KOREA

DISCOVERY SUMMIT

EXPLORING DATA
INSPIRING INNOVATION

김광희, 옥창수, 정순우 (삼성전기)

비정형 Text Data 시각화

Python Crawling & JMP Text Explorer

옥창수 (품질 관리 기술사)
정순우 (품질 관리 기술사 필기합격)
김광희 (Big Data 분석 전문가)

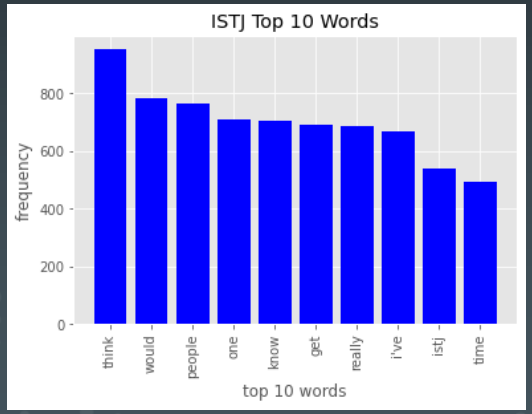
Content

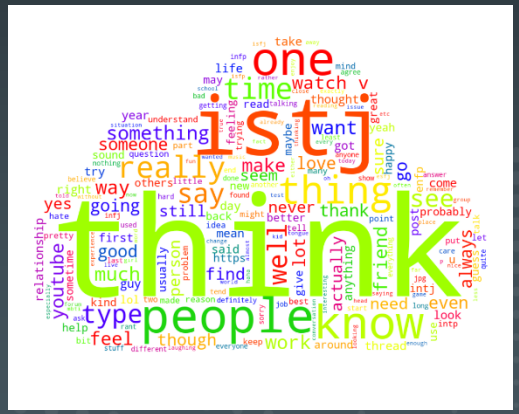
1. TEXT 분석 정의와 절차
2. Data gathering 과 처리
3. 시각화

1. TEXT 분석 정의와 절차

Text 분석이란

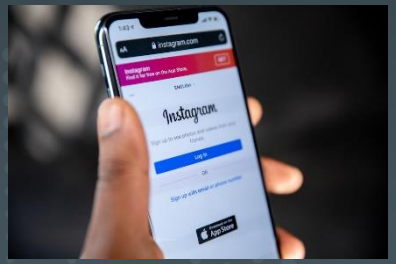
- 대량의 비정형 Text Data를 AI 또는 언어 처리 기술을 활용하여 패턴과 인사이트를 도출하는 분석 방법

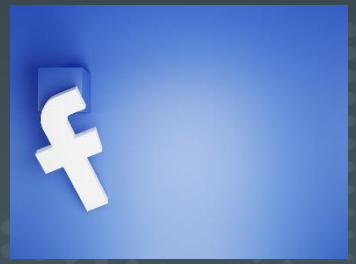




Text 분석의 의미

- Text Analysis : Text를 읽고 인사이트를 얻기 위함
- Text 정보 분류, 추출 → 감정, 패턴, 관련성 등을 파악
- E-Mail, SNS, 제품리뷰 등 다양한 Text 데이터를 활용




Text 분석의 중요성

- 기업은 다양한 비정형 데이터에서 인사이트를 도출
- 효과적인 의사결정을 위해 SNS, Blog, News 데이터 분석
- 비정형 Data를 인간이 직접 처리하는데 한계가 있음

Text 분석의 필요성

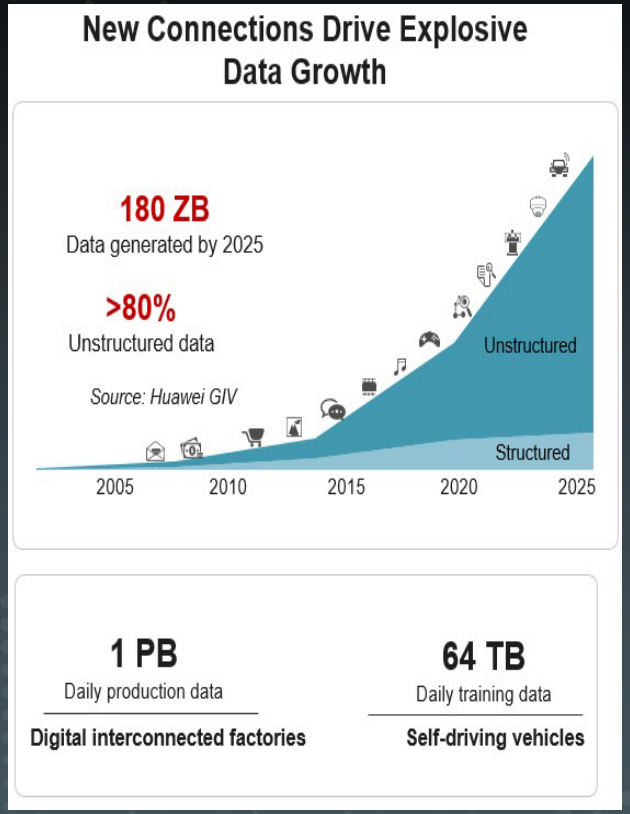
2020년 50억 명 인터넷 사용

세계 인구의 59%

49%가 SNS 활용

SNS, Blog에 대량 Text 데이터 생성

전체 Data의 80%가 비정형 Data



Text 분석의 어려움

Text Data는 Unstructured Data (비정형)

Text Data는 다양한 Platform에 분산

Text Data를 수집, 처리하려면 전문적인 지식 필요

(Crawling, Deep Learning, Coding)

Text 분석의 필요성

진정한 고객 Needs 파악 가능

신속한 의사결정 가능

비용 절감

★ **대부분의 기업은 고객이**
원하지 않는 물건을 만들어 망한다.

Text 분석 종류

감정분석 : 서비스 피드백 및 문제점 파악, 소비자 이해

주제 탐색 : 방대한 양의 데이터에서 중요한 키워드 파악

문서 분류 : 스팸메일 분류, 수신 메일 주제별로 구분

Text Data 분석 단계

1. Data Gathering (Data 수집)

- 내부: 기업 내부의 자료 활용
- 외부: 블로그, 뉴스, 리뷰, 댓글 파악 (Crawling을 통한 수집 필요)

2. Data Preparation (Data 준비)

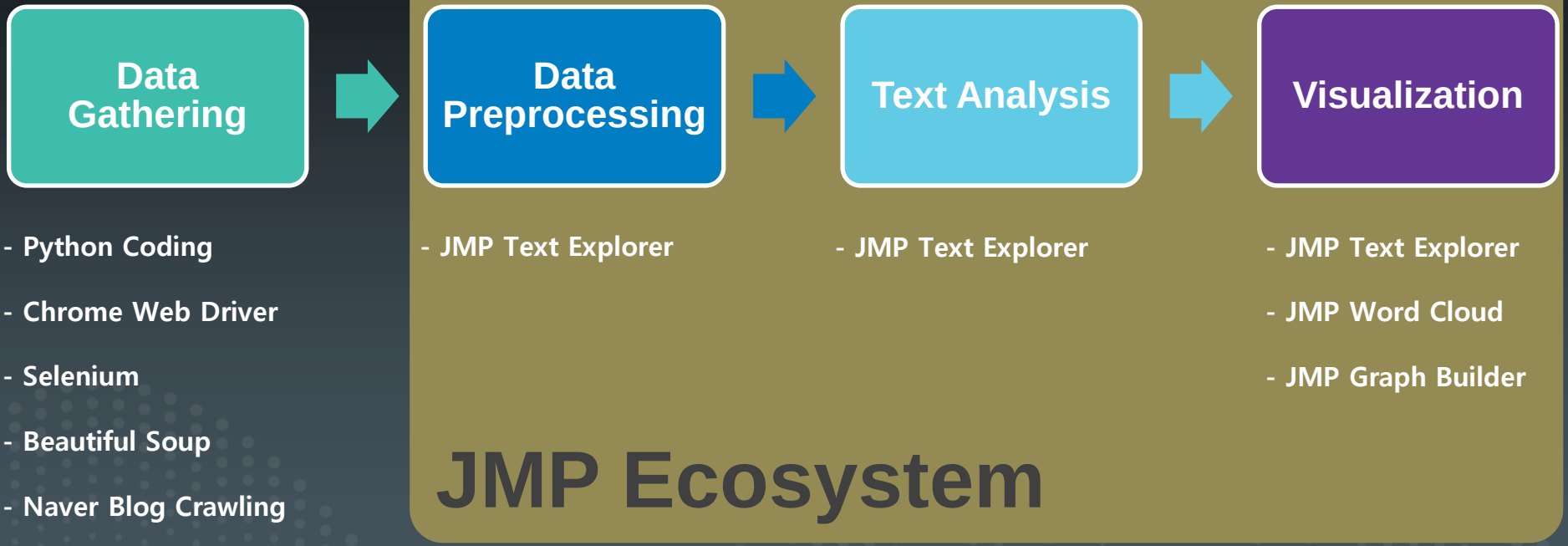
- JMP Explorer 또는 Coding을 통한 Preprocessing (토큰화, 품사태깅, 형태소 분석, 불용어 제거)

3. Text Analysis

- 키워드 추출, 문서요약, 감성분석, 토픽 분석, 군집화

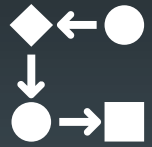
4. 시각화

Text Data 분석 Process



Why JMP Explorer?

JMP 17 EA에 Text Explorer 한글 지원



KonIpy와 같은 별도의 Java 환경 Setup 불필요

Code 없이 대량의 비정형 Text Data 분석 가능

JMP로 Text 분석하는 것이 경제적

2. Data gathering 과 처리

Data Gathering

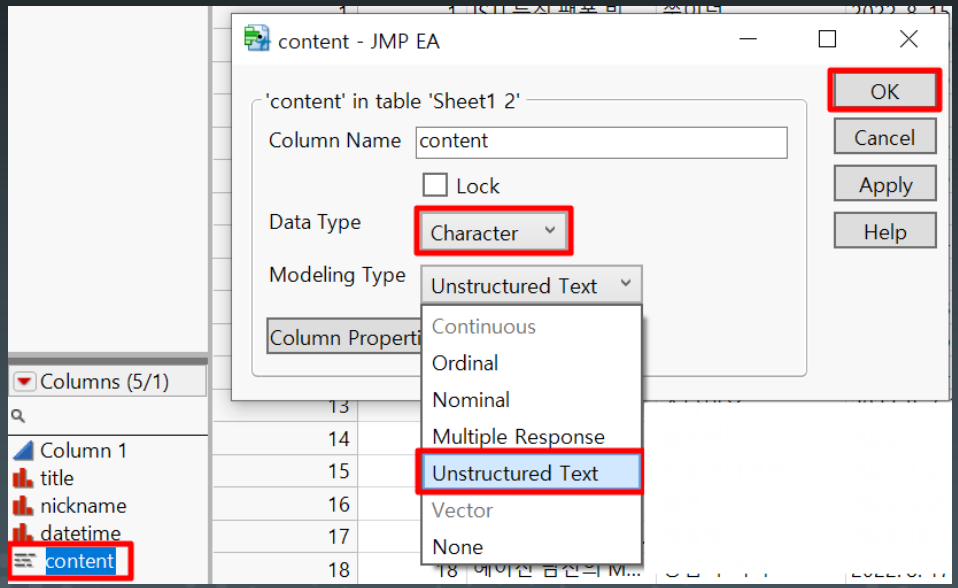
- Google Colaboratory 활용
- Blog 상위 150개 URL을 Crawling
- for 문으로 URL 접속 후 블로그 본문 Crawling 및 Excel로 저장

No.	URL
0	https://blog.naver.com/mjeez/222886513215
1	https://blog.naver.com/qmfosej/222859446821
2	https://blog.naver.com/endend5233/222884489976
3	https://blog.naver.com/sonafox/222845213401
4	https://blog.naver.com/cosreader/222851550703
...	...
145	https://blog.naver.com/yogurt0216/222850153591
146	https://blog.naver.com/aracho25/222877601056
147	https://blog.naver.com/audwn4261/222899583568
148	https://blog.naver.com/hyobbang99/222864874684
149	https://blog.naver.com/yeji2552/222900995687

Data Preprocessing & Text Analysis

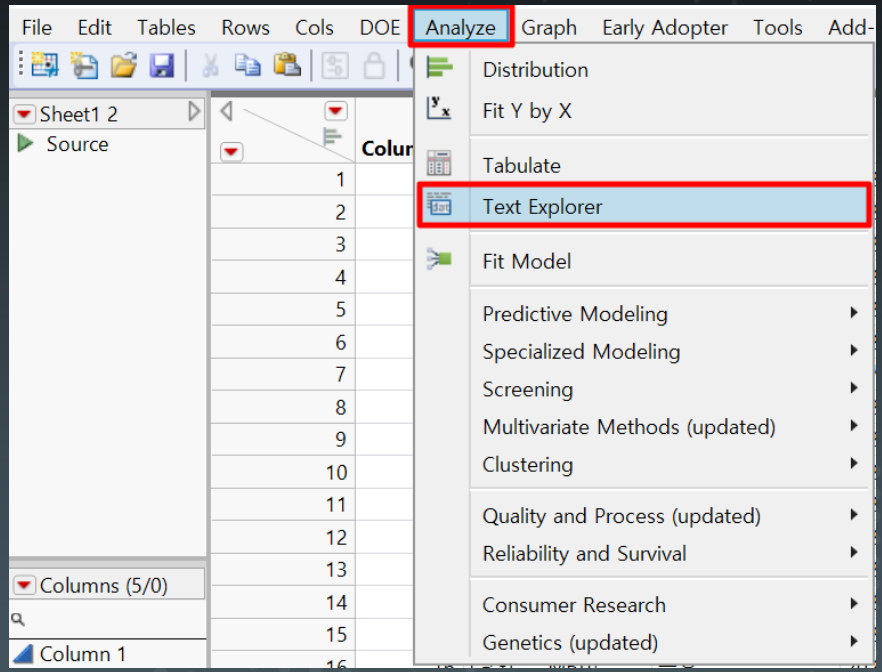
- JMP 17에 데이터 복사 (16 버전 이하 한글 처리 불가)
- Column Info에서 Modeling Type 변경

  Nominal → Unstructured Text



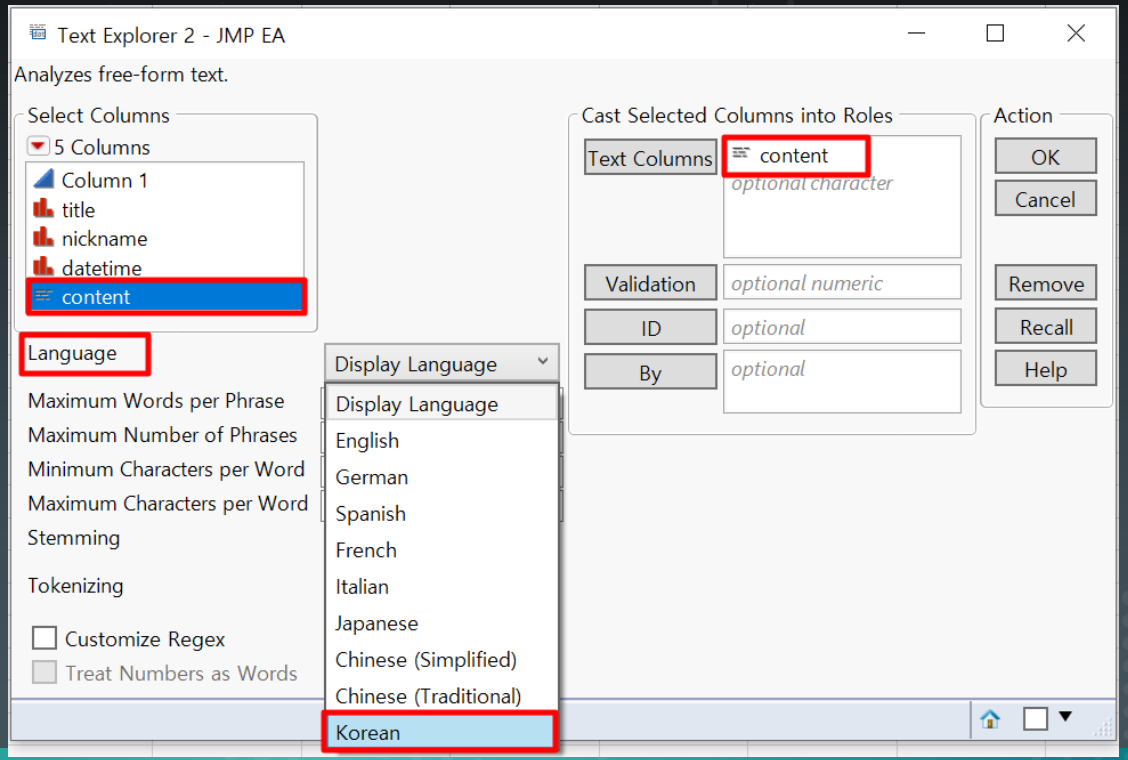
Data Preprocessing & Text Analysis

- Analyze > Text Explorer 선택



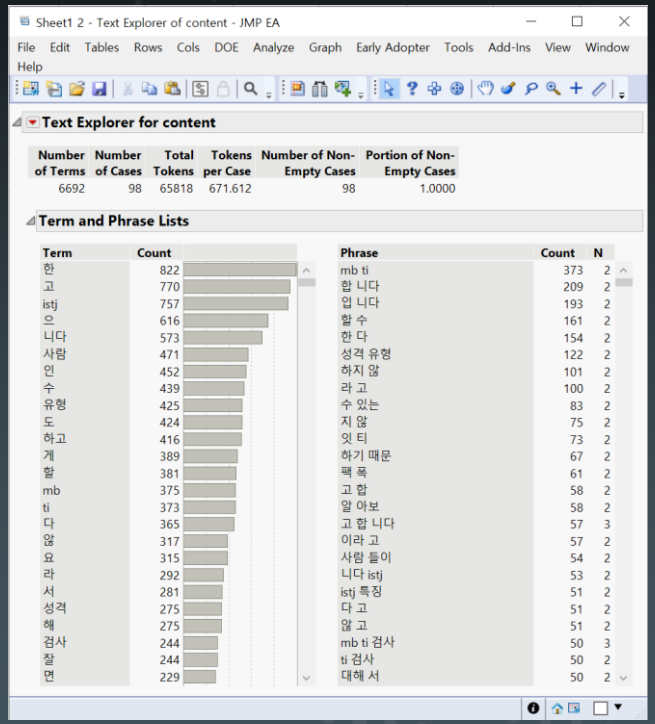
Data Preprocessing & Text Analysis

- Text Columns 선택
- Language 선택
- 'OK' 클릭



Data Preprocessing & Text Analysis

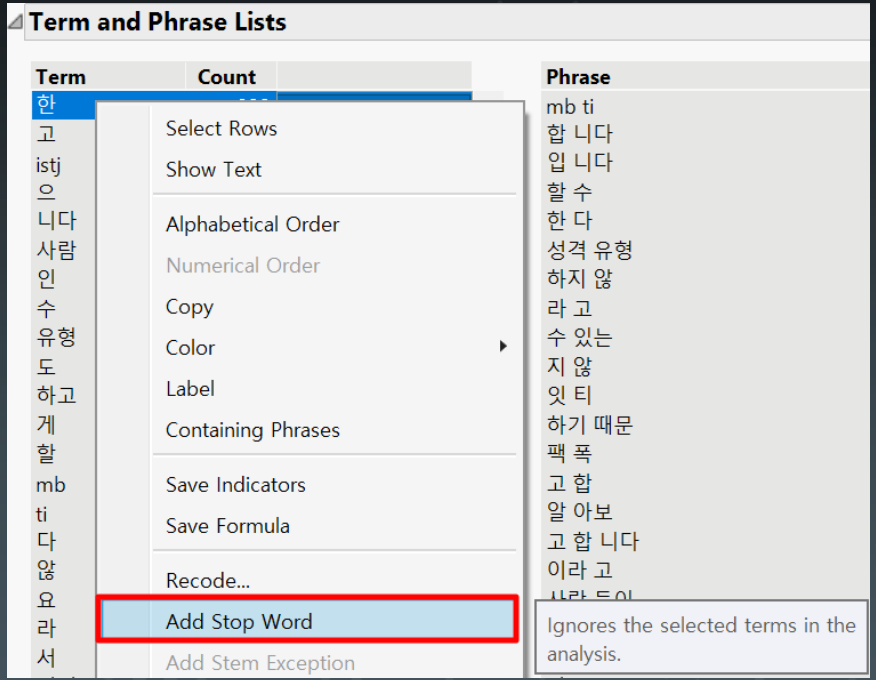
- Text Explorer 실행



Data Preprocessing & Text Analysis

- 불용어 선택 : Add Stop Word

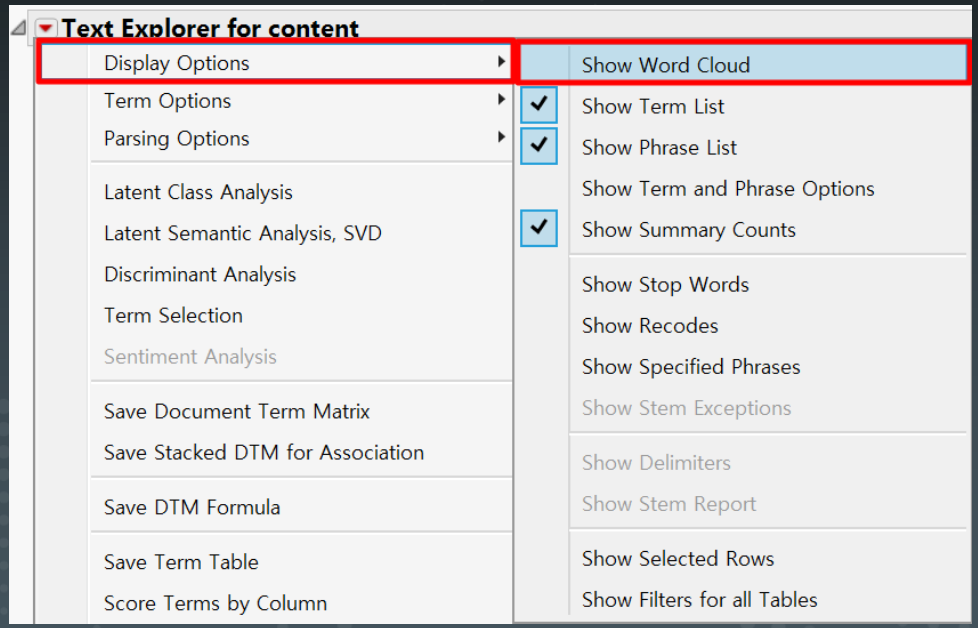
  Shift를 누른 후 중복 선택 가능



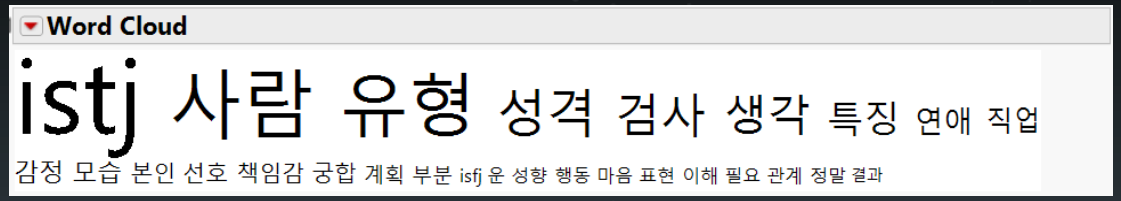
3. 시각화

Visualization (Word Cloud)

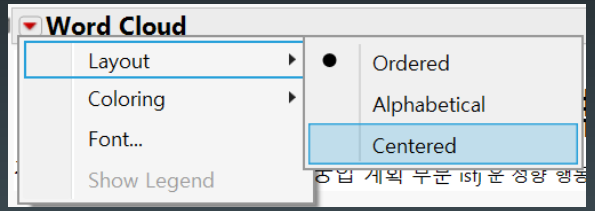
- Display Options > Show Word Cloud

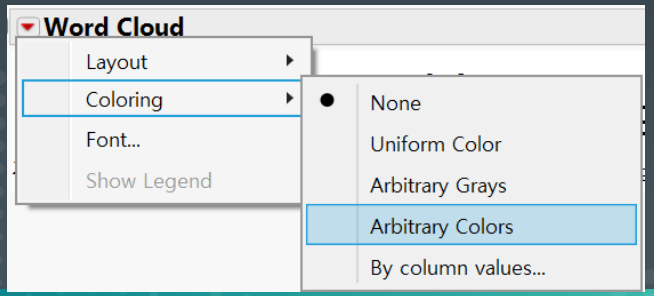


Visualization (Word Cloud)



- Word Cloud > Layout > Centered
- Word Cloud > Coloring >

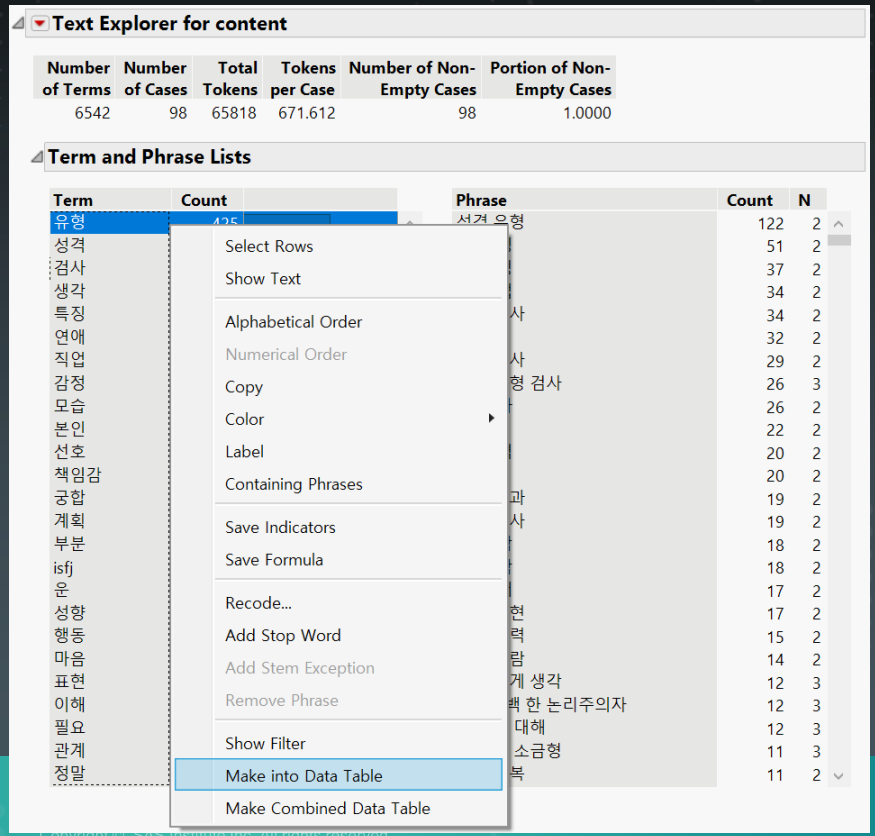




Visualization (Word Cloud)

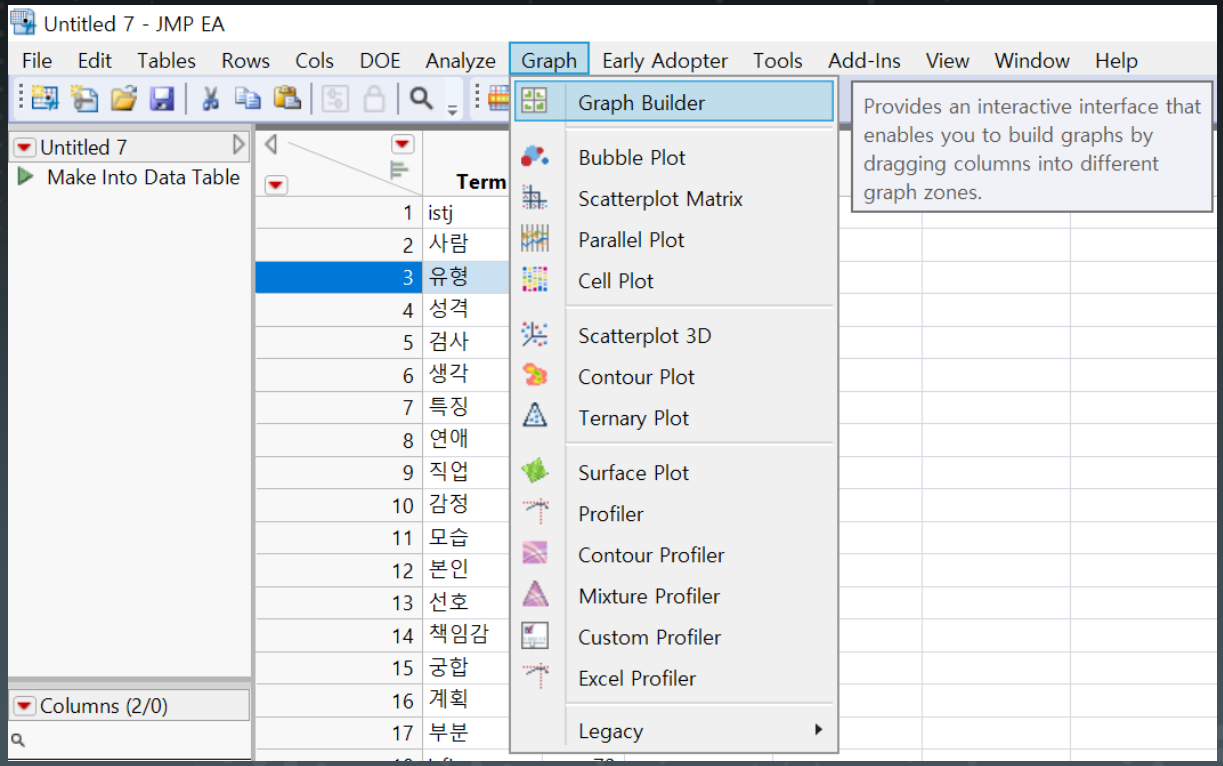
Visualization (Graph Builder)

- Make into Data Table



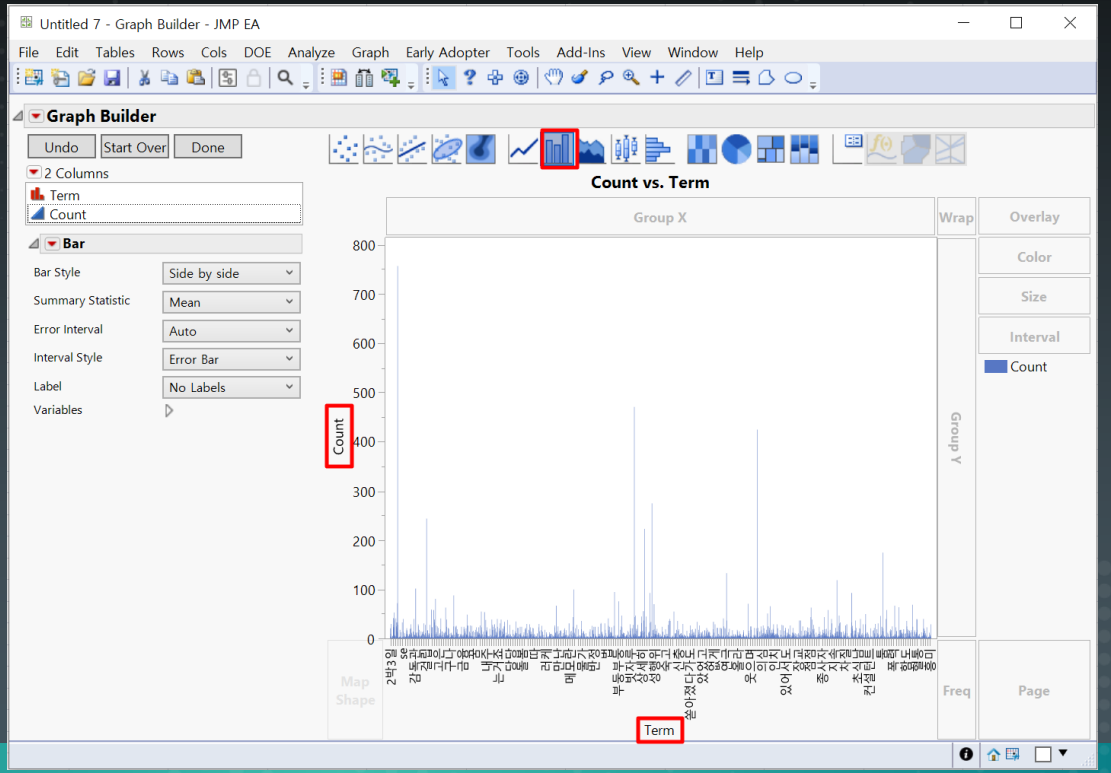
Visualization (Graph Builder)

- Graph > Graph Builder



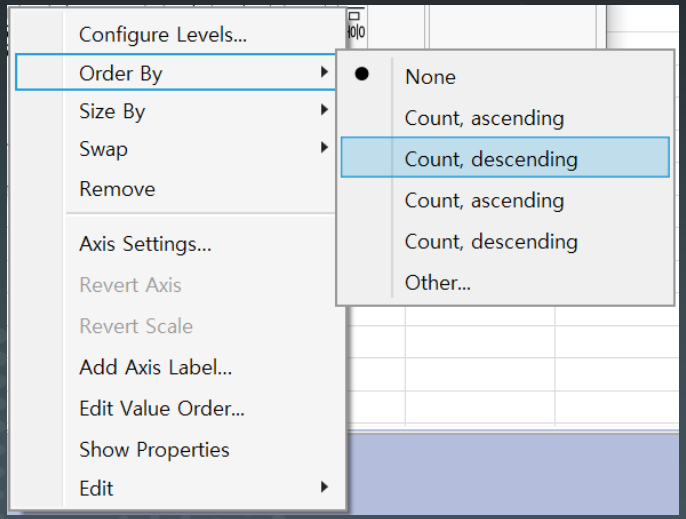
Visualization (Graph Builder)

- X : Term, Y: Count
- Graph : Bar



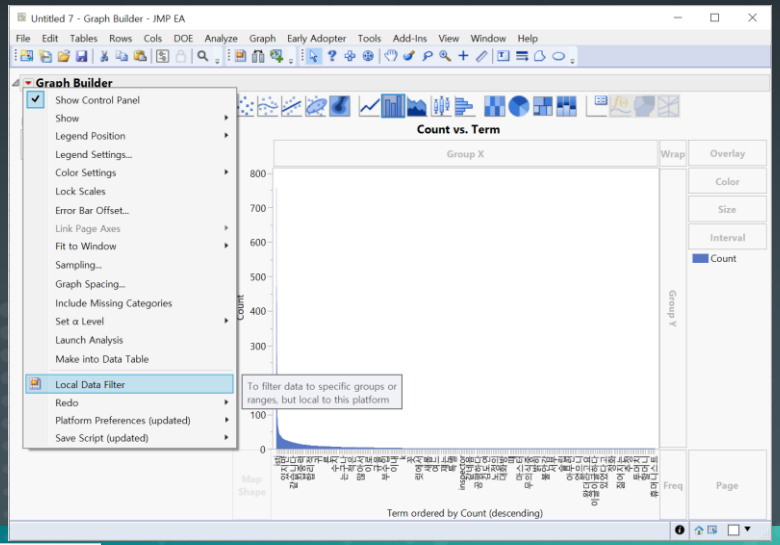
Visualization (Graph Builder)

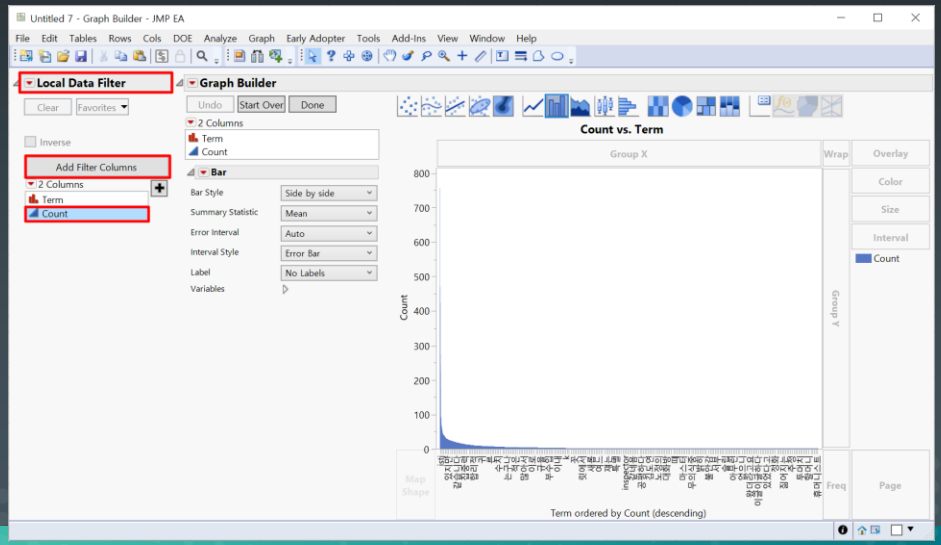
- X축에 마우스 우 클릭 > Order By > Count, descending



Visualization (Graph Builder)

- Graph Builder > Local Data Filter
- Local Data Filter > Add Filter Columns > 'Count'





Visualization (Graph Builder)

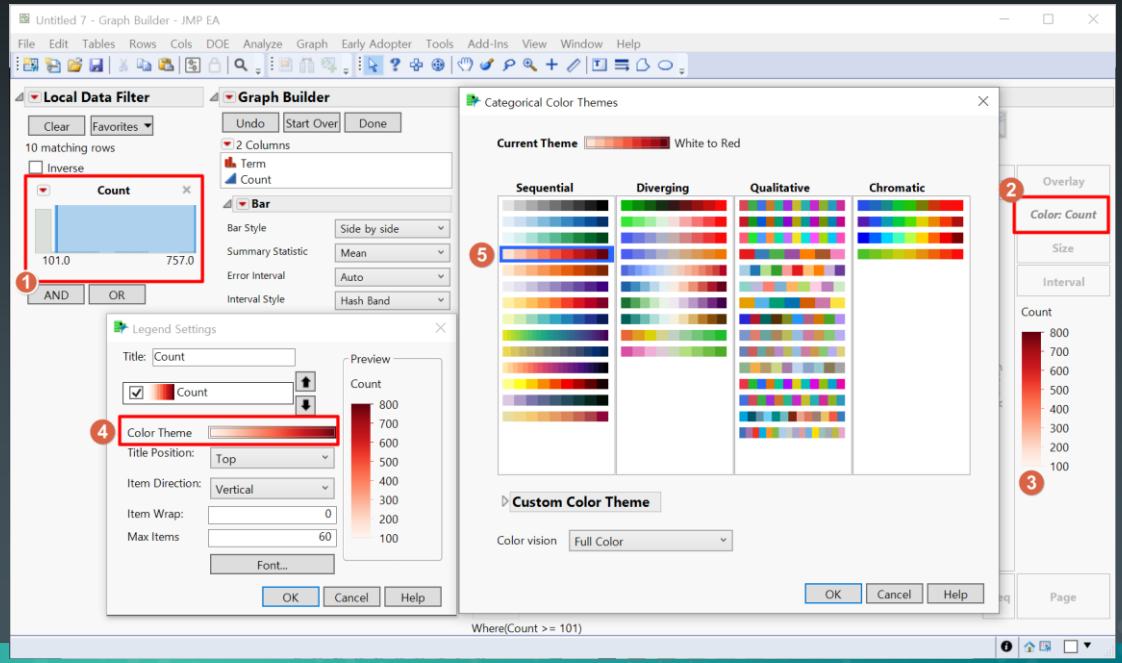
① Count의 값을 수정하여 Term 수를 변경

② 'Count'를 Color에 Drag & Drop

③ Count에 마우스 더블 클릭

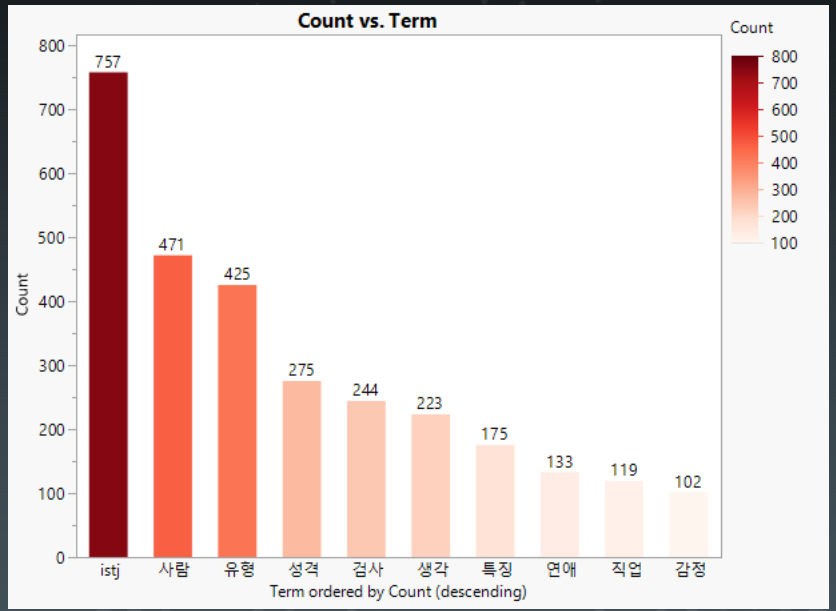
④ Color Theme 클릭

⑤ Color Theme 선택



Visualization (JMP 최종결과)

[Word Cloud (상위 100개 항목)]



[Bar Chart (상위 10개 항목)]

KOREA

DISCOVERY SUMMIT

EXPLORING DATA
INSPIRING INNOVATION