**2022학년도 2학기 중간과제물(온라인제출용)**

**교과목명 : 빅데이터의이해와활용**

**학 번 : 202135-367895**

**성 명 : 김태정**

**연 락 처 : 010-41**

**과제유형(공통형/지정형) :**

- 이하 과제 작성

**1번문제**

**(1)빅데이터**

과거에는 정형 데이터의 시대였습니다. 정형데이터는 처음부터 정해진 틀을 가지고 모으는 데이터를 의미하며 빠른 속도로 데이터를 수집할 수 있고 수집한 데이터는 바로 사용이 가능 하였습니다.

반면 현재는 비정형 데이터의 시대로 트랜드가 넘어 왔습니다. 비정형 데이터란 소셜네트워크나 IoT, 그 외 사용자 행동 정보를 가진 데이터를 의미합니다. 이 데이터는 과거에는 컴퓨터와 인터넷의 느린 속도, 그리고 데이터 수집의 문제성으로 크게 대두되지 못하였으나 현재는 컴퓨터와 인터넷의 성능이 증가하고 IoT나 스마트폰을 이용하여 데이터의 수집이 유용하게 되었으며 하둡, 스파크등의 빅데이터 솔루션이 등장하여 비정형 데이터의 문제점들이 극복되고 있습니다.

이러한 비정형 데이터와 같은 데이터를 빅데이터라고 부릅니다. 빅데이터는 기본적으로 3V를 충족하면 빅데이터라고 부릅니다. 3V는 각각 규모(Volume), 다양성(Variety), 속도(Velocity)를 의미합니다. 요즘은 정확성(Veracity)와 가치(Value)를 추가해서 5V라고도 부릅니다.

5V의 각요소를 좀 더 자세히 보자면 규모(Volume)의 정의는 딱 정해진 것이 없습니다. 다만 앞으로 데이터가 더 커질 것이라고 예상하고 이에 대비하는 서비스가 개발될것으로 예상됩니다. 다양성(Variety)의 경우 과거에는 RDB의 데이터를 바탕으로 한 처음부터 정보를 위해 탄생한 정형데이터에 대비해서 사용자의 정보를 분석으로 바탕으로 얻어낸 비정형 데이터를 사용합니다. 속도(Velocity)의 경우 기기와 인터넷 속도 및 분산컴퓨팅을 사용하여 극복해 나가고 있습니다. 마지막으로 정확성(Veracity)와 가치(Value)는 어떻게 보면 당연한 이야기인데 데이터가 아무리 많아도 그에 맞는 통찰이나 정보를 얻지 못한다면 의미가 없기 때문입니다.

그리고 빅데이터는 결국 AI와 연결됩니다. AI가 반드시 필요한 것은 아니지만 빅데이터의 발전이 AI를 가능하게 한 것은 사실입니다. AI에 사용하기 위해서는 데이터의 생성, 가공, 처리, 분석 모두 중요합니다. AI와 빅데이터는 서로 쌍으로서 발전해 나가고 있습니다.

**(2)데이터경제**

빅테크 기업(네이버, 카카오, 애플, 구글 등)데이터를 수집, 가공, 분석등을 거쳐 서비스 하고 있으며 이러한 데이터들은 21세기의 원유의 역할을 하고 있습니다. 또한 이러한 데이터들은 기업과 개인의 부가가치를 높이는 경제활동의 중요한 생산요소가 되고 있습니다.

위의 예처럼 데이터 생산자, 데이터 서비스 제공자, 데이터 비즈니스 이용자, 최종소비자 끼리 데이터를 생산, 소비하며 이는 마치 사슬처럼, 또한 현재의 경제처럼 얽히고 섥혀 있습니다. 이는 경제의 기본 생산요소인 자본, 노동에 데이터가 결합이 되었습니다. 2011년 가트너 보고서에 뉴먼이 데이터경제를 처음 언급하였습니다. 데이터 경제는 위와 같은 현상이 고용 창출에 주요한 역할을 하는 경제를 의미합니다.

생산요소로서 데이터는 한계비용이 없는 특징을 가집니다. 데이터가 다른 데이터와 결합할수록 그 가치는 높아지는 특징을 가집니다.

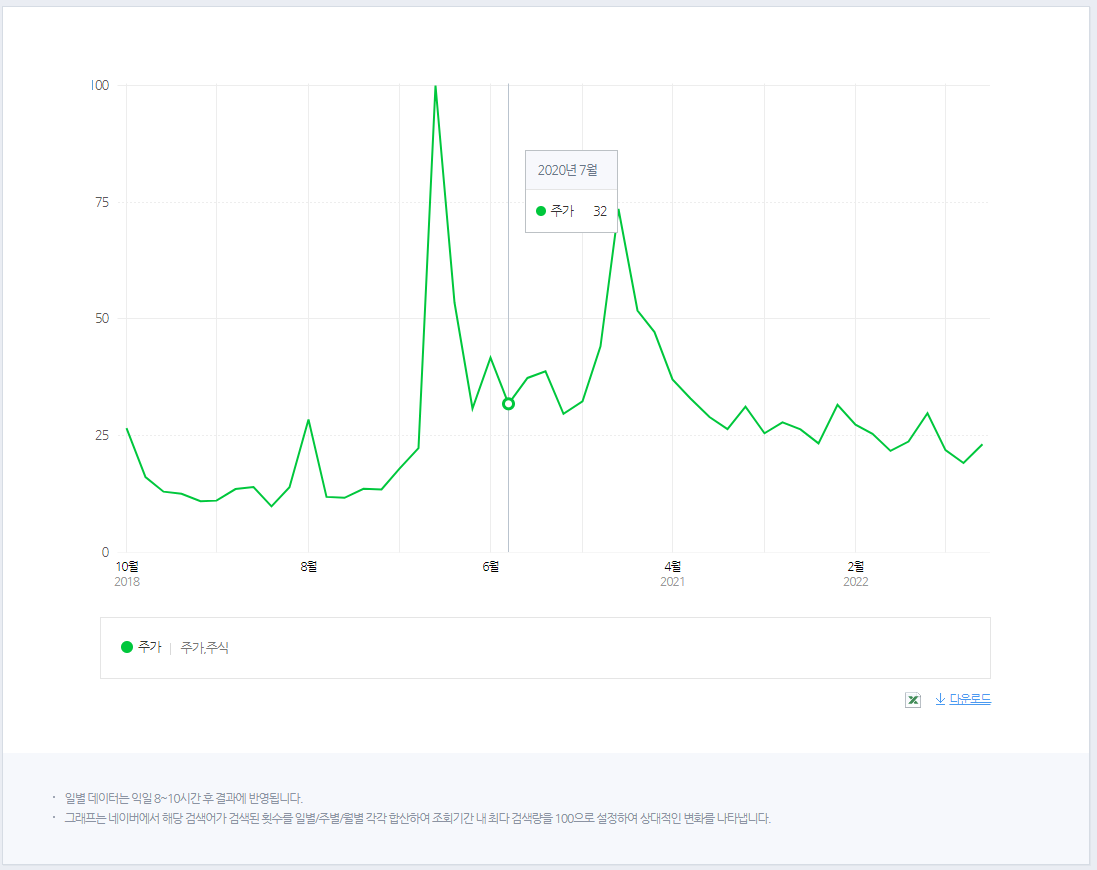
데이터경제는 데이터와 경제가 결합한 요소이므로 이처럼 데이터가 경제에 영향을 미치기 위해서는 서비스와 같이 소유권도 있고 거래도 가능하며 보호되어야 합니다. 각 나라에서는 각자에 맞는 데이터 관련 제도가 존재합니다. 한국에서는 개인정보 보호법, 신용정보의 이용 및 보호에 관한 법률, 데이터 산업진흥 및 이용촉진에 관한 기본법, 데이터 3법등으로 보호받고 있습니다.

데이터는 플랫폼 기업과 정부가 보유하고 있습니다. 21세기는 글로벌 시대고 글로벌 플랫폼은 미국에 주로 있기에 정보의 불균형이 발생합니다. 이는 데이터경제에서는 경제의 차이로 나타납니다. 한국도 이 격차를 줄이기 위해서 데이터 경제에 많은 지원을해야하며 또한 플랫폼 기업의 과도한 시장 지배력을 완화하기위해서 많은 제도 및 대비책을 마련해야합니다.

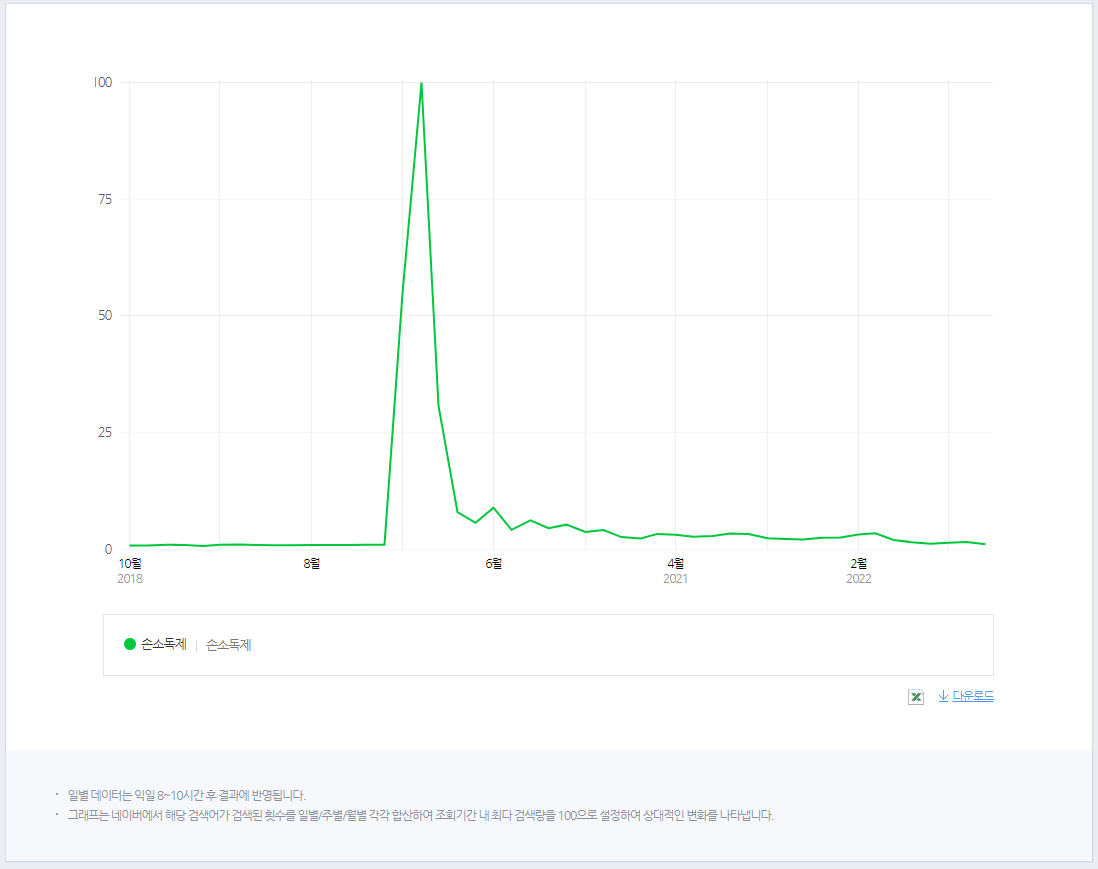
데이터경제는 아직까지는 기업위주로 언급되는 경향이 있습니다. 그러나 현재 새로운 바람이 불고있습니다. 개인이 데이터의 주체가 되며 그 데이터를 경제주체로서 거래를 할 수 있게 하는 것입니다. 이를 마이데이터라고 부릅니다. 여러 관공서에서도 마이데이터 세미나를 진행하고 있으며 앞으로는 플랫폼 빅테크 기업들이 데이터를 그냥 사용하지 못하고 일종의 대가를 지불하고 사용하게 되는 시대가 오게 될 것입니다.

**2번문제**

저는 주식과 손소독제를 검색했습니다.



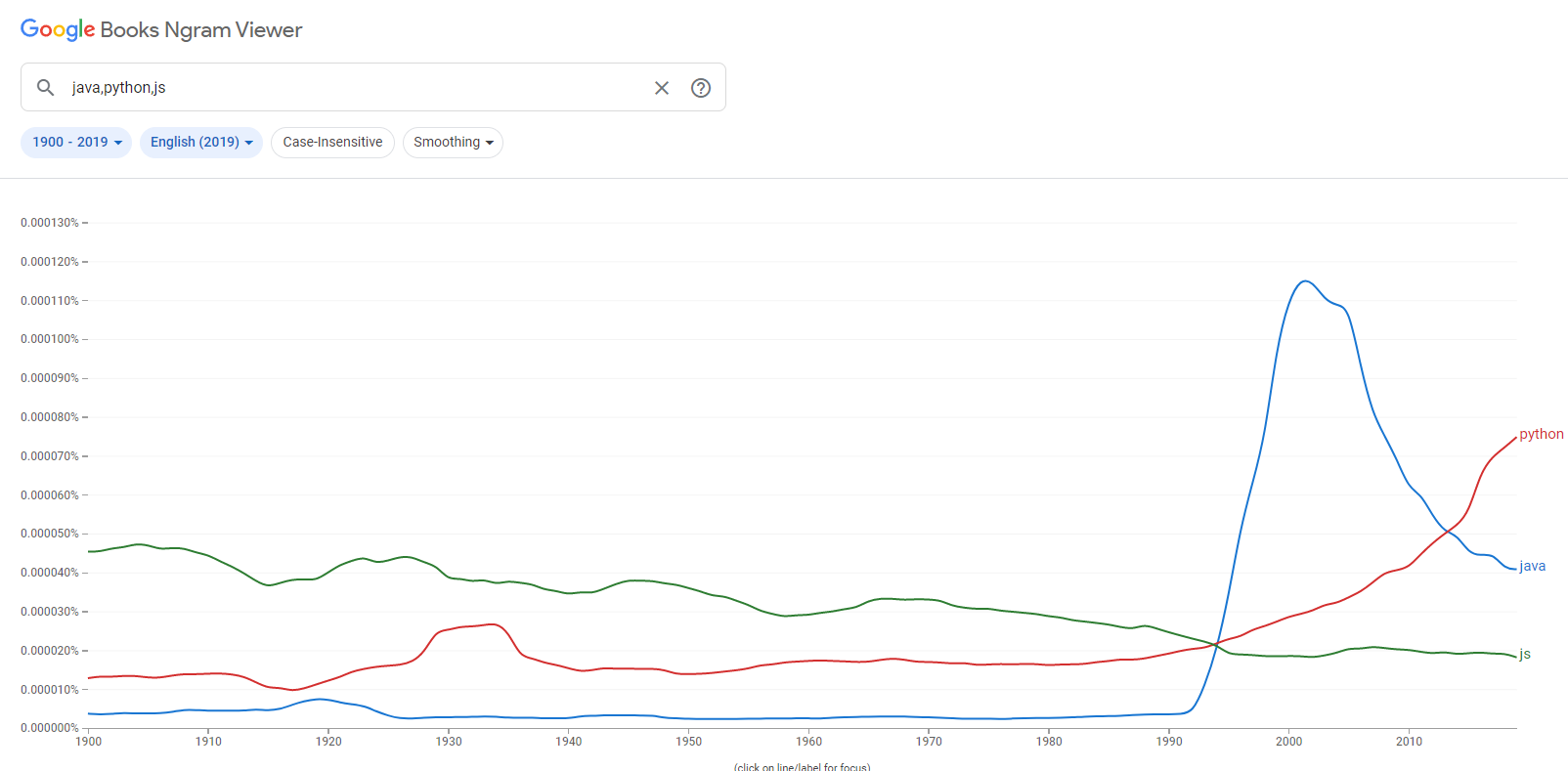
주식의 경우 코로나가 심해지기 시작하는 10월을 중심으로 검색량이 폭증하는 것을 확인할 수 있습니다. 후에도 과거에 비해서 검색량이 높아진것을확인할 수 있습니다. 코로나가 직접적으로 주식과 연관이 있진 않을것입니다. 하지만 코로나로 유동성이 증가하였고 증가한 유동성 자산이 투기자산에 몰려서 주가가 폭등한 것을 확인할 수 있습니다.



손소독제의 경우 코로나가 증가함에 따라 손소독제의 검색량이 폭증했습니다. 그 후에도 과거에 비해서는 검색 트렌드가 일정하게 유지되는 것을 확인할 수 있습니다. 따라서 코로나와 손소독제는 상관관계가 있으며 초기 이후에도 손소독제에대한 검색량이 꾸준히 지속되고 있음을 알 수 있습니다.

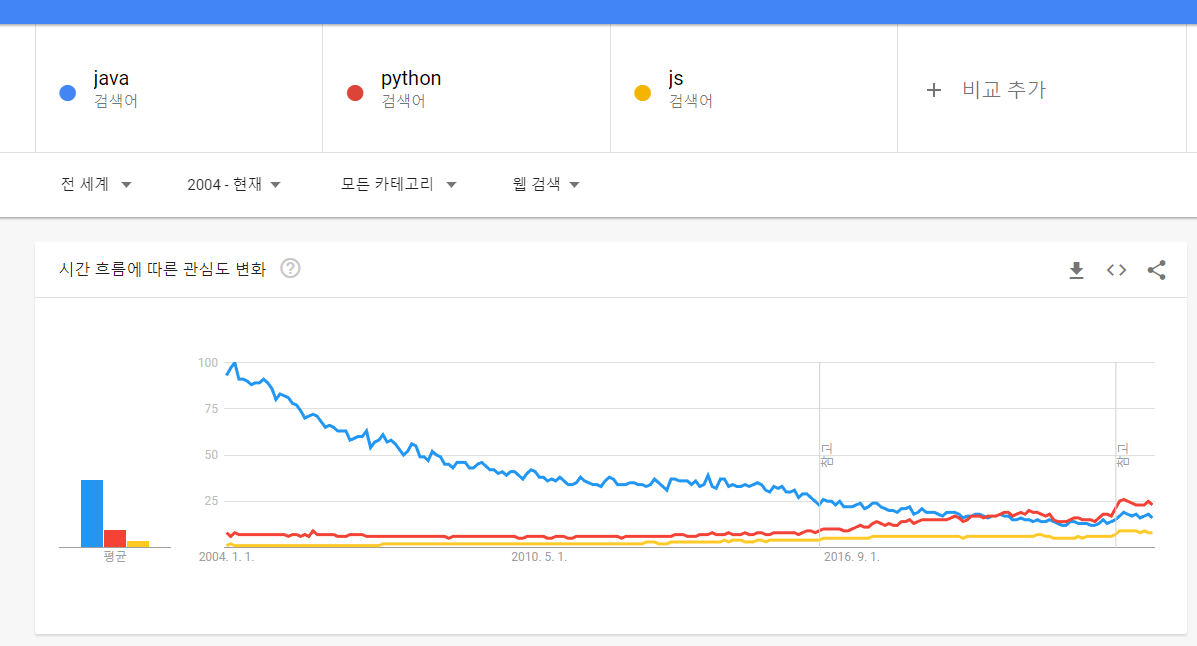
**3번문제**

**(1)번문제**



구글 ngram viewer의 경우 책 내용을 뷰어로 보는 것인데 아직도 자주 사용되는 언어인 java와 python, js를 검색했습니다. 문제는 각각의 언어가 나오기 이전인 세대들에도 표시가 되는데 이는 java의 경우 java 섬이, python의 경우 뱀 파에톤, js의 경우 문장에 js가 들어간것으로 생각됩니다. Java와 python의 경우 최근에 책에서 언급량이 늘어났는데 이는 프로그래밍언어로서 유명해진것과 일치합니다. 다만 js의 경우 오히려 예전에 비해서 줄었는데 이는 js라는 키워드 자체는 어느 책에나 있는 것이게 의미있는 정보를 얻지 못하는 것으로 판단됩니다.

**(2)번문제**



같은 키워드를 검색했을 때 최근 트렌드를 잘 반영하는 것을 알 수 있습니다. Java의 경우 인기가 점점 떨어져가는 모습을 도식화해서 보여주며, 파이썬과 js의 경우 자바에 비해 인기가 올라가는 현재의 모습을 확인할 수 있습니다. 또한 구글 트렌드는 검색엔진 결과를 바탕으로 제공하는데 검색엔진은 해당 언어의 키워드를 분석하여 알맞게 제공을 합니다. 가령 js의 경우 여러 문장에 키워드로 들어가지만 구글 내부 시스템을 통해 이 키워드가 프로그래밍언어 자바스크립트임을 알 수 있습니다. 그래서 키워드를 javascript로 바꿔도 거의 같은 결과가 제공됩니다.

**(3)번문제**

두 서비스를 종합했을 때 공통적으로 자바의 유행이 저물고 파이썬의 유행이 증가하는 것을 확인할수 있습니다. 이것은 실제로도 그러합니다. Ngram viewer의 경우 다른 언어들(C,C++.BASIC,JAVSCRIPT등)은 검색하기도 쉽지않고 검색해도 유의미한 결과를 낼 수 없는 경우가 많았습니다. 위의 예시들은 고유명사이거나 작은 단어였기 때문입니다. 따라서 ngram viewer는 뚜렷한 주제를 사용하면 더 효과적일 것 같고 구글 트렌드의 경우 거기에 구애를 받지 않습니다. 다만 인터넷이 예전부터 있던 것은 아니기에 과거의 정보들은 ngram viewer를 통해 어느정도 유추할수 있을 것 같습니다.

※ A4용지 편집 사용