에이유디 사회적협동조합에서 청각장애인분들을 위한 문자통역 서비스를 제공하고 있습니다.
-시작 전에 안내말씀을 드리겠습니다. 저희가 피티를 따로 업로드를 못했는데 왼쪽 클릭하시면 PC나 스마트폰에서 볼 수 있습니다. 위에 있는 링크를 잘 틀리지 않고 잘하면 PT를 받을 수 있고요.
그리고 저희가 안내에 보면 깃허브 주소 안내해드렸어요. 홈페이지 링크 해드렸으니까 지금 가능하시면 PT 파일을 같이 보시면 좋을 것 같고요.
연결이 안 되면 이 화면을 봐주시면 됩니다. 실제로 깃허브 가면 가지고 있는 PC에서도 데이터 분석을 실시간 할 수 있거든요. 같이 해보실 분들은 안내를 해드릴 텐데 두 가지고 같이 해주시면 좋을 것 같아요.
-혹시 주소나 사용하는 자료 같은 것들 이후에 홈페이지에서 공유를 해주시는 건아요?
-깃허브 홈페이지에 계속 올려놓을 겁니다.
-알겠습니다.
-깃허브 페이지에 링크를 올려두는 게 좋을 것 같아요.
-네.
오늘 오신 분들 파이썬이나 데이터 분석 많이 해보신 분 계신가요? 저희가 수준을 어떻게 소개해드릴지 몰라서 굉장히 쉽게 설명을 드릴까 하는데 괜찮으실까요?
-네.
-본인이 조금 능력이 있으시다 하시는 분은 안 들으셔도 되고요. 저희는 데이터 분석을 어떻게 할 수 있는지 크롤링 안내해드리고 파이썬 이용해서 많이 해보자라는 식으로 쉽게 설명해 드리도록 하겠습니다.
이제 시작하면 될까요?
인터넷으로도 생중계 되는 것으로 알고 있는데요. 제가 인터넷 주소도 많이 홍보를 해서 많이 들어오실지 모르겠는데 지금부터 시작하도록 하겠습니다.
먼저 인사드리겠습니다. 오픈넷에서 활동하고 있는 박지환 변호사라고 합니다.
(박수)
오늘 날씨도 굉장히 덥고 오전 세션 들으면서 힘드셨을 텐데 점심 끝나고 하는 세션이 항상 힘들더라고요. 많이 와주셔서 감사드리고요.
저는 주로 최근에 오픈데이터 관련된 활동을 많이 하고 있습니다. 직업은 변호사인데 IT 관련이 많아서 데이터 쪽 활동을 많이 하고 있고요. KrIGF 협회에서도 데이터 주제로 한 세션이 진행이 됐었는데 올해는 조금 더 구체적으로 데이터를 놓고 분석하는 거를 설명해 드리면 좋을 것 같아서 그동안에 정책 이슈나 큰 틀에서 말씀을 많이 드렸는데 오늘은 굉장히 유명한 개발자분을 모시고 여러분들은 굉장히 행운이십니다. 굉장히 비싼 강의료를 받고 강의를 하시는 분이신데요.
오늘 제가 어렵게 자리 모셨는데 먼저 큰 박수 부탁드립니다.
오늘 간단히 제가 설명드리고 바로 들어가도록 하겠습니다. 제목을 국민제한 플랫폼 데이터 분석해 보자 취지로 데이터 분석을 활성화하자는 측면에서 소개를 드릴까 하는데요. 분석하려고 하는 사이트는 정부혁신1번가라고 하는 여러분들은 잘 모르실 거예요. 정부혁신 관련된 활동을 하고 있는데 혁신적인 정책을 육성하는 제안상태입니다. 청와대 국민청원 만큼 활성화 되지는 않고 시민분들이 제안을 해주시는 이런 게 있어요.
예를 들면 시민분들이 직접 의견을 제안해 주신 이런 것도 있습니다. 제안내용도 볼 수 있

고 이런 것들이 데이터 분석 하기에 쉽지 않은 형태로 올라와 있어서 이것을 어떻게 보면 데이터 분석 하기 쉬운 형태로 만들어볼까 하는 게 첫 번째 저희 미션이고요. 모아놓은 데이터를 어떻게 활용할 수 있을지에 대한 아이디어를 여러분께는 소개해드릴 예정입니다.
시간이 되면 스토리얼 끝나고 관련된 법률 이슈가 조금 있기는 한데 시간이 되면 법률이슈까지 다루고 더 듣고 싶다고 하시면 스토리얼을 좀더 많이 하도록 하겠습니다. 박조은 님께 마이크 넘기겠습니다. 큰 박수 부탁드립니다.
-안녕하세요? 저는 박조은이라고 하고요. 여기 국민참여 플랫폼 데이터를 중심으로라고 되어 있는데 깃허브 들어가서 보게 되면 깃허브 하단에 발표자료라고 해서 링크가 있어요. 이 링크로 바로 들어오면 여기 있는 슬라이드를 바로 열어보실 수 있으세요.
슬라이드를 바로 열어보실 수가 있고요. 그리고 여기에 깃허브 링크라든가 클로링 링크가 같이 있어요. 깃허브만 들어오셔도 대부분 볼 수 있을 거예요. 저는 주로 개발자 활동을 하고 있고 데이터 분석을 하고 있어요. 같이 병행을 하고 있어요.
지금 소개해 주신 것과 마찬가지로 그것처럼 다양한 국민참여 플랫폼이 있어요. 대표적으로 청와대 국민청원 사이트가 있는데 작년에 청와대 국민청원 사이트를 분석해서 튜토리얼로 진행을 했어요. 파이콘이라는 곳에서. 청와대국민청원 활성화가 되어 있어서 튜토리얼 20, 30만 건 정도가 있었어요. 공개된 텍스트 데이터가 국내에 그렇게 많지 않아요. 텍스트 분석하시는 분들이 가장 필요하신 게 한국어 텍스트예요. 그런데 한국어 텍스트가 저작권이 자유롭게 쓸 수 있는 텍스트가 많지 않아서 분석을 하실 때 뉴스기사를 분석하신다거나 댓글 같은 거를 분석을 하시는데 그런 것들을 오픈소스 공개해서 다른 사람들하고 공유를 하면 좋을 텐데 저작권 문제 때문에 공유하기 어렵다는 게 애로사항이었어요. 저작권 정책에 대해서 변호사님께 문의를 드렸고 저작권들이 자유롭게 연구에 활용될 수 있도록 개방이 되었으면 좋겠다는 생각을 가지고 있어요.
그리고 광화문1번가 사이트 들어가서 보게 되면 여러 가지 제안이라든지 국민들이 올려놓은 데이터들이 많이 있는데 텍스트 데이터가 굉장히 많이 있어요. 여기 국민청원 데이터와 광화문1번가 데이터가 좋은 이유 중 하나가 현실에서 우리가 직접 필요로 하고 있는 그런 데이터들 그다음 현재 이슈 내용들이 굉장히 많이 올라와 있어요. 청와대 국민청원만 분석을 해보더라도 그 당시에 어떤 일들이 사회적으로 이슈가 되었는지 그다음에 국민들이 어떤 이슈에 많이 관여를 했었는지 투표수라든지 청원에 국민이 투표를 하게 되잖아요. 투표 수 분석해 보는 게 상당히 의미가 있었고 그리고 그렇게 청원에서 투표를 많이 받은 건들을 보면 그 당시 뉴스 기사를 같이 분석해 보면 뉴스기사도 민감하게 반응했던 이슈들을 분석해 볼 수 있어요.
광화문 1번가 정부혁신 1번가 사이트가 있는데 국민청원는 사이트와는 다르게 정책제안할 수 있는 사이트예요. 국민청원과는 다르게 데이터가 아직까지 많이 올라와 있지 않아요. 정책 제안이기 때문에 정부혁신1번가 사이트에 내가 원하는 정책을 제안할 수 있고 그게 정책이 되도록 투표를 많이 받으면 정책이 될 수도 있는데요.
이 정부혁신1번가 사이트에 있는 거를 국민청원 사이트처럼 분석을 해봐야겠다 생각을 해서 정부혁신1번가 데이터를 수집해서 분석을 했어요. 분석을 하려는 가장 큰 이유는 자연어처리 등의 연구형 텍스트가 부족한 편이에요. 해외 사례 보면 영화 리뷰라든지 국내에는 네이버 영화 리뷰 데이터고 공개가 되어 있기는 한데 사실 저작권적인 측면에서 봤을 때 영화 리뷰 쓰시는 분들에게 저작권이 있어요. 엄밀하게 따지자면 저작권 정책으로 봤을 때는 저작권 자유롭게 사용하기는 데이터이기도 해요. 국민청원 데이터도 2차 제작한다든지 상업적

으로 이용하는 게는 금지되어 있어요. 연구용으로 공개가 되었으면 좋겠다 의견을 갖고 있어요. 청와대 국민청원.
텍스트 데이터 연구를 하다 보는 어떤 정책들이 크게 이슈가 되고 있고 정책 제안으로 제안을 해볼 수 있겠다는 생각이 들어요.
그래서 지난주에 디지털혁신파크에서 제가 소개해드렸던 정부혁신 1번가 분석해 보는 해커톤을 진행을 했었어요. 다른 사이트를 개선하는 해커톤도 있었는데 데이터해커톤에서도 데이터만 크롤링 해서 함께 분석하는 작업을 했었어요.
개포디지털혁신파크에서 크롤링하고 분석하는 시간을 가졌어요. 저는 크롤링할 때 파이썬을 사용을 해요. 파이썬 외에도 다른 것들을 많이 사용하시기도 하는데 저는 주로 파이썬을 사용하고 있어요. 제가 파이썬 사용하는 이용 중 하나는 설문조사 했던 결과예요. 10년 전만 하더라도 동영상으로 되어 있거든요. 10년만 하더라도 대안언어에 가까웠어요. 10년 전에 파이썬이라는 언어가 등장했을 때 대안언어라고 했어요. 주류 언어가 아닌 거죠.
10년 전에는 파이썬 펄 무비 이런 프로그래밍 언어가 등장했었고 프로그래밍 언어가 당시만 하더라도 대안언어에 가까웠는데 지금은 거의 파이썬이 주류언어로 등극을 했어요. 같이 인기가 있었던 펄이라든지 무비는 이렇게 파이썬 만큼 인기를 누리지는 못했거든요. 파이썬이 이렇게 인기를 누리게 된 데에는 다양한 상태계, 커뮤니티 힘이 굉장히 컸어요. 다양한 생태계라고 하면 파이썬을 배우게 되면 웹개발도 할 수 있고 분석하는 것도 할 수 있고 데이터 분석 그다음에 요즘 머신러닝 딥러닝 이런 것들 많이 이슈 관심을 갖고 있는데 활용할 수도 있고 게임개발에 활용할 수 있고 다양하게 활용할 수 있는 언어라고 보시면 돼요.
10년 전에 비해서 파이썬이 주류 언어로 성장할 수 있었던 데는 커뮤니티 힘이 컸어요. 생태계. 많은 사람들이 굉장히 많이 쓰고 있다는 거죠. 다른 언어에 비교해서. 파이썬 쓰다가 뭔가 문제가 생겼어요. 궁금해서 이거를 인터넷에 물어봤어요. 그러면 파이썬을 사용하고 있는 사용자가 많기 때문에 내가 겪고 있었던 어려움에 대해서 답변을 해줄 수 있는 사람도 없고 이미 내가 어려움을 겪었던 문제를 다른 사람들도 어려움을 겪고 질문을 올리고 많은 사람들이 유사한 답변을 달아줬어요. 큰 커뮤니티에서 도움을 받으며 사용할 수 있는 언어가 파이썬이라고 보시면 됩니다.
파이썬이라는 언어는 다른 언어에 비교해서 배우기 쉬워요. 다양한 프로그래밍 언어로 개발자로 사용한 지 오래 되었는데 다른 언어에 비교해서 파이썬은 비교적 배우기 쉬운 언어라고 볼 수 있을 거예요.
파이썬의 장점은 다양한 생태계라고 말씀을 드렸어요. 그래서 데이터 분석만 하더라도 파이썬 사용하게 되면 판다스라든지 쥬피스 테렉스 이거 외에도 다양한 라이브러리와 패키지가 존재를 해요. 파이썬으로 데이터 분석한다라고 하면 이 정도 패키지 사용하면 대부분 원만한 데이터 분석이 가능한데요.
하나씩 소개해드리도록 할게요. 혹시 여기 파이썬 사용해 보신 적 있으세요? 두 분 계시네요. 파이썬 비교적 쉽다고 말씀을 드렸고 생태계 다양하다고 말씀을 드렸어요. 왼쪽에 보면 아나콘다라는 게 있어요. 아나콘다라는 거는 이름에서 유추할 수 있듯이 왼쪽에 로고가 뱀처럼 생겼는데 파이썬도 비단뱀을 의미하기도 해요 원래 파이썬이라는 프로그래밍 언어를 이름을 지을 때 비단뱀하고는 상관없이 이름을 짓기는 했어요. 그런데 아나콘다라는 거는 뱀 굉장히 큰 뱀을 의미를 하는데 아나콘다라는 뱀은 사람도 집어삼킬 정도로 큰 뱀이죠.
아나콘다 툴에는 라이브러리 중에 상당수가 아나콘다 안에 포함이 되어 있어요. 아나콘다 하나 설치하게 되면 파이썬도 설치되고 판타스 설치되고 이런 것들이 같이 함께 설치가 되

고 그다음 쥬피터라고 되어 있는데 함께 설치가 돼요. 쥬피터라는 거는 파이썬을 쓸 수 있는 편집기라고 보면 돼요. 프로그래밍 언어를 배울 때 어떤 편집기로 프로그래밍 언어를 배울지 선택을 하게 되는데 아나콘다 설치하게 되면 쥬피터 노트북까지 설치가 되기 때문에 클릭하고 열어서 바로 실습할 수 있어요. 제가 이따가 잠시 실습하는 걸 보여드릴 텐데 여기 보시면 이거를 구글 콜레보레이터 노트북인데 쥬피터 노트북을 온라인상에서 쓸 수 있도록 만들어놓은 툴이다라고 보시면 될 거예요.
그래서 쥬피터 노트북 설치되어 있기 때문에 처음에 프로그래밍 언어 배울 때 가장 어려워하시는 부분이 툴을 설치하고 호환성을 맞춘다든지 프로그래밍 배우는 것보다 더 많은 시간 쏟게 돼요. 아나콘다를 설치하게 되면 이런 버전의 의존성이라든지 설치문제 이런 것에 조금 더 자유로워질 수 있다는 게 아나콘다의 장점이 될 거예요.
그리고 이 위에 왼쪽 위에 보면 판다스라는 게 있어요. 보통 인터넷에서 검색을 해보면 팬더곰이 이미지 많이 뜨는데 팬더곰과 상관없고 판다스는 패널데이터시스템 약자라고 해요 판다스 로고 밑에 보면 이런 수학공식 같은 게 적혀 있어요. 수학공식으로 봤을 때 판다스 툴을 통해서 수치 계산을 해볼 수 있겠구나 유추해 볼 수 있고 오른쪽에 보면 그래프 3개가 있는데 막내그래프, 선그래프, 이런 것들이 있어요.
그래서 숫자를 그래프로 표현하는 툴이겠다는 생각이 들어요. 이 판다스라는 툴은 파이썬에서 쓸 수 있는 엑셀이라고 보시면 돼요. 제가 굉장히 다양한 연령대에 다양한 분들에게 파이썬 강의를 오랫동안 해왔는데요. 파이썬 처음 배우시는 분들에게 제가 다른 건 몰라도 판다스는 꼭 배워두시면 좋아요라고 많이 말씀을 드리는 편이에요. 왜냐하면 직군에 상관없이 어떤 분야에서 연구를 하고 있건 어떠한 일을 하고 있건 간에 엑셀 프로그램을 대부분 많이 사용을 하시잖아요. 판다스는 파이썬에서 쓸 수 있는 엑셀이라고 보시면 돼요.
엑셀에서 100만 줄 데이터 불러왔던 적 경험 있나요? 분석해 봤다 하시는 분 계신가요? 어떠셨어요?
-일단 주식시장 데이터를 긁어왔었는데 판다스 문법들은 하나도 모르면 진짜 오래 걸리더라고요.
-판다스는 말고 엑셀만을 사용을 해서.
-컴퓨터가 다운.
-컴퓨터가 다운이 돼요. 그렇죠. 100만 줄 이상의 파일을 엑셀로 로드를 하려면 컴퓨터가 다운이 되거나 아니면 엑셀 시스템이 멈춰버려요. 로드가 되다가 멈춰버리게 되거든요. 그런 데이터들 엑셀로 열리지 않으면 분석이 그때부터 불가능하게 되잖아요. 그런 파일을 분석하려고 하면 판다스를 가져다가 쓰시면 100만 줄 이상의 데이터도 훨씬 빠른 시간에 분석해 보실 수 있어요. 판다스를 혹시 사용해 보신 거예요?
그래서 저는 엑셀을 사용하시는 분들은 판다스 배워두시면 엑셀을 훨씬 더 자유롭게 사용할 수 있다고 많이 말씀드리는 편이에요. 그래서 저에게 들어왔던 일 중 하나가 정산을 하는데 매달 회사에서 시스템에서 데이터를 내려주는 거예요. 그러면 내려준 엑셀 파일을 가지고 매달 회계 담당자가 정산을 하게 되는데 그 정산을 할 때마다 고객사라든지 아니면 우리 회사 시스템 정책이라든지 이런 거에 따라서 정산이 주는 룰이 다 달라요. 그러면 엑셀파일로 정산을 한다고 했을 때 이틀이나 삼일 정도 정산 시즌에 매월 회계 담당자가 야근을 한다든지 밤을 샌다든지 반복하게 된다는 거죠.
엑셀 파일을 불러와요. 정산해 줘야 하는 조건을 만들어주고 매크로 같은 거를 짜놓은 게 있다 매크로 새로 적용해 주고 시트마다 다 적용을 해서 시트를 취합을 해서 공식을 적용해

주고 이런 작업들을 하는 데 3일이 걸리는 거예요.
이런 작업들을 여기 쥬피터 노트북이라는 툴에다가 스트립트로 작성해 두시는 거예요. 매월 시스템에서 데이터 엑셀파일 다운로드 받게 돼요. 파일을 쥬피터노트북으로 불러와요. 그래서 내가 지난달에 짜놨던 스크립트. 여기에 보시면 데이터 로드해 오는 부분이 있어요. 혹은 엑셀 불러올 수도 있어요. 여기에 바로 여기 한 번에 쭉 셀을 실행해줄 수 있는 기능이 있거든요. 위에서부터 끝까지 실행하는 거예요. 그러면 매달 엑셀로 반복된 작업, 대용량 파일 다룰 때마다 시간 허비했던 거를 쥬피터 노트북 사용하게 되는 2, 3시간으로 줄일 수 있어요.
대용량 파일 2기가 3기가 되는 파일들이 생각보다 현실세계에 많거든요. NC소프트에서 데이터 분석하시는 분들이 세미나 열었는데 하루에 쌓이는 데이터 양이 어느 정도 된다고 공개를 했었어요. 어느 정도 될 것 같으세요? 엔씨소프트에서 하루에 쌓이는 양. 어느 정도 될 것도? 혹시? 맞혀보실 분 있으세요??
-상품 있나요?
-상품은 없습니다.
안타깝게도 상품은 없습니다. 혹시 하루에 엔씨소프트에서 쌓이는 로그파일의 양이 어느 정도 될 것 같으세요? 한번 상상해 보시면 보통 테라 단위로 쌓인다고 해요 하루에 쌓이는 양이. 그리고 회사에서 일을 할 때 시간당 기가단위로 많이 쌓여요. 파일을 로그를 해서 시스템으로 떨구게 되는데 기가단위의 파일을 떨구게 되는데 기가단위 파일을 엑셀로 가져와서 분석하는 거는 거의 불가능하다고 보시면 돼요.
현실 세계에서 생각보다 적은 데이터도 있지만 이렇게 큰 데이터들도 많이 있는데 이런 데이터들 가져와서 내가 잘 정제하고 분석할 때 판다스를 쓰게 되면 엑셀을 쓰실 때 보다 훨씬 더 유리하고 조금 더 여유시간을 가질 수 있다.
판다스라는 툴이 아까 말씀하신 것처럼 판다스 스크립트 익숙해지지 않으면 사용하기 어려울 수 있어요. 엑셀 매크로를 전혀 몰라서 배워야 해요. 엑셀 매크로 배우는 노력 정도라면 판다스 충분히 배울 수 있다는 거예요. 파이썬을 배우고 싶다고 하시는 분들께 판다스 배우는 거를 굉장히 많이 권장을 드리고 있어요.
그래서 아나콘다 안에는 쥬피터 노트북 편집기라고 말씀을 드렸어요. 판다스 설치되어 있고 넘파이 설치되어 있어요. 넘파이는 약간 블록처럼 되어 있어요. 블록모양은 수치 행렬데이터처럼 보이기도 해요 넘파이라는 거는 수치계산용 라이브러리라고 보시면 돼요.
그래서 아나콘다 안에는 파이썬도 있고 주피터 노트북도 있고 넘파이도 있어요. 싸이판이라든지 맵트립 써보신 분 계세요? 없으세요? 맵트립에 대해서 쓸 수 있는 기능들 맵트라립에서 거의 다 사용을 할 수 있어요. 맵트라립 시각화 라이브러리예요. 시각화 해보고 싶다고 할 때 맵트라립을 사용을 하시면 다양한 시각화 할 수 있고 파이썬에는 맵트라립 외에도 훨씬 더 쉽게 사용하게 만들어본 게 있어요. 씨본이라든지 지시플로 플라나인 라이브러리도 있어요. 혹시 알 사용해보신 분 있어요?
오픈소스 , 다양한 생태계라고 말씀을 드렸는데 우리가 예를 들어 과거에 툴 사스라든지 이런 툴은 다 유료잖아요. 파이썬이라는 프로그래밍 언어는 오픈소스고 무료예요. 인터넷에서 배울 수 있는 자료를 굉장히 많이 찾을 수 있어요. 내가 이런 식으로 분석해 봤어요라고 공유해놓은 자료들이 굉장히 많아요. 그래서 저희도 여기 깃허브에다가 정부혁신 1번가 사이트를 분석했던 프로그래밍했던 분석했던 소스포드들을 다 공개해놨어요. 이렇게 공개해놓은 소스포드들이 굉장히 많아요. 다른 사람들이 분석해 놓은 자료들을 통해서 내가 배울 수 있

는 자료가 굉장히 많은 게 파이썬 말고도 다른 프로그래밍 언어 장점이기도 해요.
깃허브에는 가보면 사람들이 소스코드들을 공개해놨다. 깃허브에서 배울 수도 있어요. 밑에 보시면 텐러플로 케라스 있는데 딥러닝 이런 것들도 파이썬으로 할 수 있어요. 활용해 볼 수 있다는 게 파이썬의 큰 장점이 될 거예요.
파이썬. 사용자들이 많이 하는 얘기인데 인생은 짧으니까 파이썬을 쓰자는 얘기예요. 다른 언어에 비해서 훨씬 배우기 쉽다고 볼 수 있다는 겁니다.
정부혁신1번가 사이트 들어가게 되면 정부혁신제안목록이 있고 목록에 들어가보면 어떤 제안을 원하는지 어린이집 교사 근무환경개선 이런 것들이 올라와 있고 추천을 해주게 되죠. 청와대국민청원처럼. 깃허브 사이트에 지난주 금요일에 해커톤을 했었어요. 여러 사람들이 모여서 했던 해커톤 결과물 올려두었고요.
그리고 여기 소스코드를 한번 보도록 할게요. 깃허브 사이트에 가게 되면 발표자료는 여기에서 발표했던 슬라이드 링크예요. 관련기사는 해커톤 관련기사고 여기 보시면 온라인 토론장 제작과정부터 국민이 참여한다 기사가 있는데 정부혁신 1번가 사이트는 클라우드소싱을 통해서 운영이 되고 있어요. 이런 것들 기사도 있고 제작과정부터 국민이 참여한다고 되어 있는데 오픈소스로 공개한다고 되어 있어요. 깃허브에다가 결과물 공개하고 있다고 보시면 돼요. 크롤링 소스코드 1번 되어 있는데 클릭하게 되면 바로 이렇게 클로보레이토 넘어가는데 뒤로 가서 새탭으로 열도록 할게요.
그래서 나중에 컴퓨터 사용할 수 있는 환경이 되면 들어와서 같이 실습해볼 수 있는데 연결하면. 로그인 해야 되네요. 여기에 직접 코드는 실행해 볼 수 있어요. 그래서 여기 아까 제가 지금 소개했던 라이브러리를 로드하는 코드들이 있어요. 정부혁신1번가 사이트 크롤링 한 건데요. 여기 가게 되면 크롤링 하기 전에 확인을 해야 될 게 있어요. 어떤 걸 확인하느냐. 사이트를 내가 크롤링 해도 되는데는 저작권 정책을 해야 해요. 각 사이트에서 로고정책 텍스트가 있는데 이 사이트를 검색엔진이 읽어가도 되는지 안 되는지 여부를 작성해본 거예요. 만약에 디설라우 되어 있는는 건 가져가지 말라는 거예요. 이 유알엘은 가져가지 마세요 하는 거예요. 디설라우 되는 건 크롤링 하지 말아야 한다 확인을 하고 하셔야 해요 슬래시 표시가 되어 있는 거는 전부 다 크롤링 해야 된다. 디설라우 빼고 전부 다 크롤링 해도 된다는 뜻이에요. 크롤링 소스코드 들어가도 볼게요.
파이썬에서 쓸 수 있는 라이브러리들을 로드를 해왔어요. 화면이 작아서 키우도록 할게요.
제가 아까 소개했던 판다스 파이썬에서 쓸 수 있는 엑셀이라고 보시면 돼요. 엑셀로는 대용량 파일을 가지고 와서 분석하기 어려워요. 찍지 않으셔도 링크에서 바로 들어올 수 있어요. 크롤링 소스코드는. 크롤링 노트북은 크롬브라우저로 여는 거를 권장드려요. 콜레보레이터가 익스플로어는 안 열릴 수가 있어요. 크롬에서 오픈해 보시는 거를 권장드려요. 판다스를 aspd 불러왔어요. 앨리언스를 지정해준 거예요. 별명을 별칭을 지정해 줬다고 보시면 돼요. 판다스라고 매번 길게 쓰기가 길기 때문에 pd라고 줄여서 쓰겠다는 의미예요. 내가 아무렇게나 지정해도 상관없지만 공식문서에서 pd 지정해서 사용하고 있어요.
모르는 걸 찾아봤을 때도 pd로 답변을 해놨기 때문에 여기에 공식문서 있는 대로 사용하시는 거를 권장드려요. 그다음에 여기 아까 사이트에서 크롤링을 해와야 하는데 사이트를 크롤링 해오기 위해서 보시면 목록이 있어요.이 목록들을 페이지를 받아와야 될 텐데 받아오기 위해서 리퀘스트를 사용을 할 거예요. 리퀘스트 불어오고 뷰티풀숩은 페이지 소스 보기를 해보도록 할게요. 리퀘스트라는 걸로 이 페이지를 불러오게 되면 우리에게 이런 식으로 소스코드가 보이게 돼요. 이 소스코드 안에서 저희가 여기 안에 있는 제목 근린공원이라든

지 타고 갈 수 있는 링크라든지 소스코드 안에 제안시작 제안인 제안내용 이런 것들을 가져와야 되는데 이런 내용들을 여기에서 찾아와야 되잖아요.
여기에서 소스코드 안에서 내가 원하는 내용을 찾아오기 위해서 뷰티풀숩이라는 걸 처음한다고 보시면 돼요. 랜덤 숫자가 생성한다고 써놨어요. 이 사이트에 계속 접근을 해서 크롤링 할 때 여기 보시면 게시물을 목록으로 가서 보도록 할게요.
여기에서 정부혁신 제안 페이지로 들어가서 목록을 봤는데요. 여기 목록에 이렇게 게시물들이 쭉 있어요.
그래서 여기 게시물들을 보면 페이지가 몇 페이지까지 있는지 볼게요. 36페이지까지 있어요. 한 번에 여기 있는 거를 한꺼번에는 불러오게 되면 이 사이트에 굉장히 부담이 돼요. 한꺼번에 트래픽 발생되면 서버에 부담이 되기 때문에 주기적으로 가져올 수 있도록 랜덤이라는 것을 불러왔어요.
그래서 여기 정부혁신1번가는 페이지가 많지 않은데 국민청원 사이트는 데이터가 많다고 얘기를 드렸어요. 그런 사이트를 텀을 주지 않고 랜덤하게 간격을 두지 않고 계속 크롤링 해오게 되면 사이트에 디도스 공격을 하는 것과 유사한 효과를 발생시키게 되고 서버에 부담을 주다 보면면 막힐 수 있어요. 랜덤하게 시간을 주기적으로 두고 랜덤하게 가져온다 보시면 돼요. tqdm라이브러레가 있는데 크롤링을 하다 보면 1페이지에서 36페이지까지 하는데 몇 페이지까지 진행이 되었는지 궁금해요 중간에.
그래서 중간에 어느는 정도 진행이 되었는지 확인하기 위해서 이 tqdm 라이브러리를 사용한다고 보시면 돼요. 그리고 re 텍스트 분석을 하기 위해서 가져온 거예요. 정부혁신국민포럼 페이지 가져오기 해서 url 들어와서 소스코드 가져오게 되는데요.
여기 결과물이 안 보여서 제가 잠시 깃헙 페이지로 들어가보도록 할게요.
제가 제 pc로 작업을 하려고 했는데 맥이 연결이 안 되더라고요. 그래서 여기에 있는 pc로 발표를 하다 보니까 로그인이 안 되어서 주고 실행을 못해 보고 있는데 여기 보시면 페이지 넘버 여기에 따라서 페이지를 가져오는 거예요. 제가 잠시만 로그인을 하도록 할게요. 잠시만요.
로그인이 되었는데 여기 연결이라는 게 있어요. 연결을 하게 되면 노트북이 연결이 돼요. 그래서 내가 실행할 수 있는 상태가 돼요. 쥬피터 노트북 서버 킨 거랑 마찬가지라고 생각하면 돼요. 라이브러리 켤게요.
라이브러리를 모두 해놨고요. url이 정부혁신 이 url 그대로 가져와서 페이지 넘버원만 따로 넣어두도록 했어요. pnum 변수를 만들어놨어요. 변환의 수. 특정 값을 넣어주게 되면 그 값을 이런 식으로 불러와서 보실 수 있어요. 변수에다가 페이지 넘버 이어 먼스를 넣어줬어요. 그러면 base url 콜레보레이토리 노트북은 제일 마지막에 출력을 해요. 베이스 url 출력이 되었어요. 여기에서 위에서 써준 페이지 넘버가 여기에 url 들어가 있고요. 먼스 tp 라는 2019년 6월이라는 숫자가 들어가 있어요.
쭉 내려보도록 할게요. 여기 url 리퀘스트라는 걸로 불러보라고 할게요. 코드 하면 셀을 하나 새로 만들어요. 컨트롤 v 하시고 쉬트 엔터. 200이라고 찍혔어요. 스테이터스 코드인데요. 200은 오케이를 의미를 해요. 베이스 url 접근해서 데이터 가져왔다고 보시면 돼요. 소스코드를 제대로 가져갔는지 확인해 보고 싶어요. 리스펀스 텍스트라고 되어 있어요. 리스펀스.text 이런 식으로 소스코드가 되어 있어요. 한 줄로 예쁘게 보이지는 않는데 아까 에게 페이지 소스 해서 봤던 내용이에요.
여기에 맨 앞에 이렇게 보시면 여기 소스코드 가서 보면 비어 있는 문자가 이렇게 많이 되

어 있어요. 여기 스크롤 해서 보시게 되면 보이게 돼요. 이런 식으로 소스코드를 받아오게 돼요. 뷰티풀숩이라는 걸로 파씽을 하게 된다. 다른 사람이 짠 코드를 볼 때 내가 이거를 보는 방법은 한 줄씩 한 줄씩 뜯어서 보는 거예요. 지금 밑에 전체 라인을 보면 의미 파악하기 어려워요. 그럴 때는 한 줄씩 뜯어서 본다. html 찍어서 볼게요. 아까 한 줄로 된 소스코드가 이렇게 예쁘게 보여요. 이게 뷰티풀숩이라는 라이브러리 사용을 해서 html 형태로 소스코드가 보이는데 정부혁신 1번가에서 마우스 오른쪽 클릭하고 소스코드 보기를 했던 내용이 이렇게 여기 보인다고 보시면 돼요.
그래서 개행이 돼서 지금은 예쁘게 보여요. 여기에서 정부혁신1번가에 있는 목록을 가져오고 싶어요. 텍스트를 가져오고 싶다 경력증명을 위한 경령등록제도시행 텍스트를 가져오고 싶어요. 마우스 오른쪽 클릭하고 여기에서 어느 부분인지 소스코드를 자세하게 보여주게 돼요. 여기에서 마우스 오른쪽 클릭하고 카피 셀렉터 내가 가지고 오고자 하는 소스코드에 셀렉터 부분이 카피가 됐다고 보시면 돼요.
그래서 여기에 가서 아까 제가 이 노트북에서 이 위에다가 붙여넣기를 했는데 여기 셀렉터 가져온 거 붙어넣기 해볼게요. 그러면 이 위에 있던 게 원래 있던 거고 제가 가져온 거예요. 여기에 보시게 되면 샵은 아이디값이라고 해요. css라는 게 있고 html 있고 자바스크닙트가 있어요. 몇 가지 웹에 대한 기본적인 지식이 필요한데요. html 소스코드에 대한 이해 그다음에 제가 말씀드린 css 이해도 필요해요 여기 컨텐트 아이디를 가진 거에 서제스션 클래스에서 ur 태그 다음에 dt a 태그 라는 내용을 가져오겠다는 내용이에요.
밑에 있는 거를 위처럼 줄여서 쓸 수 있어요. 차일드 빼고 써도 상관없어요. 밑에 거부터 가져올 때 차일드 지정 번거로울 수 있어요. 세렉터를 가지고 와서 적어주시게 돼요. html 셀렉터에다가 셀렉터 리스트에 ul 태그 가져오라 하라는 거예요. 이 부분 출력을 해보도록 할게요. 컨트롤 c를 해서 여기 위에 소스코드 하나 추가를 하고요. 그다음 붙여넣기. 소스코드가 너무 길 때는 하나씩 하나씩 한 줄씩 찍어보시면 조금 이해하기가 쉬워요. 태그스라고 해서 출력을 해보라고 할게요. 아까 위에는 전제소스코드 가지고 와서 복잡했어요. 그런데 내가 원하는 여기에 있는 정부혁신페이지에 있는 제목만 가지고 올 수 있어요. 제목만 가지고 와서 제목에 있는 tr 태그 ul까지 가져온 거예요. ul에 있는 0이라는 그냥 가져오는 대괄호를 묶어서 가져오게 돼요. 파이썬에서 리스트를 의미를 하는데 내용을 리스트 안에서 가져오고 싶다. 파이썬은 인덱스가 1부터 시작하지 않고 0부터 시작을 해요 0번째 인덱스 지정을 해주고 값을 가져오게 되면 결과 론을 없애고 값을 가져오게 돼요.
포천시근린공원 이거는 진행 중이고 중앙부처산하기관, 첫 번째 페이지 볼게요. 근린공원 있고 밑에 아마 공시송달 의뢰 그다음에 중앙부처산하기관 이런 식으로 목록에 있는 거를 한꺼번에 다 읽어왔어요. 여기에서 우리는 이 제목이라든지 아니면 링크를 타고 들어가서 이 안에 있는 내용을 다시 가져오고 싶어요. 제안분야라든지 제안시작 마감일 내용 이런 내용들을 더 자세하게 가져오고 싶은데 이런 내용들을 가져오기 위해서는 밑에 보게 되면 a 태그를 통해서 링크를 가져오도록 했어요. 목록을 가져온 것처럼 여기 리퀘스트를 통해서 목록을 가져온 것처럼 다 가져오고 싶어요. 모든 게시물을 가져오고 싶기 때문에. 월별 전체 제안목록을 가져오는 것을 하나 만들었어요.
다시 목록으로 가보도록 할게요. 사람이 이거를 긁어온다고 생각을 하면 url 있을 거예요. 각각 url 클릭해서 들어가서 긁어본다고 보시면 되는데 내가 하는 게 아니라 프로그램에 시킨다고 보시면 돼요. 여기에서 서제스천 리스트를 만들고 계속 호출을 하면서 목록이 없어질 때까지 하면 우리가 손으로 하나하나 크롤링을 하지 않아도 쥬피터 노트북 스크립트 돌

리게 되면 전체 목록을 가져올 수 있게 됩니다.
태그 안에 포에서 태그 href 이것만 복사해서 보도록 할게요.
태그스라는 변수를 하나 만들어줬는데요. 여기 프린트 해보도록 할게요. 프린트.
이렇게 하고 실행을 해주게 되면 아, 오류가 났는데.
여기 태그스에서 href.
없는 내용이 있는데 제가 가져오려고 해서 그랬어요. 전체 실행을 해줘야 하는데 그래서 태그가 있으면 가져오기. href 얘만 가져오게 한 거거든요. 그래서 링크만 가져오게 밑에 작성해놓은 거를 실행해 보도록 할게요. 위에 함수가 정리가 안 돼서 함수 실행을 해주고 하면 얘가 url만 가져오게 돼요. 여기에 보시면 태그에서 href 주소만 가져오게 크롤링을 하고 있는 거예요. 전체 리스트를 가져오고 있을 텐데요.
그래서 정부혁신 1번가 페이지 들어가서 보게 되면 url 하나씩 있어요. 각각의 게시물 들어갈 수 있는 데가. 페이지 번호 하나씩 클릭하고 들어가게 되면 각각 페이지마다 게시물들이 여러 개가 있는데 이 게시물들을 내가 다 가져오는 게 아니라 크롤링 소스코드 통해서 가져오게 한다. 그리고 여기에서 잘 가져왔는지 확인을 할 텐데 지금 36페이지까지 한꺼번에 지금 실행이 되어 있어요. 시간이 조금 걸리게 될 텐데 이렇게 하게 되면 다 가지고 왔어요.
보시게 되면 url만 가지고 왔어요. 정부혁신페이지에서 소스코드 보게 되면 href 안에 이런 링크가 들어가 있어요. 링크 주소만 뽑아오도록 한 거예요. 링크 주소를 하나씩 돌면서 여기 있는 내용들을 하나씩 가져올 수 있게 되는 거죠. 여기에서 url 하나씩 돌면서 이 아래에서 게시물 내용을 가져오는데 샘플로 하나씩 가져오도록 했어요. url 이거다 하면 sid 204번이고 페이지 넘버가 1이다 전체 소스코드를 다 가져오게 될 거예요. 여기에서도 제안분야라든지 제안시작일 마감일 제안인 제안내용만 가지고 오고 싶은 거죠.
여기 소스코드에서 각각 내용을 가져오게 했어요. 셀렉터는 아까 가지고 온 것처럼 제안내용 가지고 오고 싶다 하면 제안내용에 소스코드가 어느 부분인지 오른쪽마우스 클릭하시고 카피셀렉터 하게 되면 제안내용에 셀렉터를 가져오실 수가 있습니다. 코드 안에 있는 셀렉터를 가지고 와서 텍스트만 가져오게 돼요. 그래서 제안내용는 그대로 우리가 스크랩 해올 때처럼 똑같이 가지고 온다고 보시면 돼요. 그래서 가져온 내용들 이런 식으로 보시면 여기 어떤 제안분야인지 제안시작일이라든지 마감일 누가 제안을 했는지 이렇게 다 보실 수 있어요.
그다음에 여기에서 아이디값이라든지 이런 것들을 다 추출을 해서 전체 페이지를 돌면서 데이터를 가져와요. 데이터 가져온 것을 리스트에 담아주고 그 리스트를 판다스에 있는 데이터 프레임 만들어주고 저장을 하게 되는 게 이 제안페이지 내용인데 제안페이지 내용 다시 보도록 할게요. 데이터 폴더에다가 제가 이렇게 크롤링 해서 모아놓은 파일을 쭉 실행해 주게 되면 서제스션 csv 파일을 만들게 돼요. 이런 식으로 csv 파일에 카테고리 콘텐츠 제안시작일 마감일 이거까지 다 저장이 되어 있어요. 이런 식으로 csv 파일을 저장해놨다고 보시면 돼요.
이거는 지난주에 돌려놓은 거라서 지난주까지 데이터만 있어요. 최근 것까지 크롤링 해보고 싶다고 하시면 여기 오셔서 처음부터 끝까지 쭉 실행해 주시면 돼요. 여기 최신 내용으로 크롤링 해서 csv 파일 저장할 수 있어요. 저장하면 여기 왼쪽에 있는 다운로드 받을 수 있게 파일이 생겨요. 내 로컬 드라이브 저장이 된다고 보시면 돼요.
이 내용을 통해서 분석했던 내용을 간략하게 보여드리라고 할게요. 이디에이소스코드 되어 있는데 실행을 해보도록 할게요.

여기에서도 마찬가지로 아까 소개했던 판다스를 가져왔어요. 판다스 가져오고 씨본은 시각화 도구라고 보시면 돼요. 잠시 소개를 해드릴게요.
이게 씨본 사이트인데요. 여기 보면 갤러리가 있어요. 갤러리에 들어가서 보게 되면 굉장히 다양한 시각화 파일을 볼 수 있어요. 무언가 시각화를 하고 싶다. 화려한 히트맵 같은 거 그려보고 싶다면 클릭해서 들어가면 소스코드가 나와요. 소스코드 그대로 실행해 보면 소스코드 복사해와서 새 노트북을 하나 만들도록 할게요.
그리고 여기에 소스코드 붙여넣기 하고 바로 실행해 보도록 할게요.
그러면 이렇게 우리가 씨본 사이트에서 봤던 코드가 바로 실행이 되면서 그래프를 시각화했어요. 여기에서 박스플롯이라는 것을 그렸는데 내가 다른 씨본에 있는 다른 그래프 그리고 싶다고 하면 예제를 보고 바꾸시면 돼요.
박스플롯 이아니라 바프롯으로 바꾸고 싶다 하면 바꾸실 수 있어요.
맵플라립 보다 훨씬 문법이 간결하다는 게 씨본의 장점이에요. 첫 번째는 줄은 씨본 라이브러리 불러오는 개념이고 여기 스타일 지정. 사실 이 코드는 없어도 그래프 시각화 하는 데는 그다지 문제가 되지 않아요.
그다음 여기는 데이터 로드. 씨본에는 기본적으로 실습할 수 있도록 여러 가지 되어 있어요. 밑에 있는 소스 코드만 가져와서 복사하시고 여기 밑에다가 셀 새로 만들고 붙여넣기 그다음 실행 하면 넘파이. 임포트 넘파이 이렇게 하면 이런 식으로 바로 시각화가 돼요. 그래서 시각화 도구 소스코드를 다 외우실 필요가 없어요. 씨본 사이트를 알아두시고 내가 내 데이터 연구하는 분야라든지 아니면 도메인 업무하고 있는 분야에서 시각화 하고 싶다고 하면 갤러리 보고 예제 보고 따라해 보시는 걸 권장해드리는 편이에요. 여기 씨본에 있는 시각화 소개를 했고 씨본에 있는 도구를 사용을 해서 시각화를 하고 있어요.
새로고침을 한번 해보도록 할게요.
그래서 연결을 해주게 되면 노트북 자원 할당 되게 돼요. 이 노트북 콜라보레이토리 장점이 어떤 거냐면 내 노트북 사양이 좋지 않아요. 실습을 하다 보면 데이터 분석이라든지 크롤링을 하다 보면 내 컴퓨터 느려서 답답할 때가 있거든요. 구글 콜라보레이토리는 구글 서버를 활용하는 거예요. 클라우드 서비스이기 때문에 내 컴퓨터가 느려도 빠른 구글의 서버를 사용할 수 있다는 게 큰 장점이에요.
그리고 여기에 보시면 런타임에 보시면 런타임유형 변경 보시면 하드웨워 가속기 gpu tpu 지원하고 있어요. 딥러닝 소스코드 gpu 사야겠다 생각이 들게 되거든요. 처음부터 무리하게 구입할 필요 없이 구글에서 사용해 볼 수 있다는 게 장점이기도 해요.
아까 런타임에서 보시면 파이썬3, 2 실습해서 써보실 수 있어요. 가끔 공개된 소스코드로 실습할 때 파이썬 2로 되어 있는 소스코드가 있어요. 파이썬 2 따로 설치를 하셔야 하는데 콜라보레이토리에서 파이썬2로만 변경해 주면 파이썬2로도 실습이 가능합니다. 이런 것들이 콜라보레이토리의 큰 장점이라고 보실 수 있을 거예요.
그래서 데이터 로드 필요한 라이브러리를 가져오고요. 그다음에 서제스션 csv 파일이 없어요. 그러면 깃헙에 가셔서 csv 파일을 불러와야 해요 .
라이선스 정책 말씀해 주셔야 하잖아요. 제가 10분 정도만 더 진행하고 라이선스 진행해 주시는 될 것 같아요. url로 가서 eda 하는 부분이 있고 데이터가 있는데 데이터에서 서제스천 csv 파일로 가서 로우 파일 밑에 url이 보여요. 클릭 해서 복사를 해주시고요. 분석하는 소스코드 여기에 경로를 지금 얘로 다시 변경해 주도록 할게요.
해주게 되면 데이터가 로드가 됐어요. 지난주에 수집한 데이터가 171건이었어요. 171건 가

져와서 미리보기 해서 보게 되면 여기 제목이 있어요. 업체들의 어음기간을 줄여주세요 csv 파일 앞에서 5개 불러오게 돼요. 여기에서 만약에 3 지정을 해주게 되면 3개 불러오고 1개 하면 1개 불러오고 100개 하면 100개 불러와요. 지정해준 만큼. 파일 앞에서부터 불러오고 tail 하게 되면 뒤에서부터 불러와요. 이거는 판다스의 기능이에요. 엑셀파일 되게 큰 거 로드해 오다 보면 처음에 로드해 왔을 때 전체 다 읽으면 오래 걸리잖아요. 파일에서 1분간 미리보기 해서 보신다고 생각하시면 돼요. 판다스에 있는 데이터 미리보기 기능이고요. 인포. 이 데이터가 수치형 데이터인지 문자형 데이터인지 숫자로 되어 있는지 문자로 되어 있는지에 따라서 달라지게 돼요. 순자인데도 문자로 되어 있는 경우도 있기도 하고 데이터에 맞는 형태로 맞춰주는 것도 중요해요. 기본적으로 숫자 데이터 보여줘요. id 값 의미가 없어요. 171건 투표가 되었고 편차가 큰 편 최소값 최대값 이런 것들 볼 수 있어요. 지금 이거는 판다스로 데이터를 볼 수 있다고 보시면 돼요.
밸류카운트 해서 보게 되면 각각 카테고리별로 몇 건씩 있는지 개수를 세워줘요. 보건복지 분야에서 27건, 일반행정에서 27건 이런 식으로 되어 있다. 미리 그래프 보도록 할게요. 보시면 카테고리별 투표수와 제안수. 제안수를 보게 되면 제안수가 있고 오른쪽 그래프 스크롤 하게 되면 투표수가 있어요. 좀 작게 보도록 할게요.
사이트에 올라와 있는 각각 카테고리별고객님 어떤 카테고리에 어떤 제안이 많이 올라왔는지 어떤 제안에서 투표를 어느 정도 받았는지 볼 수 있는데 왼쪽 거는 제안수고 오른쪽 거는 투표수라고 보시면 돼요. 왼쪽 거는 카운트플롯 카테고리만 넣어주게 되면 알아서 우리가 위에서 계산했던 밸류카운트를 해서 그래프로 그려줘요.
보면 보건복지분야 일반행정 기타분야 제안이 많이 들어가 있다. 투표수를 보게 되면 제안수와 투표수는 일치하지 않는다는 것을 확인해 볼 수 있고 국민청원 데이터를 분석해 봤을 때 보면 카테고리에 들어오는 청원수 투표수는 일치하지는 않아요.
보건복지분야에서 보시면 검은색바가 굉장히 길게 되어 있어요. 이 검은색바는 실내구간이라는 걸 의미하는데 95% 데이터 구간이 여기에서 이만큼 구성하고 있다 의미를 해요. 평균 투표수 출력을 한 거예요. 여기는. 평균을 봤을 때 실내구간을 봤을 때 이렇게 그려졌다 보시면 되고요.
그다음 투표수 200건 이상 받은 거를 따로 출력을 해서 봤고요. 그다음 시간이 없으니까 뒤에 거는 쭉 실행을 해보시면 보실 수가 있을 텐데 여기 월별제안수라든지 제안일 기준의 투표수라든지 시각화 해봤고요. 워드클라우드도 그려봤어요. 보시게 되면 많은 등장하는 단어들 제안해서 등장하는 단어들을 보실 수가 있어요. 전체 제안에서 보시면 정부 불법 사회 청년문제 이런 키워드들이 많이 보인다. 그다음에 일반행정분야에서는 공무원 시스템 사업 공공. 보건복지 육아 주차 장애 어린이집 이런 단어.
안전 카테고리 주차 차량 이런 것들이 보이고 워드 클라우드로 큰 의미를 찾기는 어렵지만 그 텍스트에 어떤 단어들이 등장하는지 좋아요.
정부라든지 사회 기관 이런 것들을 너무 빈번하게 등장하는 단어라서 작은 단어들을 자세히 보고 싶어요. 그러면 여기 스탑워즈라는 게 있어요. 이거는 제외하고 볼 수 있어요. 그래서 여기까지 깃헙 사이트에 지난주에 해커톤 한 자료를 올려두었는데요. 이 글로 직접 실습이 가능합니다. 관심 있으신 분들은 실습해 보시는 거를 권장해드리도록 할게요. 저는 여기까지 발표 마무리 하도록 하겠습니다. 변호사님께 저작권 얘기를 하도록 할게요.
(박수)
-큰 박수를 부탁드립니다.

시간이 얼마 남지 않아서 간단히 소개드리고 질문 받는 게 더 좋을 것 같아요. 지난주에 저희가 행안부 와 해커톤을 진행했고 일부를 소개해드렸어요. 어떠셨나요? 파이썬을 배워야겠다고 생각이 드셨나요? 이거는 적성과 맞지 않아 서 영원히 멀리해야겠다 두 가지 생각이 있을 텐데 제가 생각하는 건 전체가 파이썬을 능숙하게 다루기 어렵다고 보고 최소한의 문법이라든지 대화할 수 있는 수준을 가지고 데이터 분석을 할 때 참여할 수 있는 인사이트를 갖는 그 정도로 우리가 공부를 하면 어떨까라고 생각이 들고 그래서 제가 작년에 박조은 개발자님과 같이 시민데이터랩 했었어요. 일반시민분들과 시민사회활동가분들을 대상으로 파이썬 포함해서 엑셀 강의를 진행했었는데 올해 여름도 진행할 계획입니다. 오늘 계기로 관심 가지시는 분들 계시다면 깃헙 홈페이지 참고해 주시면 될 것 같고요.
그래서 저희가 지난주 해커톤을 한 결과로 보면 일일이 크롤링을 할 필요가 있을까 정부가 데이터셋을 잘 정리해서 제공을 하면 국민제안 정책에 대해서 크롤링을 해서 공유하는 방법도 있겠지만 정부에서 처음부터 제공해 주면 힘들여서 크롤링 할 필요가 없겠다고 생각이 들어서 실시간으로 쓸 수 있는 것들을 개방하시고 정기적으로 모여서 분석할 수 있게 제안드릴 예정이에요. 그런 자리를 앞으로 계속될 것 같고 그 계획 역시 깃헙 홈페이지 들어오시면 조만간 소식을 들을 수 있을 것 같고요.
그리고 마지막으로 제가 말씀드릴 거는 1분 내로 정리될 수 있을 것 같은데 지난주에 했던 논의 결과는 이거입니다. 그래서 데이터를 크롤링 하면서 저작권 이슈에 대해서 많이 걱정을 많이 하세요. 글을 올려주신 분들이 국민분들에게 저작권이 있기 때문에 긁어서 데이터 분석을 했을 때 저작권 침해소지를 없을지 고민을 했는데 저작권 관련해서 정부혁신 1번가 보면 왼쪽이 현행입니다. 특별히 써도 된다 써도 되지 않는다. 얘기가 없어서 지난주에 논의를 해서 오른쪽과 같이. 자세한 말씀은 드리지 않겠는데 이 사이트에다가 게시물을 올리면 결과물은 정기적으로 데이터 분석을 위해서 사용된다 동의를 하시고 데이터 분석을 통해서 도출할 수 있도록 하는 이용약관개정안을 도출을 했습니다. 구체적인 내용까지 포함해서 도출을 했고요. 이거를 정부혁신 담당하는 정부부처에 제안을 해서 정부가 운영하고 있는 모든 사이트에 이러한 형태로 저작권 이슈 해결해 주는 이러한 형태로 개정해달라는 일종의 청원을 할 예정입니다. 여러분께서도 도움을 주시면 좋을 것 같고요.
논의하면서 기본 자료들을 정리를 했어요. 해외에서 저작권 정책이나 라이선스 어떻게 생각하고 있는지 제가 깃헙에 있는 구글박스 자료로 링크해드렸는데 보시면 시간나실 때 들어가 보시면 굉장히 자세하게 정리를 해놨기 때문에 데이터 분석을 편하게 할 수 있도록 저작권 정책 라이선스 정책에 개선되는 데 도움이 되고자 할 예정이고 내용을 자세하게 말씀드리면 3, 4시간 설명을 드려야 해서 다음 기회에 알려드리도록 하겠습니다. 저희가 10분 정도 남았는데 저작권 정책 궁금하신 분 저한테 말씀해 주셔도 되고 궁금하실 사항, 제안하실 사항 있으신가요? 5분 정도 받도록 하겠습니다.
-질문인데 보통 플랫폼 이라고 하면 여러 가지 의미가 있지만 정부참여 플랫폼이라고 하면 이걸 보면서 저는 변호사님도 비슷한 얘기 해주셨지만 이런 식의 정부 홈페이지나 기관들이 일일이 크롤링 하는 방식이 아니라 에이피아이 통해서 자료 접근하고 있는 사례라든지 미래 계획이라든지 이런 게.
-일단은 국민이 정부에 제안하는 사이트 관련된 에이피아이는 확인은 못해 봤고 제공하고 있지 않은 걸로는 알고 있고 공공데이터 포털 접속을 해보셨는지 모르겠는데 정부가 운영하고 있는 공공데이터 에이피아이 종합적으로 제공되고 있습니다. 필요하신 에이피아이 굉장히 많아요. 미세먼지 실시간 데이터 미세미세라든지 네이버 미세먼지 다 정부가 제공되고

있는 공공 에이피아이 사용되고 있을 거예요. 여러 가지 서비스가 제공이 되고 있으니까 검색해 보시면 굉장히 많은 에이피아이가 제공되고 있는 걸 알 수 있습니다. 품질이라든지 이슈는 많이 있어요. 이슈제기를 많이 해주시면 좋은 에이피아이가 제공될 수 있지 않을까. 정책 제안 국민청원 관련된 에이피아이 실시간으로 제공이 되었으면 좋겠다 생각을 하고 있는데 아직은 제공되지 않고 있습니다. 데이터 공부를 어떻게 해야 할지 이런 것도 좋고요. 어디 가면 수업을 들을 수 있는지 이런 질문도 좋고요.
일단은 조만간 무료로 들을 수 있는 시민 대상으로 하는 데이터 교육은 서울 엔피오 지원센터 지원을 받아서 8월, 9월 내로 강좌 개설할 예정입니다. 조금더 자세하게 천천히 여러 프로젝트를 하면서 데이터 공부를 할 예정이니까 오늘 계기로 이런 형태로 데이터가 관리되고 분석된다는 정도의 아이디어만 얻으셨으면 저는 충분하다고 보고 오늘 이 내용들을 다시 복습을 하고 싶으신 분들은 유튜브로 생중계가 되고 있고 녹화가 올라올 거예요. 천천히 다시 들어보시면 아마 여기 있는 코드를 직접 실행해 보시면서 따라갈 수 있을 것 같고요.
코드 자주 쓰는 거 몇 개만 활용하셔도 업무에 도움이 될 것 같고 크롤링 관련해서 업무 자동화 손이 줄어드는 코드기 때문에 잘 활용해 보시면 여기에도 있고 검색하면 코드가 굉장히 많이 나오기 때문에 잘 활용해 보시면 좋을 것 같아요. 파이썬을 많이 사용하는 분들 공무원이나 공공기관에서 도전을 많이 해보시면 좋겠다고 기회가 날 때 제안을 드리는데.
-청원제안 사이트 한해서 저작권 문제가 있다고 들리거든요. 개인적으로 쓴 글이 저작권 글이라는 건데 제가 문득 생각을 하기로는 일단은 목록에 다 나와 있는 전부 다 공개임에도 불구하고 저작권이 어떤 문제가 있는지.
-아까 크롤링을 하면 복제라는 행위가 일어나는데 저작권법에 보면 저작권 강의를 할 건 아닌데요. 저작권자 이용 허가를 받고 복제라는 행위를 할 수 있어요. 원칙이. 원칙대로 하면 10만 명이 글을 쓰셨다가 허락을 받고 복제를 해서 분석을 하는 원칙입니다. 예외가 있습니다. 학술연구목적, 공익목적으로는 이용 허락을 받지 않고 복제를 할 수 있는 예외조항이 있는데 데이터 분석 목적으로 자기만 보려고 했던 저작물이 아니라 복제해서 분석하는 것은 허락할 수 있겠다는 게 제 판단이고 돈을 벌기 위한 형태가 아니기 때문에 안심하고 분석하셔도 좋습니다. 다만 조금 분명히 하기 위해서는 데이터 분석을 위한 복제는 들어가면 좋겠다 한 가지가 있고 법개정이 어려우면 이용약관을 통해서 해결하는 두 가지가 있는데 박정 의원님께서 저작권법 개정안이 올라와 있는데 동참을 해주시면 좋을 것 같은데 국회가 일을 잘 안 하고 있어서 당분간은 어려울 것 같고 당장 할 수 있는 거는 이용약관을 통해서 내가 올린 저작물이 데이터 분석 형태로 분석될 수 있구나 미리 알려드리고 게시하시는 분들에게 동의를 받는 이용약관을 개정하는 정도는 걱정없이 쓸 수 있지 않을까 법개정은 장기 과정. 이용약관개정은 단기 과정으로 해서 국회를 열심히 일하도록 할 수밖에 없을 생각요.
없을 것 같고요. 조만간 공론화 될지 않을까. 혹시 문제가 생기면 저한테 연락을 주세요. 따로 끝나고 명함을 드릴 테니까 저작권 때문에 무서워서 데이터 분석 못하겠다는 분들은 저한테 연락주시면 오늘 오신는 분에 한해서.
-저는 학술 연구 측면에서 알 포함인지.
-알이요? 알하고는 전혀 다른 거라서.
씨본도 파이썬이에요. 제가 오늘 소개한 거는 다 파이썬이고요. 아까 지지플로지라고 한 건 알에서 시각화 하는 거라고 보시면 돼요. 그런 식으로 사용을 할 수 있는데 알과 파이썬은 완전히 다른 거라고 보시면 돼요. 영어와 한국어 차이점 정도로 보시면 될 것 같아요.

-통계 프로그램과 연관시켜서?
-통계프로그램 연계시키는 거는 별개의 문제일 것 같아요.
-파이썬은 확실히 하는 걸로 말씀을 드리고 별도의 비용 없이 박조은 개발자님도 참여해 주실 수 있을 것 같은데. 제가 자꾸 홍보를 말씀해 주셔서 죄송한데 파이콘이라는 컨퍼런스가 있는데 올해는 8월 17일인가요 광복절쯤에서 파이콘이라는 행사가 열리게 되는데요. 거기 오시면 공공데이터 관심있는 분들도 계시고 민관에서 많이 오실 텐데 국회에서 파이썬 이용하시는 분이 계셔서 국회에서 한 분이 발표해주시는 걸로 알고 있어요. 저도 신청해서 가볼 예정인데 거기 오시면 저를 만날 수 있을 것 같고 국회에서도 데이터 분석을 열심히 하고 계신다는 걸 전해드리고요.
-파이콘 얘기 해주셨는데 파이콘에서 튜토리얼 진행을 해요 서울시 자전거 따릉이 튜토리얼 하는데 관심있는 분들은 파이썬으로 따릉이 분석 같이 해보실 수 있어요.
파이콘.kr 사이트로 들어오시면 돼요.
-또 질문 있으신가요? 3분 정도 시간이 남았는데 그러면 식사하시고 너무 졸린 시간에 어려운 걸 가르쳐드려서 많이 힘드셨을 텐데 저희는 준비한 건 이 정도로 마무리 하도록 하고 다음 데이터 분석 주제로 만났으면 좋겠습니다. 장시간 들어주셔서 감사드립니다.