Monthly User Guide from JMP Korea

제 29호 (2019년 12월)

Text Explorer

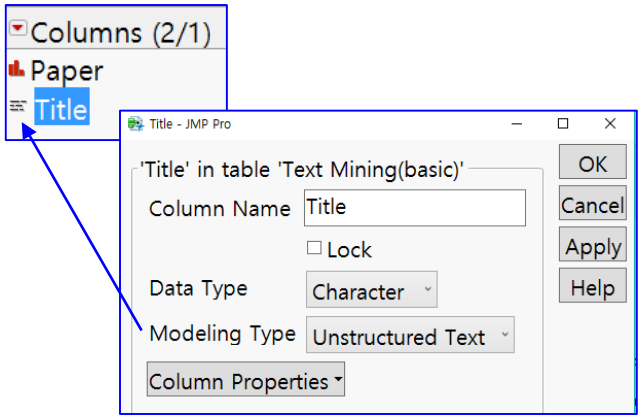
* 본 Guide 의 내용과 관련한 문의는 ikju.Shin@jmp.com 으로 연락 바랍니다

** Monthly Guide 전체 내용(지난 호 포함)은 아래 Site에서 확인 가능합니다
(https://www.jmp.com/ko_kr/newsletters.html)

1. Text Explorer 소개

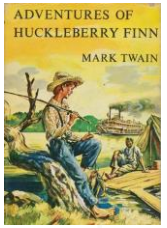
이번 호에서는 비정형 텍스트(Unstructured Text) Data에 대한 분석을 하는 Text Explorer에 대해 알아보도록 하겠습니다. Field Claim, 고객 상담 기록, 시/소설, Survey 결과, 사건 기록 등 다양한 형태의 비정형 Text 데이터가 있을 수 있겠습니다.

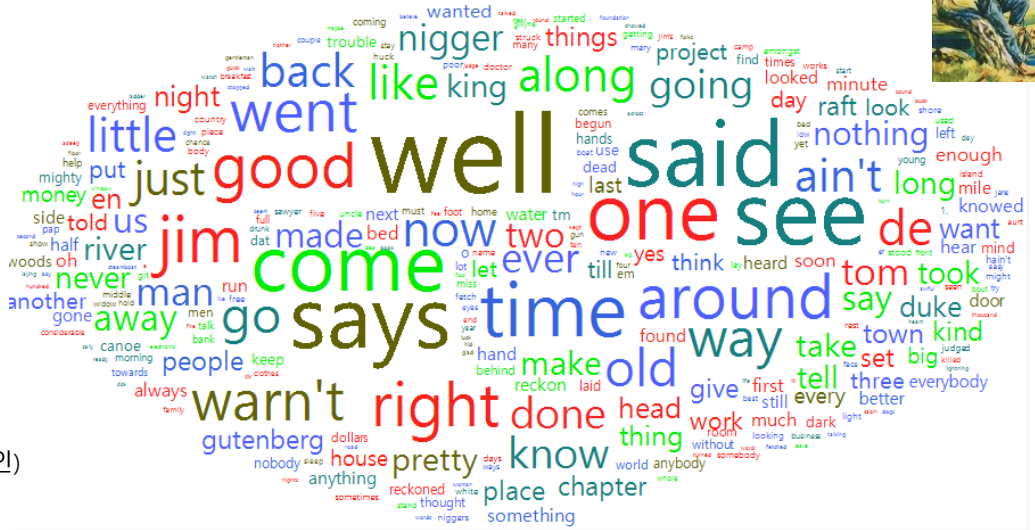
JMP에서의 Text Data 인식 (Column Info)



(Column 명 위에서 우측 마우스 클릭 / Column info에서 확인)

Text Explorer 방법 중의 하나인 Word Cloud 예시*





* 출처 : Gutenberg Project(https://www.gutenberg.org)에서 Adventures of Huckleberry Finn

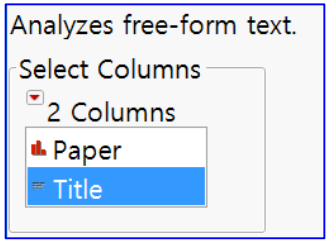
2. 기본적인 분석 Process 및 방법

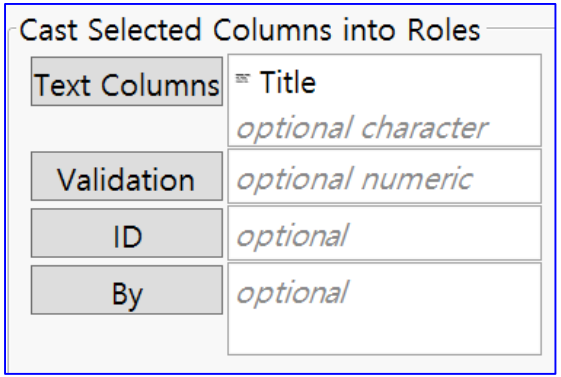
간단한 Sample Data로 JMP를 이용한 기초적인 Text Mining 방법에 대해 설명 드립니다.

1. 아래와 같은 Text Data 가 있다고 가정

	Paper	Title
1	A	I am a boy.
2	A	I am handsome.
3	A	You are a girl.
4	A	You are beautiful.
5	B	She is girl, beautiful.
6	B	He is boy, handsome.
7	B	Look at the Picture.
8	B	They are boys.
9	B	They are girls.
10	A	They are handsome boys.
11	A	They are beautiful girls.
12	A	Big data
13	C	Big data analysis
14	C	data analysis
15	A	Statistical analysis
16	C	Text analysis
17	C	
18	C	K is boy
19	C	M is girl
20	B	J is data

2. (일단 다른 Option은 그대로 두고)
'Title' Column 을 'Text Columns' 에 선택, OK 클릭





2. 기본적인 분석 Process 및 방법

3. Term and Phrase list 확인
-Term : 하나의 단어(단위), 조사, 접속사 등 불용어(Stop words) 제외함
(Terming : 관사 등 의미 없는 단어를 제거하는 것)
-Phrase(구) : 두 개 이상의 단어로 구성된 의미 단위
(Phrasing : 예) big, data → big data)

	Paper	Title
1	A	I am a boy.
2	A	I am handsome.
3	A	You are a girl.
4	A	You are beautiful.
5	B	She is girl, beautiful.
6	B	He is boy, handsome.
7	B	Look at the Picture.
8	B	They are boys.
9	B	They are girls.
10	A	They are handsome boys.
11	A	They are beautiful girls.
12	A	Big data
13	C	Big data analysis
14	C	data analysis
15	A	Statistical analysis
16	C	Text analysis
17	C	
18	C	K is boy
19	C	M is girl
20	B	J is data

Empty case

Text Explorer for Title

Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-Empty Cases	Portion of Non-Empty Cases
16	20	60	3	19	0.9500

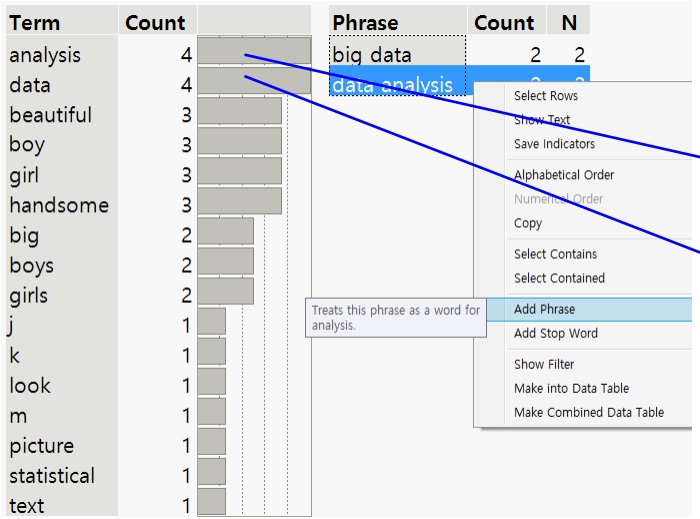
Term and Phrase Lists

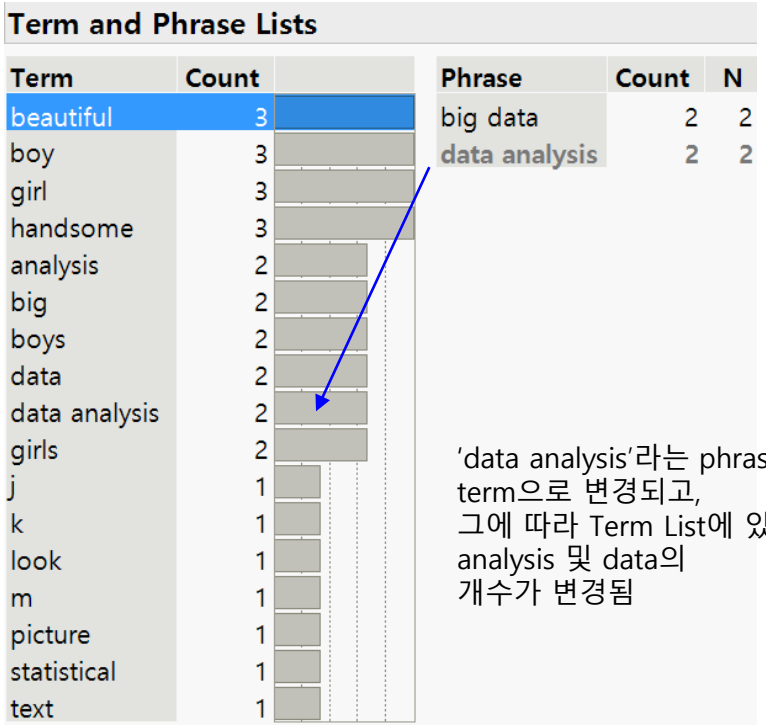
Term	Count
analysis	4
data	4
beautiful	3
boy	3
girl	3
handsome	3
big	2
boys	2
girls	2
j	1
k	1
look	1
m	1
picture	1
statistical	1
text	1

Phrase	Count	N
big data	2	2
data analysis	2	2

2. 기본적인 분석 Process 및 방법

4. 특정한 Phrase를 하나의 단위(Term)로 하고자 한다면
해당 Phrase를 선택 후 우측 마우스 클릭 후 add phrase 선택
→ 해당 phrase 가 좌측 Term list로 이동됨

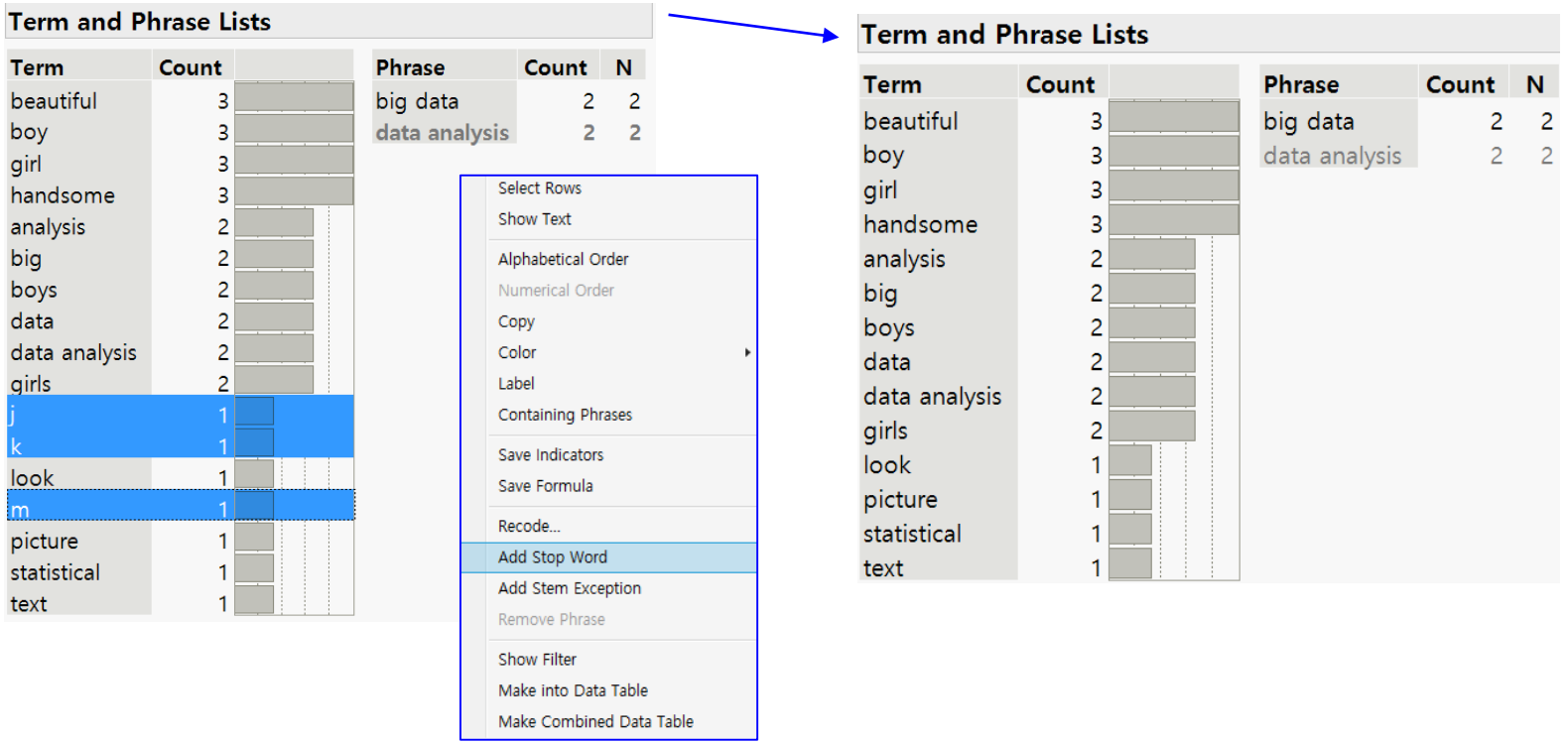




'data analysis'라는 phrase가 term으로 변경되고,
그에 따라 Term List에 있던 analysis 및 data의 개수가 변경됨

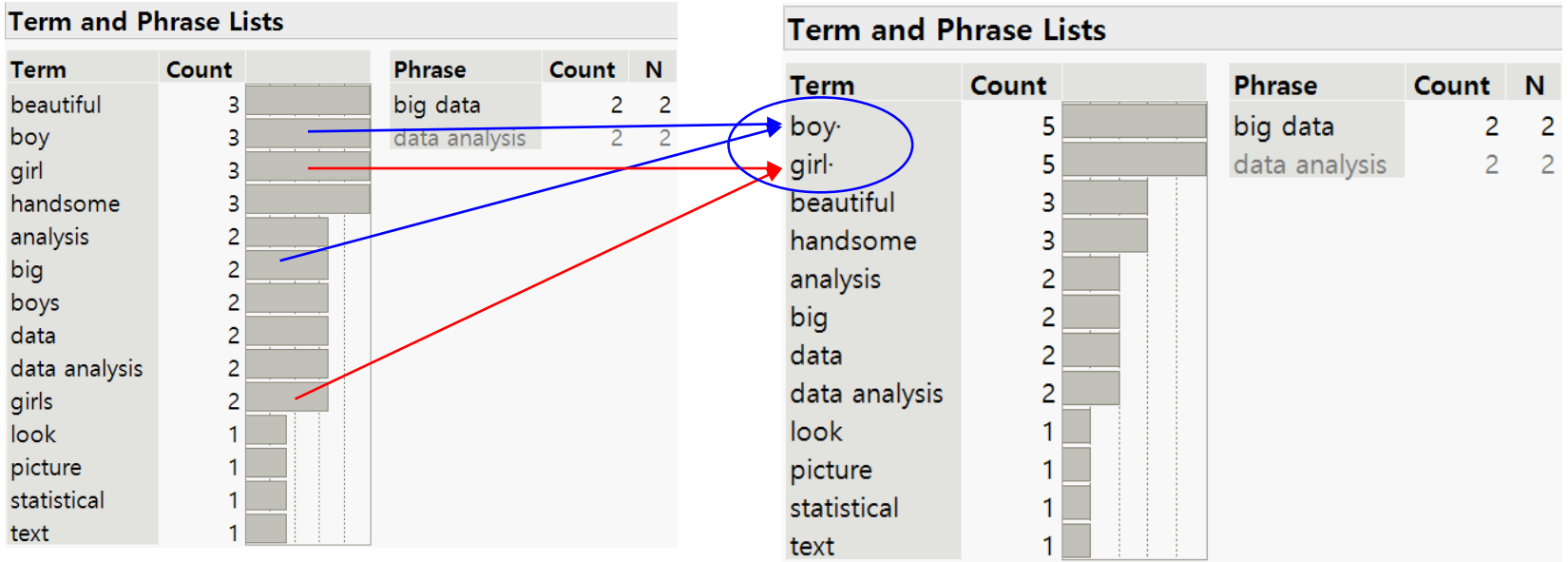
2. 기본적인 분석 Process 및 방법

5. 불용어(stopword) 선택
: 의미가 없다고 판단되는 단어(term)를 선택한 후 우측 마우스 클릭 후
'Add Stop Word' 클릭하면 Term list에서 사라짐

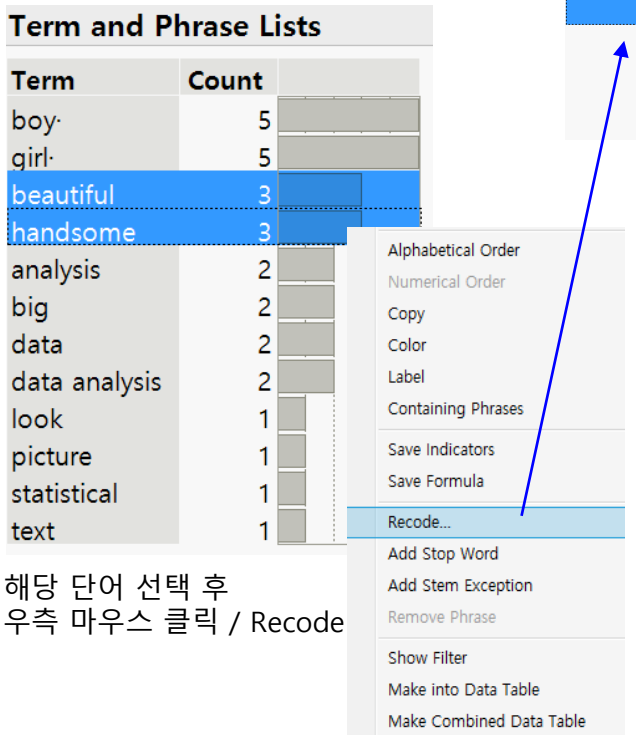


2. 기본적인 분석 Process 및 방법

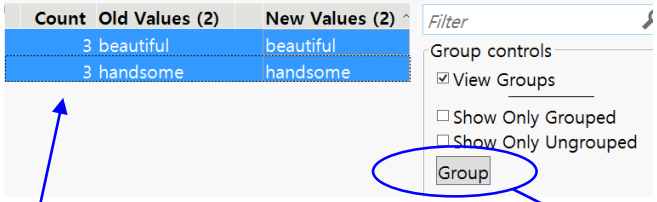
6. Stemming(어간 추출) : 어간만을 추출하여 하나의 Term화하고자 한다면 아래와 같이 하면 됨
   Term Options / Stemming / Stem for Combining
   Stemming : 동일한 기본단어로 결합(변경)하는 것을 말함
   예) processes, process, processing → process



7. 다른 단어(term)를 같은 뜻의 단어로 재구성하고자 한다면 아래와 같이 실행

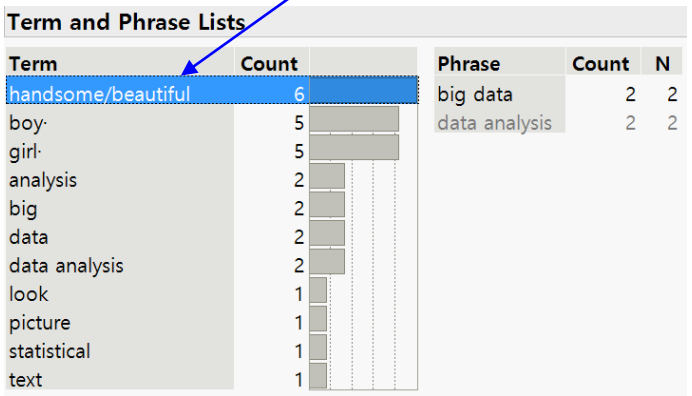


해당 단어 선택 후
우측 마우스 클릭 / Recode



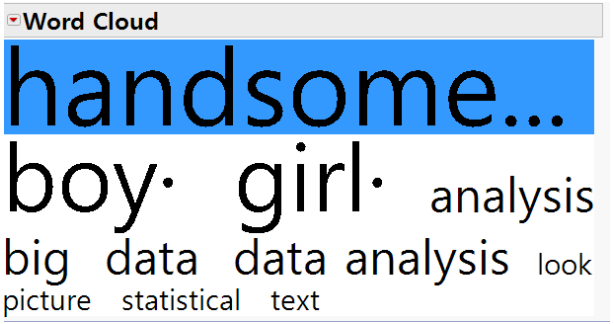
해당 단어 선택 후
'Group' 버튼 클릭

New Values에 새로운
단어 입력

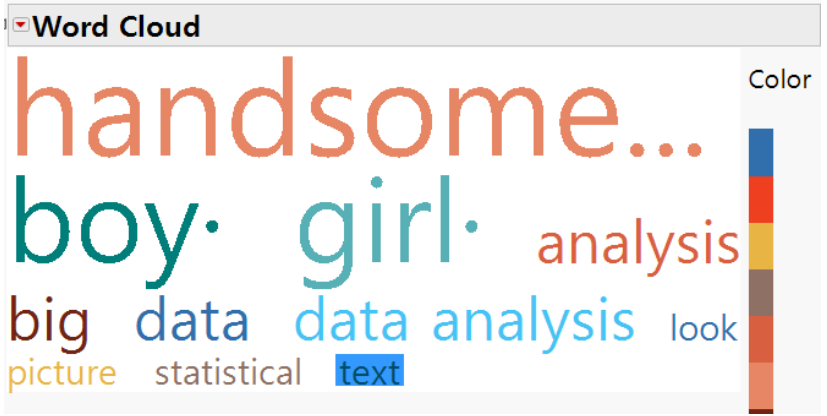


2. 기본적인 분석 Process 및 방법

8. Word Cloud Style로 표현하고자 한다면
   1) Display Options / Show Cloud 를 선택한 후

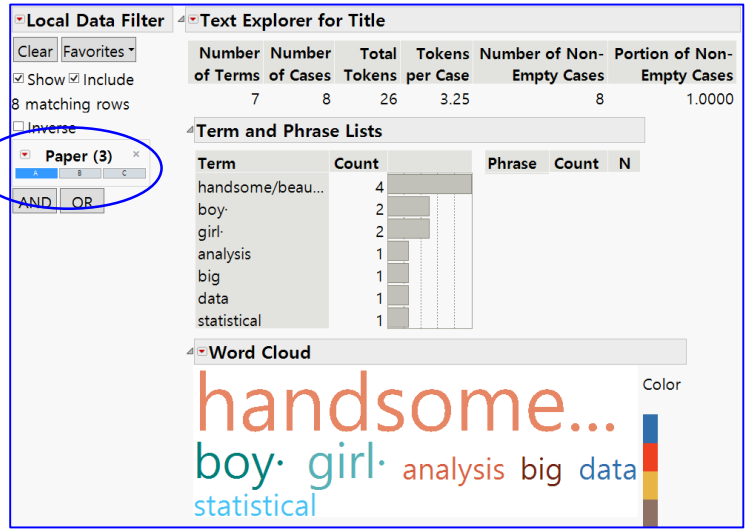


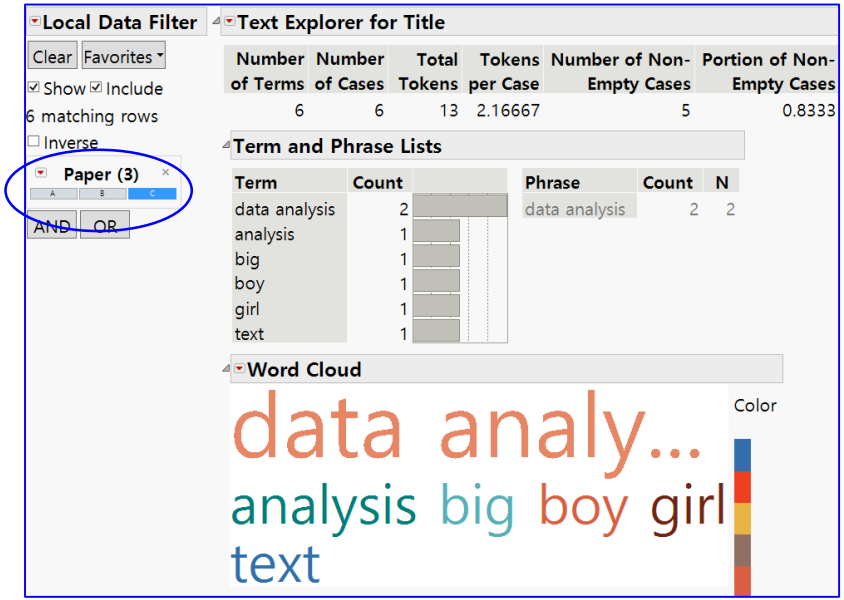
2) Word Cloud에서
   -Layout / Ordered 또는 Centered
   -Coloring / Arbitrary Colors 를 선택한 뒤 결과는 아래와 같음



2. 기본적인 분석 Process 및 방법

9. 층별(Stratification) 변수별로 확인해 보고자 한다면
Text Explorer의 왼쪽 붉은 색 역삼각형 클릭 후
Local Data Filter 기능을 활용하면 됨

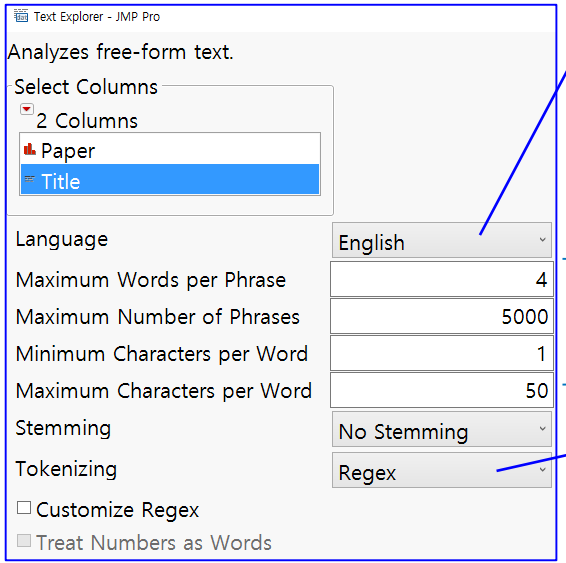




3. 실행 화면에 대한 간단한 설명

Stemming : 어간 추출의 방법 결정
(method for combining terms with similar beginning characters but different endings)

1) No Stemming
2) Stem for Combining
3) Stem All terms



JMP의 Text Explorer 기능은 아쉽게도 아직은 한글이 지원되지 않습니다(15 Version 기준)

일본어, 중국어는 Stemming, Tokenizing 지원 안 됨.

Phrase 당 최대 단어 수,
Word 당 최소/최대 글자 수
등을 지정

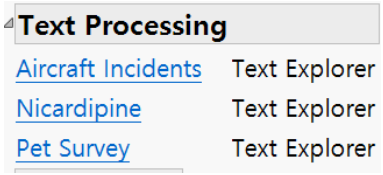
Tokening(Text를 어근(Term)으로 분리하는 것)
: 소문자(lowercase로 전환, 마침표/쉼표 등 구분 기호(delimiter) 제거

1) Regex : 정규 표현식(문자열과 일치할 수 있는 패턴의 정의)
2) Basic Words

* 보다 상세한 사항은
  1) Help / JMP Documentation Library의 basic analysis / text explorer (15 version) 또는
  2) Help / books / basic analysis 의 text explorer 편 참조(14 version 이하)
* JMP Homepage에서도 확인 가능(https://www.jmp.com/support/help/en/15.0/?os=win&source=application#page/jmp/text-explorer-overview.shtml#)

4. Text Explorer에 대한 추가적인 Study

1. Help / Sample Data 에서 왼쪽 sample filers categorized by type of analysis에서
   Text Processing을 클릭하면 Text Explorer 관련 추가적인 Sample data를 찾아볼 수 있습니다



2. Youtube에 있는 JMP Text Explorer 소개 동영상(3분)
   (https://www.youtube.com/watch?v=JnigzY7gl_o)

3. JMP Text Explorer에 대한 동영상 강의 자료(약 50분)
https://www.jmp.com/en_us/events/ondemand/mastering-jmp/text-explorer.html