빅데이터 분석 전문가

04차시

한글 자연어 분석
시각화
데이터베이스

조성진 강사

❑지난 시간 복습

❑데이터 수집해주세요.

- ❍keyword: 빅데이터
- ❍수집데이터: 50개
- ❍추출파일명: my_crawling.csv

❑언어 유형

❍굴절어

- 영어(라틴어), 유럽권, 인도

❍교착어

- 한국, 일본어, 터키어, 터키어

❍고립어

- 현대 영어, 중국어, 베트남어

❍포함어

- 메이저 언어 없음

❑한국어 분석의 어려움

❍고립어(굴절어)인 영어로 감사합니다의 출현 정도를 알려면?

- Thank you, Thanks의 반복 횟수만 알면 됨.

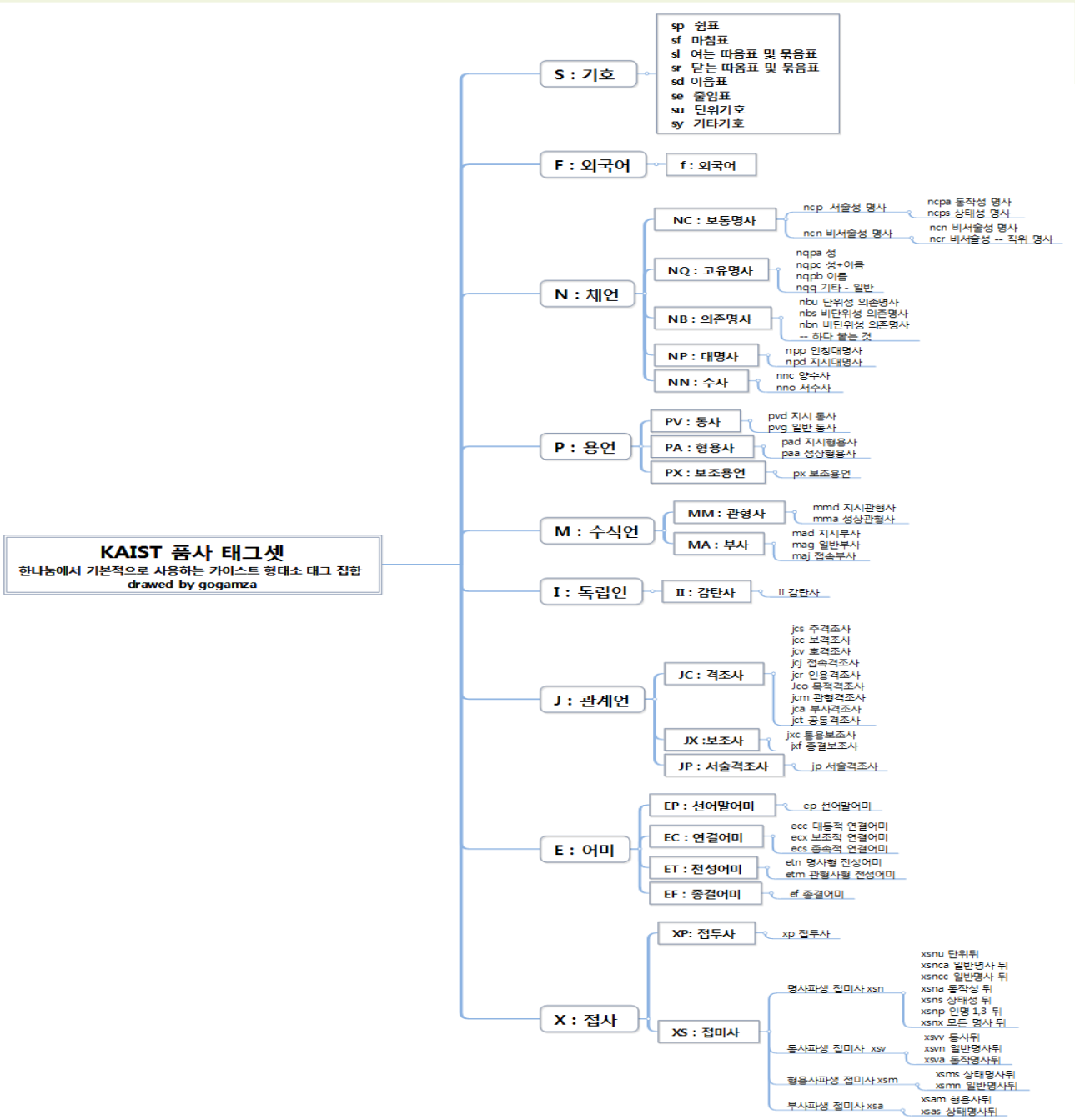
❍교착어인 한국어로 감사합니다의 출현 정도를 알려면?

- 감사+하+ㅂ니다 의 형태로 형태소 분석이 선행되야함.
- 감사라는 단어의 출현 빈도 로는 찾을 수 없음
- ex) 국정 감사, 외부 감사, 내부 감사, 감 사먹자

분석해서 뭐 할껀데??

	noun	n
1	데이터	31
2	규제혁신	9
3	개인정보	7
4	산업	7
5	우리	7
11	서비스	4
12	정부	4
13	결합	3
14	규제	3
15	나라	3
16	데이터경제	3
17	빅데이터	3
18	신기술	3

보기에 어떠신가요?



출처: https://github.com/haven-jeon/KoNLP/blob/master/etcs/KoNLP-API.md

❑데이터 시각화(Data Visualization)

- ❍데이터를 그림이나 그래프를 통해 시각적으로 표현하는 모든 과정
- ❍단순히 예쁘거나 멋있어서 하는 작업이 아님
- ❍데이터를 쉽게 이해할 수 있게 도와줌

❑데이터 시각화(Data Visualization) 분류

❍정보 시각화(Information Visualization)

- 보통 대규모 데이터를 색체, 통계(도표, 그래프 등), 이미지 등을 활용해 요약적으로 표현하는 것을 의미

❍과학적 시각화(Scientific Visualization)

- 실험결과나 시뮬레이션 데이터 등 복잡한 데이터를 쉽게 탐색할 수 있도록 3차원 그래픽 기술 등을 활용하여 시각화하는 기술

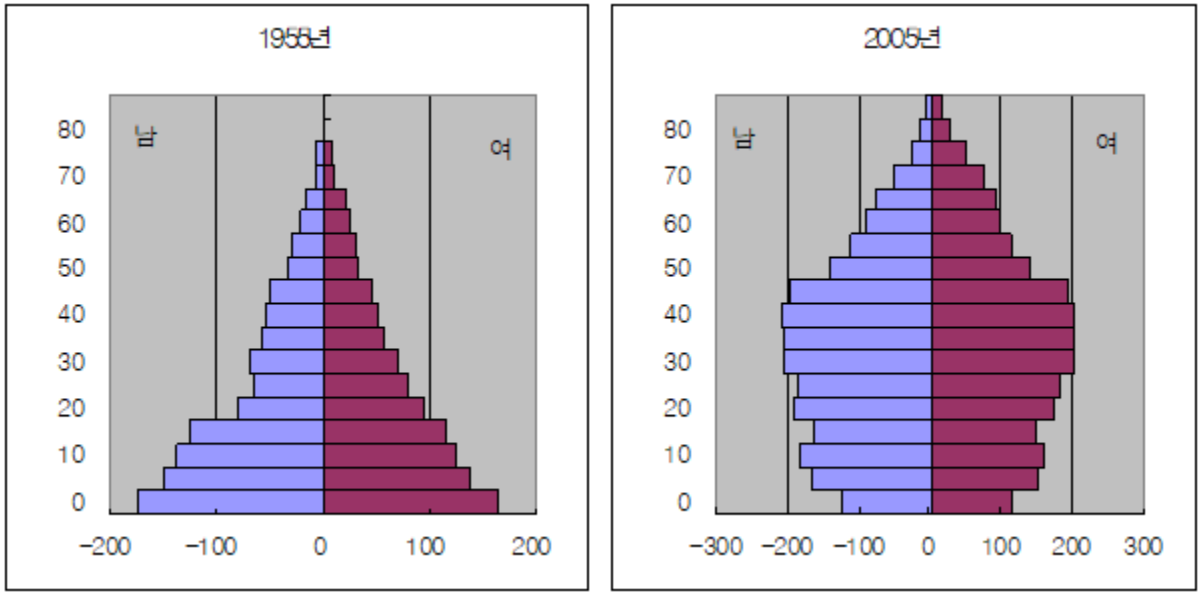
❍인포 그래픽(Information Graphic)

- 인포메이션과 그래픽의 합성어로 복잡한 수치나 글로 표현되어 있는 다량의 정보를 차트, 지도, 다이어그램, 로고 등을 활용하여 한눈에 파악 할 수 있도록 하는 디자인

❑데이터 시각화(Data Visualization) 이유

❍정보전달의 효율성

- 인간은 정보를 받아 들일 때 시각에 대한 의존도가 높음
- 인간은 패턴화 된 자료를 받아들일 때 높은 인식 속도를 보임



대한민국 인구 피라미드

❑데이터 시각화(Data Visualization) 이유

❍새로운 발견

– 데이터로는 보이지 않던 자료들이 시각화 이후 보이게 됨

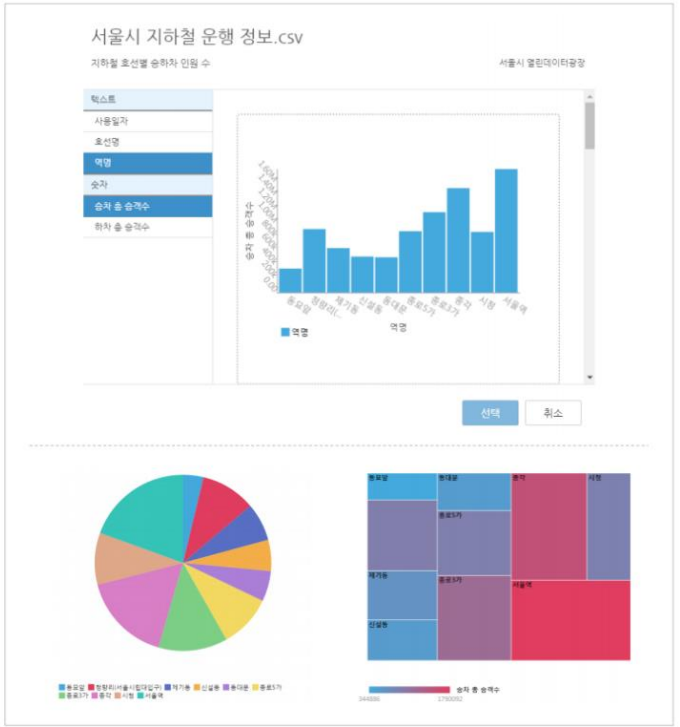
A	B	C	D	E
사용일자	호선명	역명	승차 총 승객수	하차 총 승객수
2017-03-31	1호선	동묘앞	8594	9569
2017-03-31	1호선	청량리(서울시립대입구)	29781	31052
2017-03-31	1호선	제기동	20393	20432
2017-03-31	1호선	신설동	19214	18900
2017-03-31	1호선	동대문	16343	19045
2017-03-31	1호선	종로5가	30488	30931
2017-03-31	1호선	종로3가	35751	33902
2017-03-31	1호선	종각	51676	50327
2017-03-31	1호선	시청	28314	29207
2017-03-31	1호선	서울역	62986	62036
2017-03-30	1호선	동묘앞	10432	11162
2017-03-30	1호선	청량리(서울시립대입구)	30637	31369
2017-03-30	1호선	제기동	22767	22931
2017-03-30	1호선	신설동	18675	18373
2017-03-30	1호선	동대문	16457	18624
2017-03-30	1호선	종로5가	30808	31242
2017-03-30	1호선	종로3가	36939	33812
2017-03-30	1호선	종각	52312	50132
2017-03-30	1호선	시청	28214	28743
2017-03-30	1호선	서울역	55322	54630
2017-03-29	1호선	동묘앞	10400	11178

열 (Attribute)

셀 (Cell)

행 (Item)

▲ 데이터 테이블의 형식적 구조, 서울시 지하철 운행 정보(데이터 출처 : 서울 열린데이터 광장)



▲ 열 선택 및 조합에 따른 시각화 예 (데이터 시각화 솔루션 DAISY 활용)

❑Database

- ❍체계화된 데이터의 모임
- ❍여러 사람이 공유하고 사용할 목적으로 통합 관리되는 정보의 집합
- ❍데이터의 중복을 없애고 구조화해 저장
- ❍데이터를 필요할 때 사용할 수 있도록 저장하는 공간
- ❍데이터를 지속적으로 관리하고 보호하는 것이 주목적

❑DBMS: DataBase Management System

❑데이터베이스를 관리하는 시스템

- ❍데이터의 추가/조회/변경/삭제
- ❍데이터의 무결성 유지
- ❍트랜잭션 관리
- ❍데이터의 백업 및 복원
- ❍데이터의 보안