 끼리의 연관성을 알 수 있는 시각화 자료이다.

3.4회차별 멘토링 결과

|  |  |
| --- | --- |
| 회차 | 내용 |
| 1회차 | 우리 팀은 다양한 의견중 2가지의 주제, 내가 찾는 동네는 어디일까? 한눈에 보는 핫 이슈, 중 어떠한 것이 적절 할 지 멘토를 통해 정하기로 결정했다. 멘토링 결과 한눈에 보는 핫 이슈를 선정했고, 멘토 분들의 사용했으면 하는 기술, 시각화 방법 등을 추천받고 이를 추합하고 조원들끼리의 토의를 거쳐 프로젝트 시작. 추천받은 기술은 No SQL 이다. |
| 2회차 | 좋은 방향으로 가고 있음을 확인받고, 프로젝트시 추천받은 기술은 python pickle, , h5 , python 라이브러리 태그 바꾸면 알아서 가져오는, python 스케쥴러, APSScheduler 등 이다.  스케쥴러의 경우 try -catch절로 보완하여 문제없이 돌아가고 있으니 예상치 못한 문제로 인해 전체 프로세스가 멈추면 서비스에 문제가 가기 때문에 커널에 의존적인 외부 모델이긴 하지만 안정적으로 정해진 시간에 다시 새로 실행시켜주는 크론탭을 사용, 알 수 없는 오류가 발생 하더라도 정상적으로 동작하도록 했고, No SQL 의 경우 적용하고 싶어 proto type으로 먼저 RDBM로 구현후 추후 변경, 계획 세웠으나 일정이 좀 어긋나 적용하지 못해 추후 업데이트 사항에 들어갈 예정이다. |

**4. 기대 효과**

4.1 향후 개선 사항

- 1. 다음 키워드 10개, 네이버 키워드 10개를 사용한 분석 방법으로 이를 특징화 하여 나타내었다. 이를 확대하여 다음, 네이버, 구글, 트위터, 4개의 대형 실시간 검색어와 그 분석을 따로 보여주고 이용자는 각 사이트별 성격을 한눈에 파악 할 수 있게 한다.

- 2. 시계열 분석 자료를 넣어 각 검색어 별 시간별 변화 추이를 살 펴 보는 기능을 넣으려 한다. 이를 통해 검색어의 동향 파악을 할 수 있게 된다. 예를 들어 사건이 생기게 되면 시간에 따라 그 검색 양상이 달라진다. 경찰 관련 사건이라면 처음에는 사건 발생, 두번쨰는 과정, 세번째는 결과이다. 이를 검색어와 시계열 분석을 통해 시각화한다. 검색어의 변화양상으로 그 흐름 파악이 가능해지는 것이다.

- 3. 지금은 2시간 단위로 보여주지만 월별, 일별, 계절 등으로 보여주는 기능. 긴 기간 유지되는 키워드 분석을 통하여 시국의 변화 양상을 확인하는 서비스를 제공하려 한다.

-4 . 다양한 기기(모바일, 패드,탭 등)에서 적용 가능하도록 최적화 한다.

4.2 기대 효과

정보화 시대로 다양한 정보의 홍수 속에서 빅데이터라는 단어가 나오는 세상이 되었다. 이 와중 우리는 실시간 검색어를 통해 간략하게 현재의 이슈를 알고 있지만 이는 일부에 지나지 않는다. 실시간 검색어는 매 시간 바뀌고 있으며 이는 잠시 한눈을 팔면 지나갈 물에 지나지 않는다. 또한 사이트별로 모이는 사람의 성향에 따라 검색어도 다르기 때문에 전체적인 이슈를 한눈에 그리는 것은 사실상 불가능 하다. 잇슈왓슈는 그런 의미에서 이용자들에게 전체적인 나라의 이슈를 보여주며, 지나간 자료 또한 제공하여 이용자에게 현재의 트랜드의 변화 양상을 확인 할 수 있고, 한눈에 보여주기 때문에 아침, 점심, 저녁, 어느 시간이든 클릭 한번과 길지 않은 시간으로 충분히 전체적인 맥락 확인이 가능하다. 또한 정말 궁금 한 것이 생긴다면 이를 자세하게 찾을 계기를 마련해 주어 뉴스도 보기 힘든 바쁜 현대인들에게 보여주는 이슈서비스라 할 수 있다.

**5. 개발 후기**

|  |  |  |
| --- | --- | --- |
| 테이블, 노트북, 앉아있는, 컵이(가) 표시된 사진  자동 생성된 설명 | | |
| 성명 | 후기 |
| 김재현 | IT 업계를 두 번이나 탈주했던 경험이 있습니다. (눈물 없이 못 들어 중략……).  이러한 와중에 멀티캠퍼스에서 훌륭한 선생님의 지도하에 열정 넘치는 청년 분들과 함께 배움의 기회를 얻게 되었고, 특히 이번 프로젝트 기간 동안 팀원들에게서 까맣게 잊고 살았던 열정과 도전이라는 멋진 단어들을 새로이 선물로 받게 된 것 같습니다. 뜬금없이, 왠지 다시 한 번 달려볼 수 있을 것 같습니다.  고맙습니다. 모두! 정말! |
| 문진한 | 이번 프로젝트를 통해서 데이터 분석가를 조금이나마 경험을 해 본 것 같아 더욱 뜻 깊은 것 같습니다. 데이터 분석가는 데이터를 다루기 위한 기술이 필요하고, 전처리를 한 데이터를 가지고 통계적인 접근으로 인사이트를 주는 직업입니다. 마치 모든 틀이 정해져 있고, 데이터만 잘 처리해서 넣으면 나올 수 있다는 생각은 버리게 되었습니다. 학원에서 진행한 프로젝트였지만, 어떤 통계 방식을 선정하고, 어떻게 데이터를 수집하는 방식에 대해서 깊이 생각을 하면서 앞으로 컴퓨터 언어와 통계 분석에 초점을 두고 공부를 해야겠다는 생각이 들었습니다. |
| 박종선 | 이번 프로젝트는 실무에서 하듯 분업화를 경험하였고 또한 서비스를 실제로 상용화하며 평소 잘 몰랐던 서버구성의 전체적인 흐름을 배웠습니다,  그리고 가장 중요한 것은, 교육수료 후 내가 무엇을 해야 좋을지 좋은 가이드가 되었습니다.  앞으로는 React나 Vue.js도 공부해볼까 합니다. |
| 이희수 | 다양한 분야의 사람들이 모여 한 프로젝트를 하며 배운점이 많았습니다. 같은 주제여도 각자가 생각하는 가 달라 ‘저런 생각을?!’ 이란 생각이 드는 다양한 의견이 있어 이를 추합하는 과정이 너무 재미있었고, 협동하는 일이니 만큼 깃과 깃플로우를 이용해 코드와 일정을 정리하고 공유하는 과정에서 정말 한 줄 한 줄 내가 이 팀과 일하고 있구나 실감했습니다. 교육받은 과정을 거의 다 넣은 활동이라 이를 활용할 수 있었던 주제를 뽑아 만족스러웠고, 각각의 장점이 달라 이를 서로서로 배우며 지내는 과정에 이것이 협업이구나, 코딩은 진짜 협업을 해야하는 구나를 꺠우치는 과정이었습니다. |

**6. 강사 의견**

|  |  |  |
| --- | --- | --- |
| 평 가 요 소 | 배점 | 평 |
| 1. 아이디어 :  유사한 서비스 존재 유무 및 체계성 | /20 |  |
| 2. 개발 :  실제 구현 정도 및 배포 유무, 코드의 무결성 및 난이도, 현업적용도, 실무기술 반영정도 | /30 |  |
| 3. PJT 수행력 :  일정관리 및 역할분담, 목표 일정 달성도, 팀내 참여도 등 | /30 |  |
| 4. 준비도 :  프리젠테이션 및 프로젝트 준비 정도 | /20 |  |
| 계 | /100 | 강사 의견 필수 |