

⋮ 문제해결을 위한 데이터 분석 및 시각화

빅데이터와 데이터의 이해

한성대학교 노은희 교수

서울세계도서관정보대학 도서관방문기관으로 선정

“미래로 향하는 새로운 이정표”



오늘의 학습

학습 내용

1. 빅데이터/정보와 데이터/데이터베이스/DBMS 이해
2. 빅데이터 활용과 데이터 제공 사이트

빅데이터 개념

- 디지털 환경에서 다양한 멀티미디어 콘텐츠 증가, SNS 서비스의 확산, 사물 인터넷의 증가, 스마트폰 기기의 이용 확대에 따라 새롭게 생성되는 데이터의 규모가 방대하고, 생성 주기도 짧고, 형태도 수치 데이터뿐 아니라 문자와 영상 데이터들이 기하 급수적으로 생성되어 기존의 방식으로서는 저장, 관리, 분석이 어려운 대규모 데이터
- 빅데이터 환경은 과거에 비해 데이터의 양이 폭증했다는 점과 함께 데이터의 종류도 다양해져 사람들의 행동은 물론 위치정보와 SNS를 통해 생각과 의견까지 분석하고 예측할 수 있다.
- 기존 데이터의 양이나 종류가 너무 커서 기존 방법으로는 도저히 수집, 저장, 검색, 분석 등이 어려운 데이터를 총칭해 일컫는 용어

[네이버 지식백과] [빅데이터 정의](#) (빅데이터, 2013. 2. 25., 정용찬)



빅데이터의 환경

SNS 이용 급증으로 인한 비정형 데이터 생성, 개인정보 -> 활용가치

스마트폰과 태블릿 PC의 빠른 확산 -> 개인의 위치정보 수집으로 인한 마케팅 분야에서 데이터 활용

각종 센서로부터 데이터 수집, 저장 - IOT



정보와 데이터

정보와 데이터

자료(Data)

가공되지 않은 일련의 사실들

정보
(Information)

의미있고 유용한 형태로 가공된 데이터

자료(Data) : 현실 세계에서 관찰이나 측정을 통해 수집한 단순한 사실이나 결과값으로 가공되지 않은 상태
정보(Information) : 의사 결정에 도움을 줄 수 있는 유용한 형태로, 자료를 가공(처리)해서 얻을 수 있는 결과



데이터 이해

■ 데이터 정의

이론을 세우는데 기초가 되는 사실이나 바탕이 되는 자료
관찰이나 실험, 조사로 얻은 사실이나 자료
컴퓨터가 처리할 수 있는 문자, 숫자, 소리, 그림 따위의 형태로 된 자료

■ 데이터 구분

- ① 정량적 데이터(Quantitative Data)는 주로 숫자로 이루어진 데이터
- ② 정성적 데이터(Qualitative Data)는 문자와 같은 텍스트로 구성되며 함축적 의미를 지님




데이터 이해

■ 데이터의 유형

정형화에 따른 분류

데이터 분석에서는 정형 데이터를 이용한다. 반정형 데이터는 수집된 데이터를 정형화하는 과정이 필요해서 정형 데이터보다 데이터 처리에 더 많은 시간과 비용이 발생

데이터 유형	설명
정형 데이터	정해진 형식과 구조에 맞게 저장 되도록 구성된 데이터, 연산이 가능(관계형 데이터 베이스의 테이블에 저장된 데이터로 회원정보 데이터, 매출 데이터, 상품 데이터 등)
반정형 데이터	데이터의 형식과 구조가 비교적 유연하며 연산이 불가능(XML, HTML, 웹 로그, 센서 데이터 등)
비정형 데이터	구조가 정해지지 않은 대부분의 데이터로 연산이 불가능(이미지, 동영상, 음성, 이메일, SNS 등)



데이터베이스

여러 사람에 의해 공유되어 사용될 목적으로 통합하여 관리되는 데이터의 집합

자료항목의 중복을 없애고 자료를 구조화하여 저장함으로써 자료 검색과 갱신의 효율을 높임.

데이터베이스

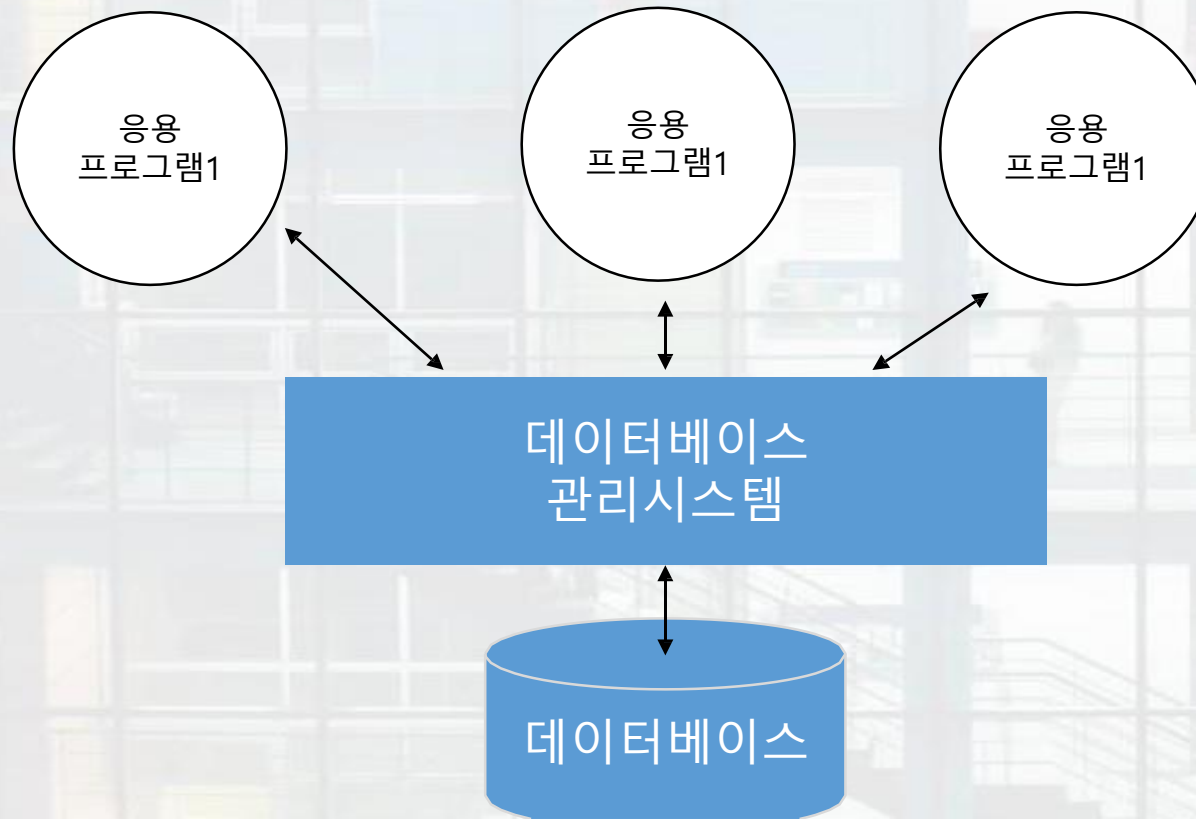
데이터베이스(database)

- 데이터들의 중앙집중화와 중복 데이터에 대한 통제를 통해 다수의 응용프로그램들을 효율적으로 지원



DBMS(DataBase Management System)

- 데이터베이스를 조작하는 별도의 소프트웨어(DBMS : DataBase Management System)
- 데이터베이스 관리 시스템이란 데이터베이스를 관리하며 응용 프로그램들이 데이터베이스를 공유하며 사용할 수 있는 환경을 제공하는 소프트웨어
(예: DBMS는 Oracle사의 Oracle, My_SQL MicroSoft사의 MS_SQL)



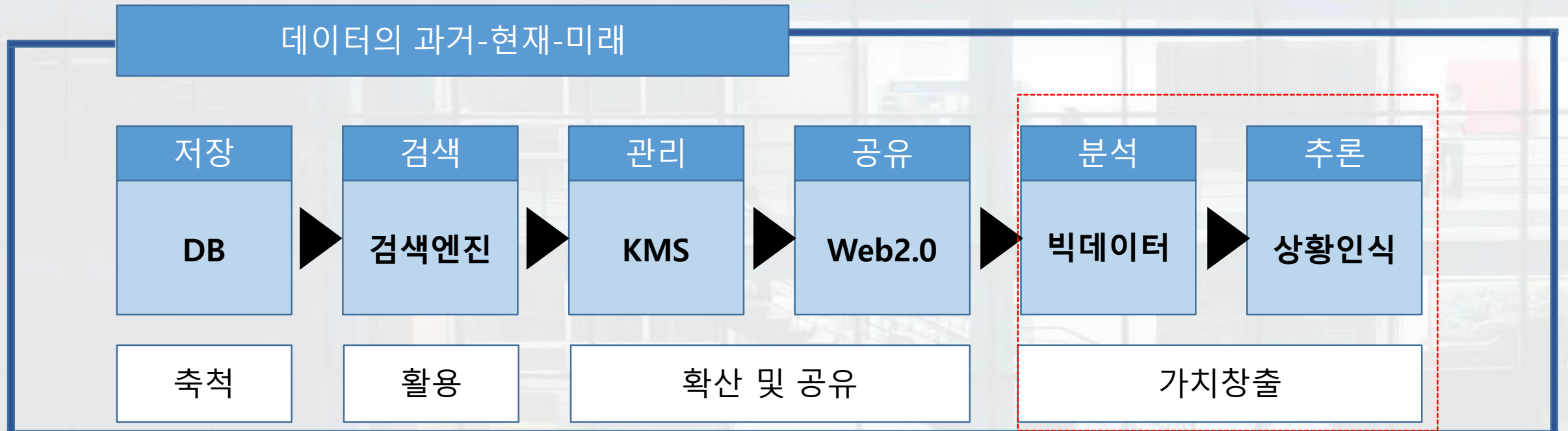
왜 빅데이터인가?

정보통신 기술의 주도권이 데이터로 이동

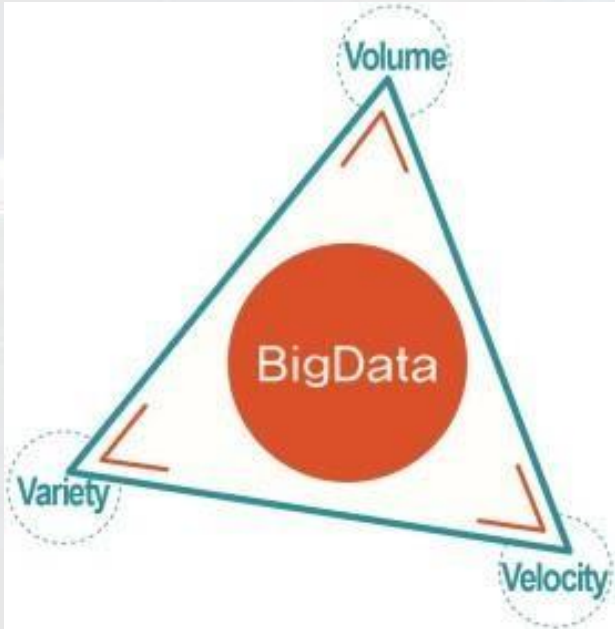
- 인프라, 기술, SW -> 데이터

공간, 시간, 관계, 세상등을 담은 빅데이터

빅데이터는 미래 경쟁력과 가치 창출의 원천



빅데이터의 특징_3V



- 규모(Volume) : 크기는 일반적으로 수십 테라 바이트 혹은 수십 페타바이트 이상 규모의 데이터 속성을 의미(MB,GB->TB,PB,EB)
- 속도(Velocity) :속도는 대용량의 데이터를 빠르게 처리하고 분석할 수 있는 속성. 융복합 환경에서 디지털 데이터는 매우 빠른 속도로 생산되므로 이를 실시간으로 저장, 유통, 수집, 분석처리가 가능한 성능을 의미(유통활용주기 몇시간~몇주-> 분,초 단위로 유통 활용)
- 다양성(Variety): 다양성(Variety)은 다양한 종류의 데이터를 의미하며 정형화의 종류에 따라 정형 데이터(고객 신상 데이터, 매출, 재고, 회계데이터등), 비정형 데이터(동영상, 음악, 메시지, 위치정보등)로 분류

<출처>[네이버 지식백과]

빅데이터의 특징_4V

빅데이터란



데이터 형식이 다양
(variety)



데이터 생성 속도가 빠름
(Velocity)



대규모 데이터의 집합 및 관련
기술과 인력을 통칭
(Volume)



분석 가치와 활용효과 측면으
로 의미가 확대되는 추세
(Value)



빅데이터 활용을 통해 국가 및 기업의 경쟁력 강화

- 공공: 국가 미래전략 수립, 사회현안 해결, 공공서비스 혁신
- 민간: 서비스 질 제고, 생산성 향상 등 경쟁력 강화



**빅데이터를 활용하여 데이터 중심의 정책수립을 지원
선제적 공공서비스를 개발하여 유능한 정부 구현을 위한 사업**

- 빅데이터 표준분석모델 정립 및 확산
- 빅데이터 이용 활성화 추진

<출처> <http://www.bigdata.go.kr/intro.html>

빅데이터 사례 영상



빅데이터 플랫폼

빅데이터 기술의 집합체이자 기술을 잘 사용할 수 있도록 준비된 환경



<출처>[네이버 지식백과]

빅데이터 활용사례



아마존은 경쟁업체의 가격, 주문내역, 예상 이익률, 웹사이트에서의 활동 등 방대한 데이터를 수집해 가격을 10분마다 최적화
-> 가격을 관리하여 매년 25%의 수익을 올림.



스타벅스 매장은 매장을 내기전에 빅데이터를 기반으로 상권 분석, 다른 스타벅스의 위치, 교통 패턴, 지역인구 통계등의 데이터를 다량 수집하여 이를 분석하여 최상의 입점 위치 찾기

빅데이터를 활용하여 고객에게 최상의 경험을 제공, 자체 어플리케이션을 통해 소비자들의 정보 수집하여 개별 고객의 커피 취향부터 방문 예상 시간까지 알아내서 고객의 취향에 맞게 서비스

빅데이터 활용사례

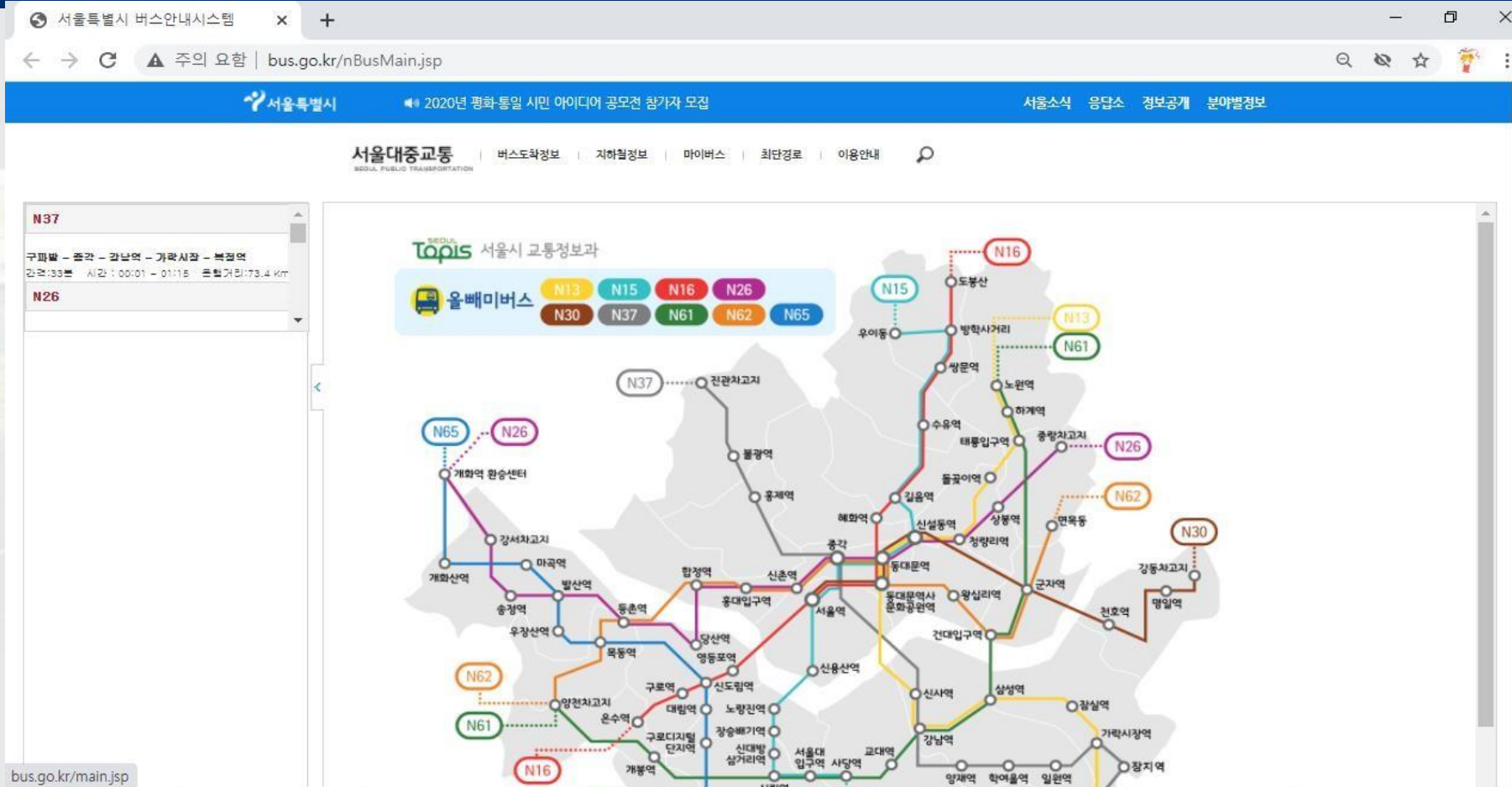
구글의 독감 트렌드 서비스

빅데이터를 이용하여 검색창에서 발열, 기침 등의 검색 빈도로 독감 유행수준 파악

구글은 자사 검색 데이터를 활용해 독감 유행 수준을 예측한다. 독감이 유행하면 관련 검색어가 많아지는 등 독감 관련 주제를 검색하는 사람 수와 실제로 독감 증상이 있는 사람 수 사이에 밀접한 관계가 있다고 보았기 때문이다.

실제로 미국 질병통제예방센터 데이터와 비교해본 결과 구글의 예측은 신뢰성 높게 나왔다. 이같은 관계를 바탕으로 구글은 **2008**년 부터 '독감 트렌드'라는 독감 예측 서비스를 선보였다

빅데이터 활용사례



**서울시가 빅데이터를
활용하는 방법**

<빅데이터 활용사례: 서울시 심야버스 노선 최적화>

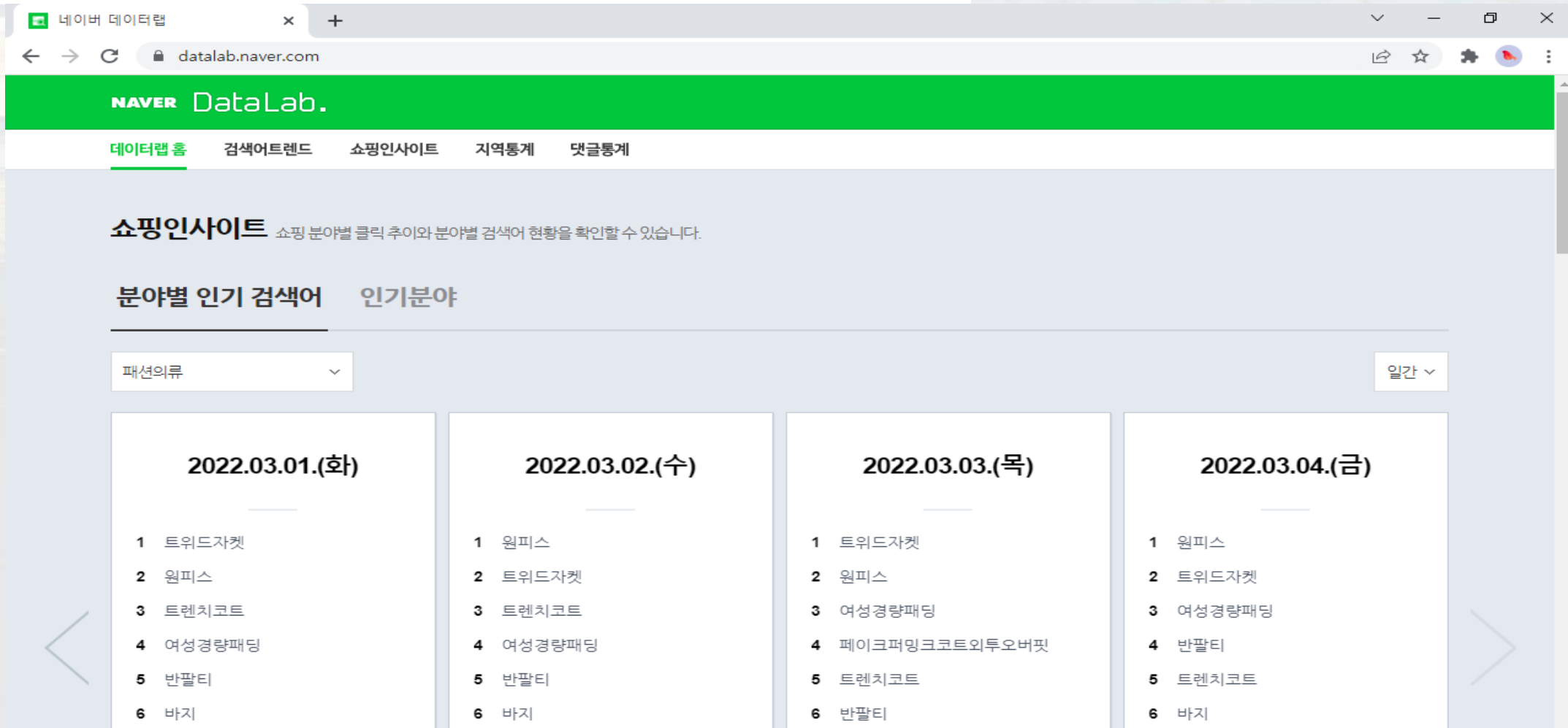
-서울시는 시민들의 휴대폰 통화량을 기반으로 한 KT의 유동인구 데이터
-시가 보유한 교통 데이터를 융합·분석

->최적의 심야버스 노선을 구축

네이버 데이터 랩 시장의 트렌드 분석시스템

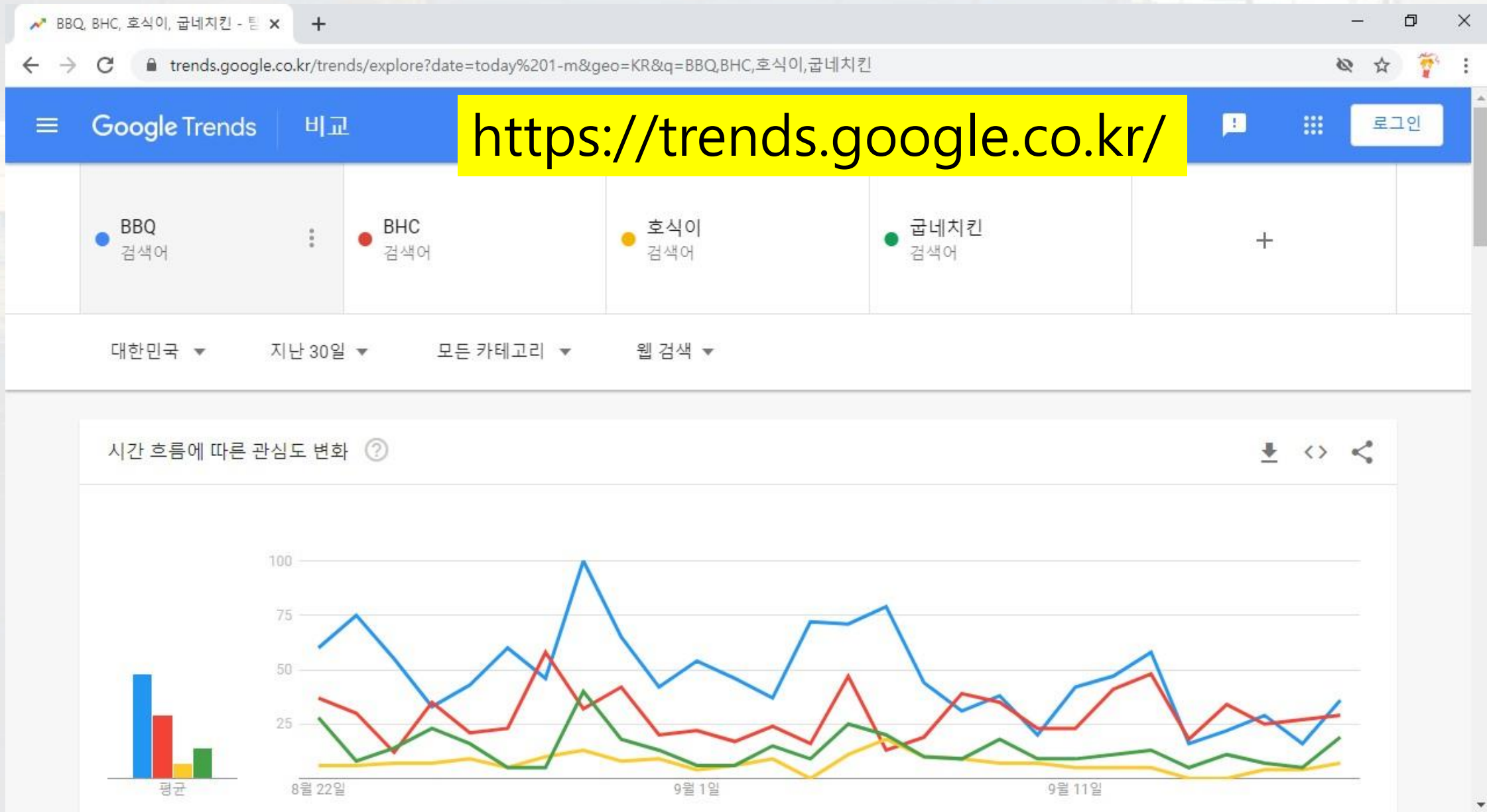
<https://datalab.naver.com/>

데이터랩은 주제별, 기간별, 연령별 등으로
분야별 인기 검색어와 인기분야를 그래프로 추이를 보여 주는 프로그램



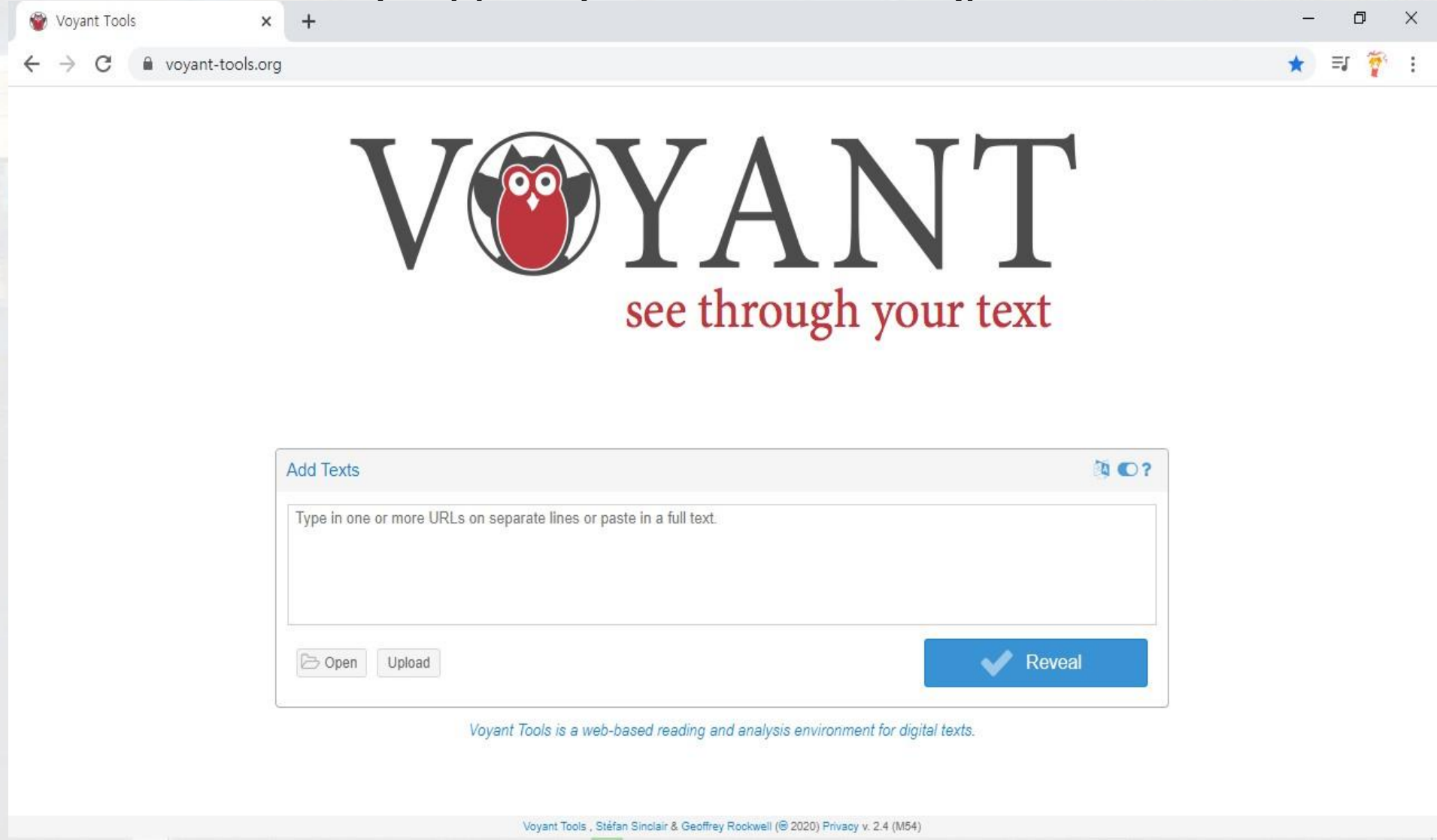
구글 트렌드

구글에서 서비스 중인 검색어나 동영상을 기반으로 하는 빅데이터 분석 서비스
구글을 통해 검색하는 데이터를 수집해서 분석하여 데이터로 보여주는 서비스



Voyant Tools – 빅데이터 기반의 텍스트 분석

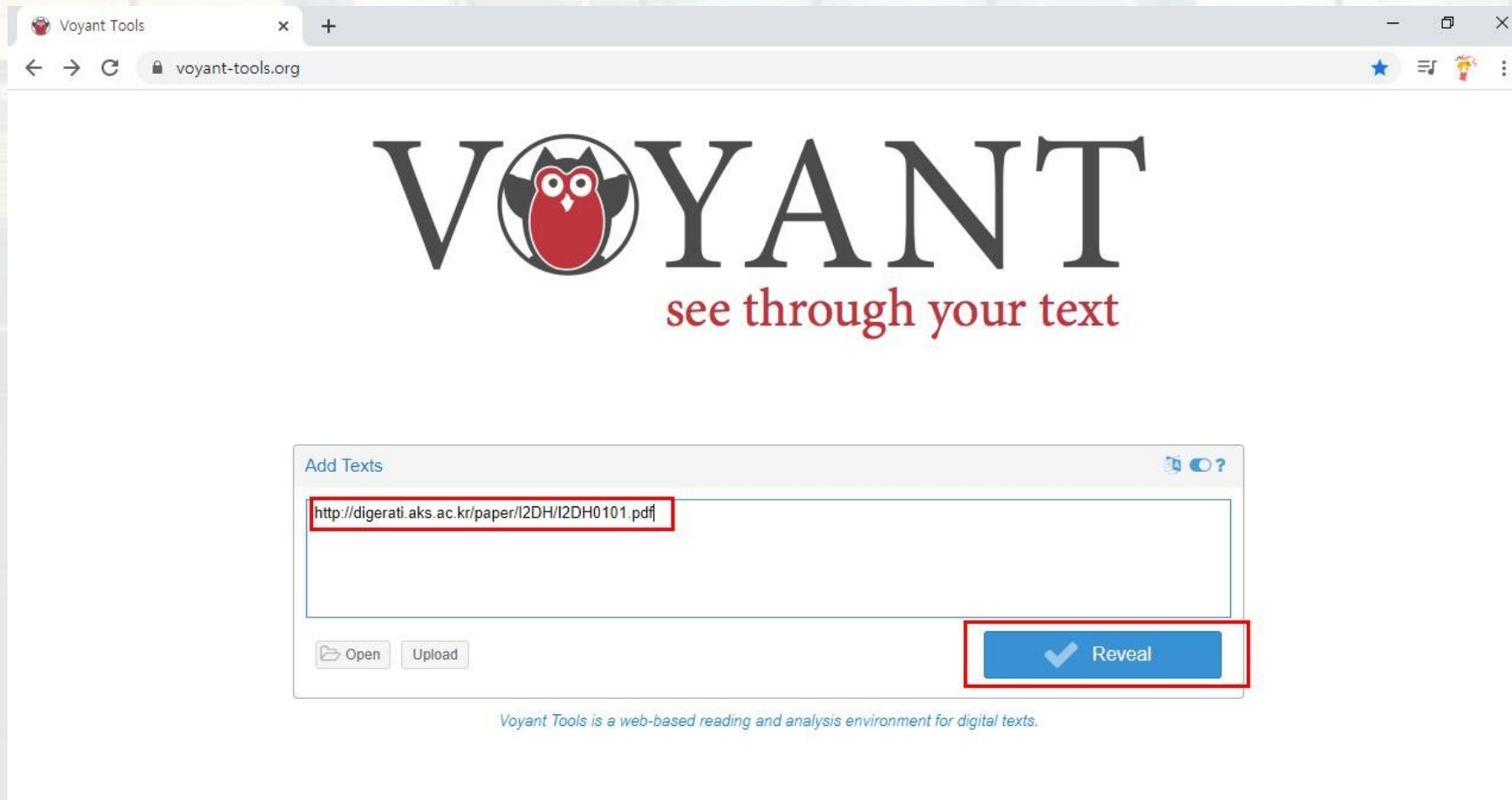
<https://voyant-tools.org/>



Voyant Tools - 온라인 영어 형태소 분석기

<https://voyant-tools.org/>

텍스트 구하기 : <http://digerati.aks.ac.kr/paper/I2DH/I2DH0101.pdf>



Voyant Tools - 온라인 영어 형태소 분석기



공공데이터포털 <https://www.data.go.kr/>

공공데이터포털

data.go.kr

목록등록관리시스템 로그인 회원가입 사이트맵 ENGLISH

DATA 공공데이터포털 .GO.KR

데이터찾기 국가데이터맵 데이터요청 데이터활용 정보공유 이용안내

공공데이터포털은 행정안전부에서 운영하는 공공데이터 통합제공 시스템 대한
민국 정부가 보유한 다양한 공공데이터를 개방하여 누구나 편리하고 손 쉽게 활
용할 수 있게 하는 것을 목적

인기검색어

어떤 공공데이터를 찾으시나요?

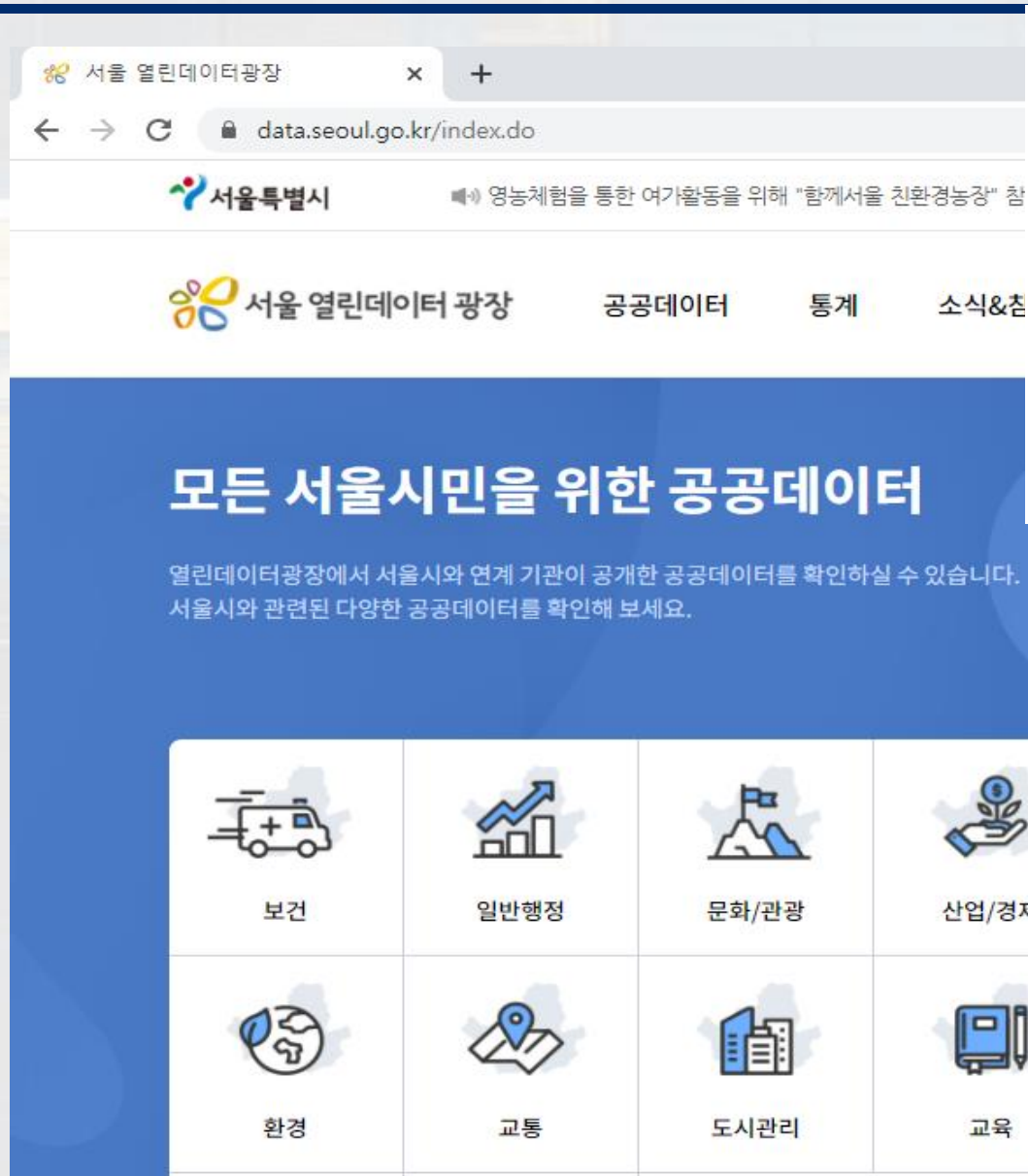
3. 지역화폐

검색조건 분류체계 서비스유형 확장자

① 검색도움말

콘텐츠 바로가기

서울열린데이터광장 <http://data.seoul.go.kr/>



The screenshot shows the Seoul Open Data Plaza website. At the top, there's a browser tab labeled '서울 열린데이터광장' and the address bar shows 'data.seoul.go.kr/index.do'. Below the browser, the Seoul Special City logo is visible. The main navigation bar includes '서울 열린데이터 광장', '공공데이터', '통계', and '소식&참'. The main content area has a blue header with the text '모든 서울시민을 위한 공공데이터'. Below this, a paragraph states: '열린데이터광장에서 서울시와 연계 기관이 공개한 공공데이터를 확인하실 수 있습니다. 서울시와 관련된 다양한 공공데이터를 확인해 보세요.' At the bottom, there's a grid of 10 categories with icons: 보건 (Health), 일반행정 (General Administration), 문화/관광 (Culture/Tourism), 산업/경제 (Industry/Economy), 복지 (Welfare), 환경 (Environment), 교통 (Transportation), 도시관리 (Urban Management), 교육 (Education), and 안전 (Safety).

서울 열린데이터광장

공공데이터 통계 소식&참

모든 서울시민을 위한 공공데이터

열린데이터광장에서 서울시와 연계 기관이 공개한 공공데이터를 확인하실 수 있습니다.
서울시와 관련된 다양한 공공데이터를 확인해 보세요.

보건	일반행정	문화/관광	산업/경제	복지
환경	교통	도시관리	교육	안전

서울열린데이터광장(data.seoul.go.kr)

개인과 가정, 부동산과 재산, 세금과 재정, 기업과 경제, 농림수산
과 천연자원, 제조건설과 개발, 서비스와 생활, 사회보장과 복지,
자연과 환경, 문화와 여가, 통일과 국방, 교통과 물류에 관한 민원
안내 및 신청서비스

데이터셋
7,056

서비스
14,539

Open API
5,512

인기 그룹 데이터
20종 생활이슈

자세히보기 >

뉴스레터
구독안내

Newsletter
열린데이터광장을
신청하세요!

국가통계포털 <https://kosis.kr>

KOSIS국가통계포털(kosis.kr)

국내·국제·북한의 주요 통계를 한 곳에 모아
이용자가 원하는 통계를 한 번에 찾을 수 있도록
통계청이 제공하는 One-Stop 통계서비스

The screenshot shows the KOSIS homepage with a green header and a white search bar. The search bar contains the text "검색어를 입력해주세요." and a magnifying glass icon. Below the search bar, there are several tags: # 인구, # 소비자물가지수, # 1인가구, # GRDP, and # 국내총생산. On the right side, there is a chatbot icon and a button that says "과거에 조회했던 통계를 다시 볼 수 있나요?". Below the search bar, there are three main sections: "추계인구('22)" with a value of 5,163만명, "출생아수('21)" with a value of 260,500명, and "합계출산율('21)" with a value of 0.810명. To the right of these, there is a section titled "물가" (Inflation) with a subtitle "2022년 02월 기준". This section contains two rows: "소비자물가지수" with a value of 3.7% (전년동) and "생활물가지수" with a value of 4.1% (전년동).

KOSIS 국가통계포털
Korean Statistical Information Service

보다나은 정부
대한민국 대정환
한국판뉴딜

국내통계 국제·북한통계 쉽게 보는 통계 온라인

검색어를 입력해주세요.

인구 # 소비자물가지수 # 1인가구 # GRDP # 국내총생산

과거에 조회했던 통계를 다시 볼 수 있나요?

내가본통계표
통계놀이터

추계인구('22) 5,163만명
출생아수('21) 260,500명
합계출산율('21) 0.810명

물가
2022년 02월 기준

소비자물가지수 3.7% (전년동)
생활물가지수 4.1% (전년동)

빅데이터/ 정보와 데이터/ 데이터베이스/ DBMS에 대해 이해하고
빅데이터 활용사례와 데이터를 제공하는 사이트에 대해 학습하였습니다