

Chapter 9

Word Cloud

오 세 종

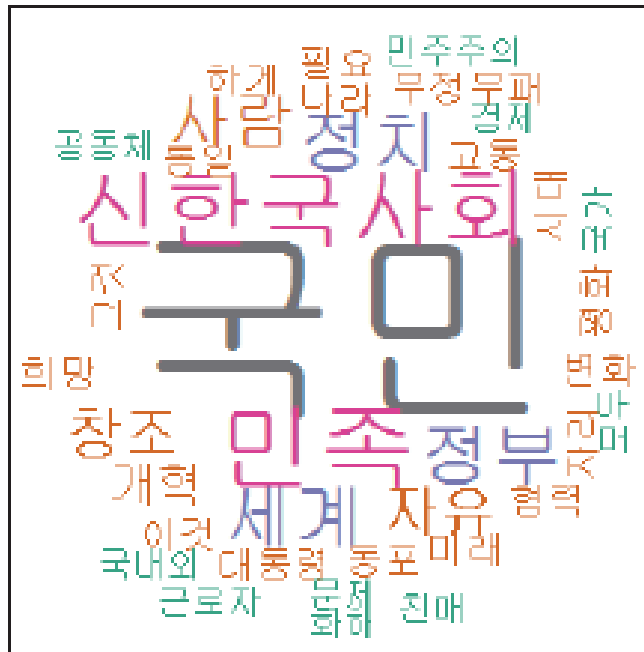
Contents



1. 워드 클라우드
2. 네이버 데이터랩
3. 구글 트렌드

1. 워드 클라우드

- 텍스트 마이닝: 비정형 텍스트에서 의미 있는 정보를 찾아내는 기술
 - 단어 분류 또는 문법적 구조 분석 등의 자연언어 처리 기술에 기반
 - 문서 분류, 관련 있는 문서들의 군집화, 정보의 추출, 문서 요약 등에 활용



문서내에서 발생 빈도가
높을수록 폰트의 크기를
크게 하여 배치

워드 클라우드

1. 워드 클라우드

- 필요한 패키지
 - ◉ wordcloud : 워드클라우드 작성
 - ◉ KoNLP : 한국어 처리
 - ◉ RColorBrewer : 단어의 색을 처리

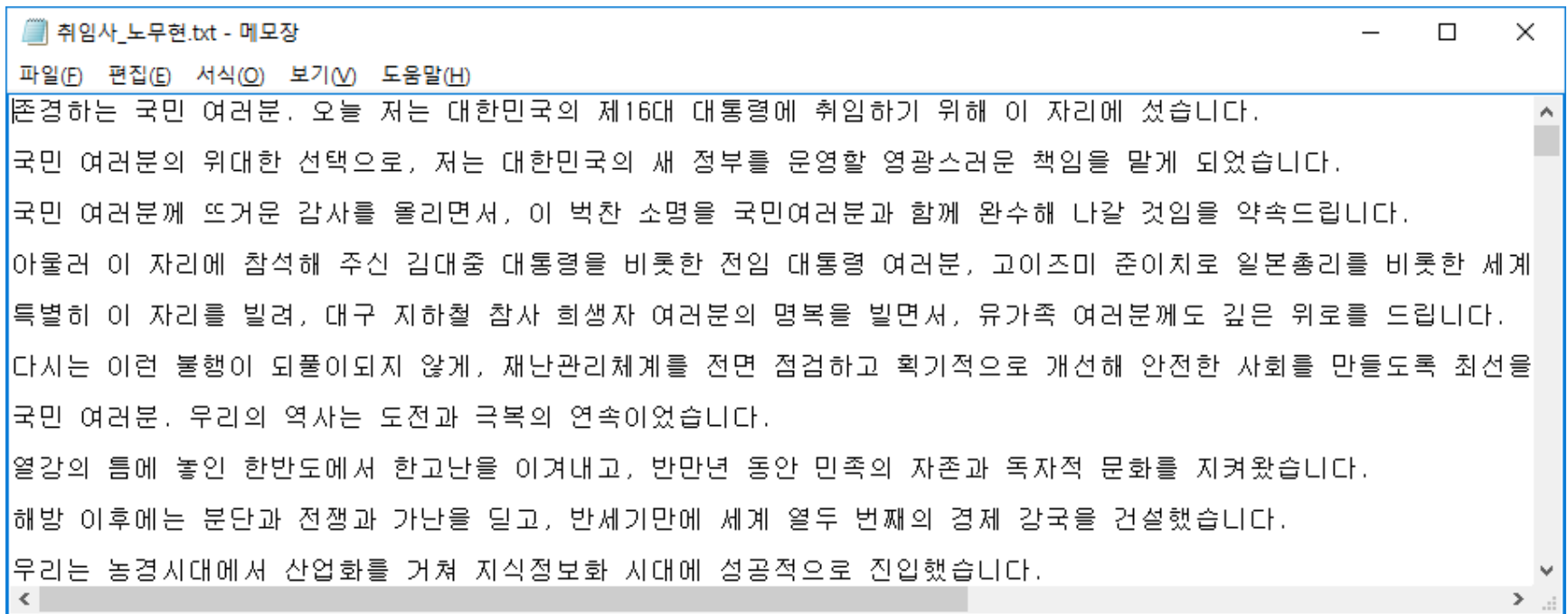
자료 출처 : “R 로 배우는 코딩(생능출판사)”

1. 워드 클라우드

- 대통령 연설문 분석

- 데이터 다운로드

인터넷에서 노무현, 이명박, 박근혜 대통령의 연설문을 검색하여 취임사 내용 전체를 마우스로 선택하여 복사한 후 **"취임사_노무현.txt"**와 같이 저장



1. 워드 클라우드

```
library(wordcloud)
library(KoNLP)
library(RColorBrewer)

useSejongDic() # 세종 한글사전 로딩
pal2 <- brewer.pal(8, "Dark2") # 팔레트 생성
text <- readLines(file.choose()) # 파일읽기
noun <- sapply(text, extractNoun, USE.NAMES=F)
```

④ 파일의 각 행에서 명사만 추출

☞ KoNLP의 extractNoun 함수 사용. sapply는 결과를 벡터 또는 행렬 형태로 반환함. "USE.NAMES=T"로 설정하면 단어 결과 위에 본문의 각 행이 포함됨

file.choose() : 파일 불러오기 윈도우

readLines() : 텍스트 파일 읽기

1. 워드 클라우드

```
> noun
```

```
[[1]]
```

```
[1] "존경" "국민" "여러분" "오늘" "저" "대한" "민국" "제16대"
```

```
[9] "대통령" "취임" "하기" "자리"
```

```
[[2]]
```

```
[1] ""
```

```
[[3]]
```

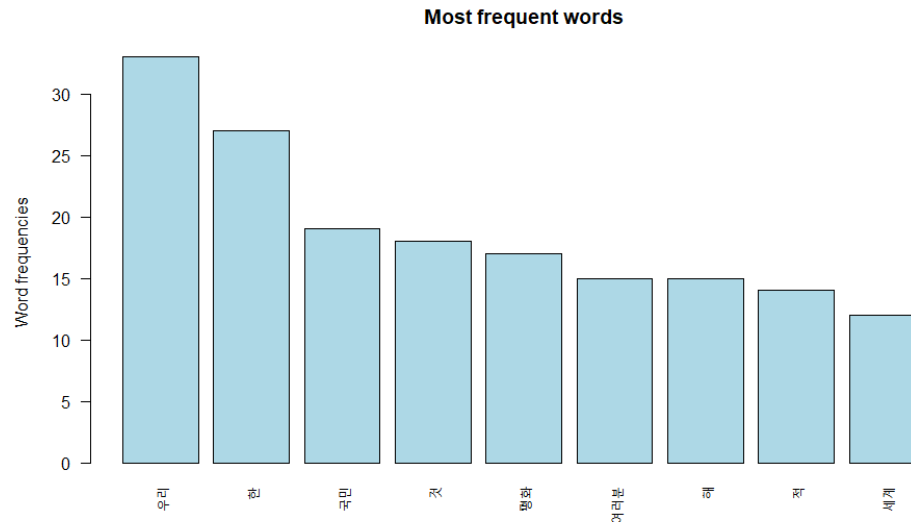
```
[1] "국민" "여러분" "위대" "한" "선택" "저" "대한" "민국"
```

```
[9] "정부" "운영" "영광" "책임"
```



1. 워드 클라우드

```
noun2 <- unlist(noun)           # 추출된 명사 통합
wordcount <- table(noun2)       # 단어 빈도수 계산
temp <- sort(wordcount, decreasing=T)[1:10]
temp
temp <- temp[-1]                # 공백단어 제거
barplot(temp, las = 2, names.arg = names(temp),
col = "lightblue", main = "Most frequent words",
ylab = "Word frequencies")
```



1. 워드 클라우드

```
wordcloud(names(wordcount),  
           freq=wordcount,  
           scale=c(6,0.7),  
           min.freq=3,  
           random.order=F,  
           rot.per=.1,  
           colors=pal2)
```

단어들
단어들의 빈도
단어의 폰트 크기 (최대,최소)
단어의 최소빈도
단어의 출력위치
90도회전 단어 비율
단어색

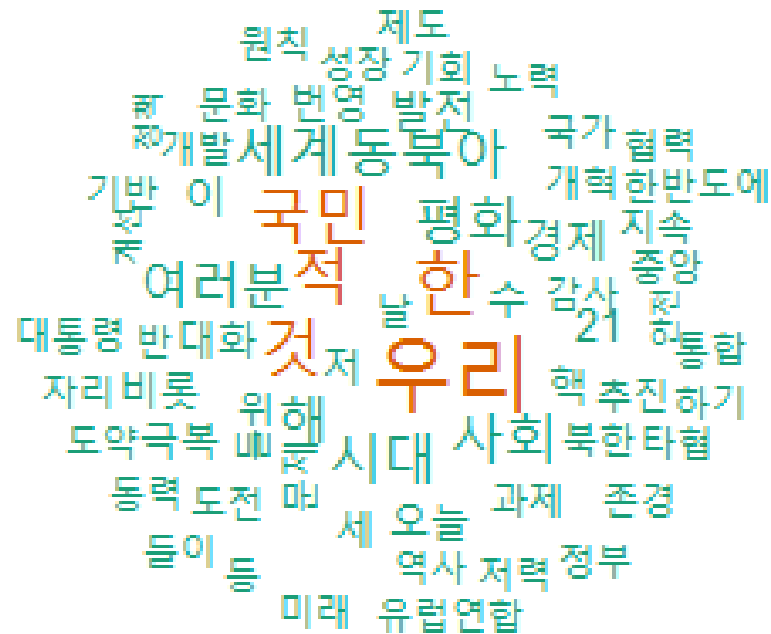
1. 워드 클라우드

● wordcloud

Parameter	설명
word	단어
freq	단어들의 빈도
size	가장빈도가 큰 단어와 빈도가 가장 작은 단어 폰트 사이의 크기 차이
min.freq	출력될 단어의 최소 빈도
max.word	출력될 단어들의 최대개수
random.order	TRUE 이면 랜덤으로 단어출력, FALSE 이면 빈도수가 큰 단어일수록 중앙에 배치
random.color	TRUE 이면 단어색은 랜덤순으로 정해지고, FALSE 이면 빈도순으로 정해짐
rot.per	90 도로 회전된 각도로 출력되는 단어의 비율
colors	가장 작은 빈도부터 큰 빈도까지의 단어색

1. 워드 클라우드

- 1차 결과



불필요한 단어도 섞여 있고, 사전에 없는 단어는 빠져있다.



1. 워드 클라우드

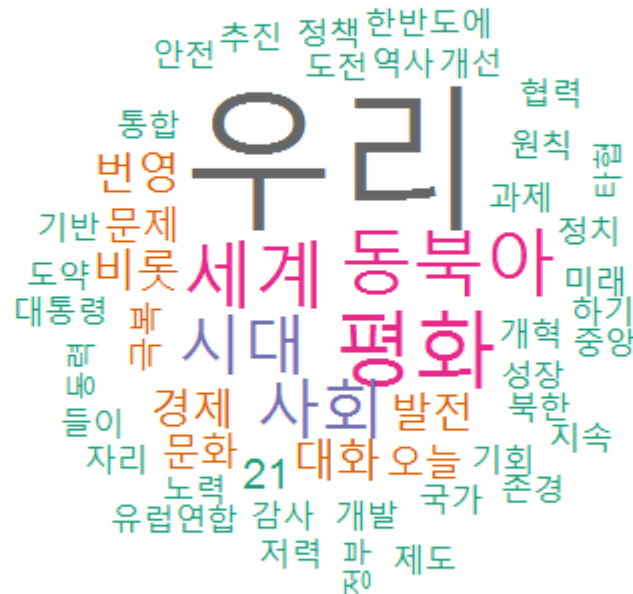
```
# 빈도수 높은데 워드클라우드에 없으면 사용자 사전에 추가
mergeUserDic(data.frame(c("정치"),c("ncn")))
noun <- sapply(text,extractNoun, USE.NAMES=F)
noun2 <- unlist(noun) # 추출된 명사 통합

# 무의미 단어 제거
noun2 <- noun2[nchar(noun2)>1] # 1글자 단어 제거
noun2 <- gsub("국민","", noun2) # '국민' 제거
noun2 <- gsub("여러분","", noun2) # '여러분' 제거

wordcount <- table(noun2) # 단어 빈도수 계산
wordcloud(names(wordcount),
          freq=wordcount,
          scale=c(6,0.7),
          min.freq=3,
          random.order=F,
          rot.per=.1,
          colors=pal2)
```

1. 워드 클라우드

- 노무현 대통령 취임 연설문 워드클라우드



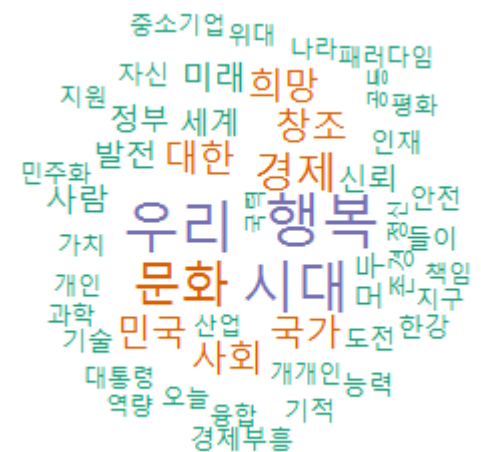
1. 워드 클라우드



노무현



이명박



박근혜



[연습문제 1]

- 20대 국회 개원-여·야 3당 대표 국회연설문에 대해 wordcloud 를 작성하시오
- https://www.taxtimes.co.kr/hous01.htm?bigcode=1&middlecode=0&smallcode=0&r_id=221813

2. 네이버 데이터랩

- <http://datalab.naver.com/>

NAVER DataLab. *beta*

[데이터랩 홈](#) [검색어로 알아보는 대한민국](#) [지역통계](#) [공공데이터](#)

검색어로 알아보는 대한민국

실시간 급상승 검색어와 검색어별 추이 그리고 TV/오락, 쇼핑, 영화 등 분야별 검색어 순위를 확인할 수 있습니다.

급상승 트래킹 NEW 분야별 인기 검색어

2017.11.06.(월) 14:47:00 기준

- 1 만수르
- 2 진해수
- 3 집단성매매
- 4 이정후
- 5 손승락
- 6 유상무
- 7 코코소리

2017.11.06.(월) 14:47:30 기준

- 1 만수르
- 2 진해수
- 3 손승락
- 4 집단성매매
- 5 이정후
- 6 유상무
- 7 코코소리

2017.11.06.(월) 14:48:00 기준

- 1 만수르
- 2 진해수
- 3 손승락
- 4 집단성매매
- 5 이정후
- 6 유상무
- 7 바른정당

2017.11.06.(월) 14:48:30 기준

- 1 만수르
- 2 손승락
- 3 진해수
- 4 집단성매매
- 5 이정후
- 6 유상무
- 7 바른정당

2. 네이버 데이터랩

주제

☒ 네이버 통합 검색어 ☐ 네이버 쇼핑 클릭수

전체선택 초기화

주제어1

가을

주제어 1에 해당하는 모든 검색어를 콤마(,)로 구분하여 최대 20개까지 입력

주제어2

단풍

주제어 2에 해당하는 모든 검색어를 콤마(,)로 구분하여 최대 20개까지 입력

주제어3

주제어 3 입력

주제어 3에 해당하는 모든 검색어를 콤마(,)로 구분하여 최대 20개까지 입력

주제어4

주제어 4 입력

주제어 4에 해당하는 모든 검색어를 콤마(,)로 구분하여 최대 20개까지 입력

주제어5

주제어 5 입력

주제어 5에 해당하는 모든 검색어를 콤마(,)로 구분하여 최대 20개까지 입력

범위

☒ 합계 ☐ 모바일 ☐ PC

기간

전체 1개월 3개월 1년 직접입력

일간

2017 08 05

2017 11 05

· 2016년 1월 이후 조회할 수 있습니다.

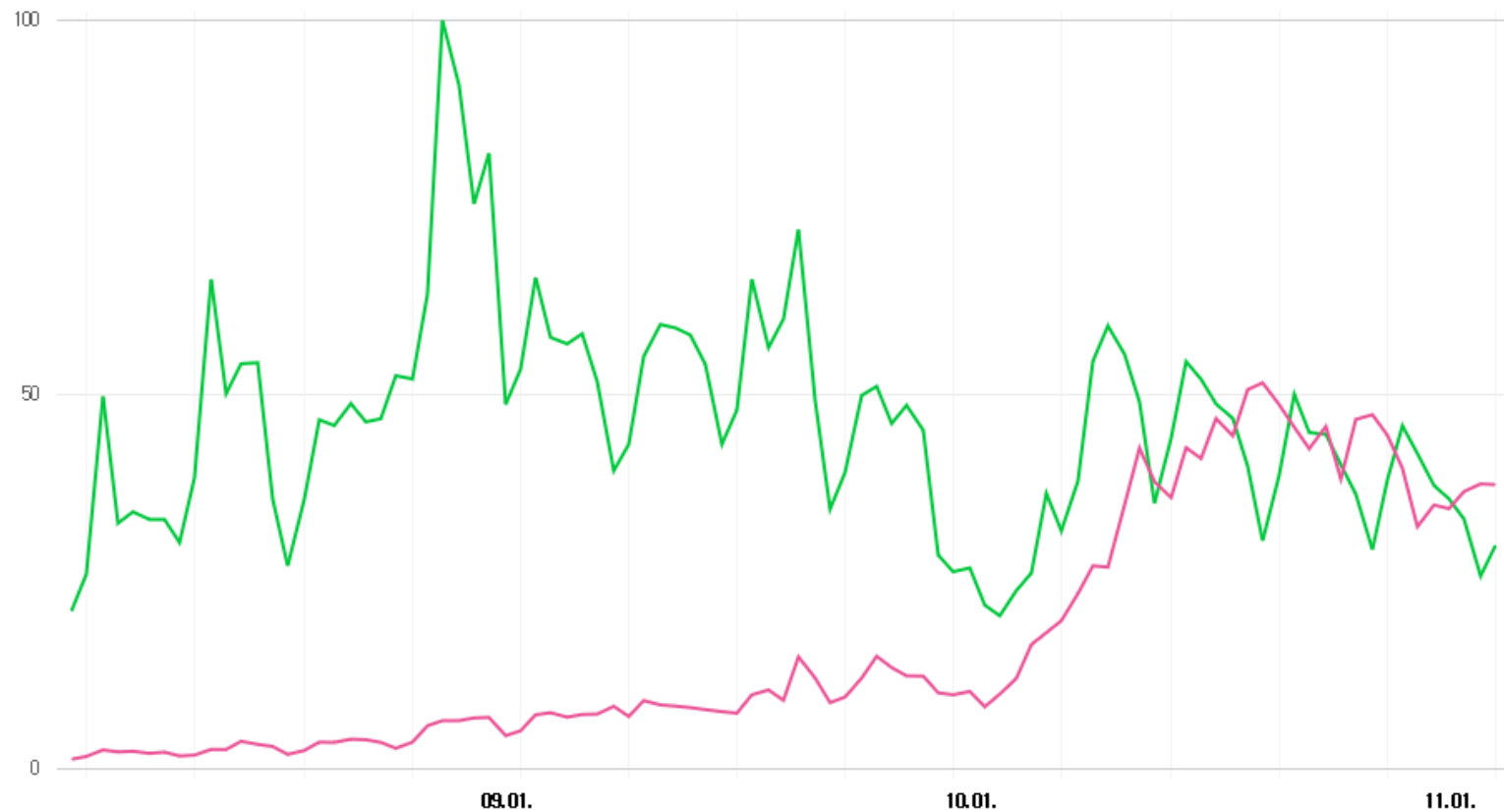
성별선택

☒ 전체 ☐ 여성 ☐ 남성

연령선택

☐ 전체 ☐ ~12 ☐ 13~18 ☐ 19~24 ☐ 25~29 ☐ 30~34 ☐ 35~39 ☐ 40~44 ☐ 45~49 ☐ 50~54 ☐ 55~60 ☐ 60~

2. 네이버 데이터랩



● 가을 | 가을
● 단풍 | 단풍

2. 네이버 데이터랩

지역통계 네이버 검색데이터와 다른 기관/기업 데이터를 통해 만들어진 정보로 지역별, 업종별 추이를 확인할 수 있습니다.



지역별 관심도

지역별 관심도는, 네이버에서 매일 발생하는 수 억건의 검색어와 네이버가 가지고 있는 수 백만건의 지역 데이터를 기초로, 조회기간내 지역별/업종별 관심도 정도를 확인할 수 있는 서비스입니다.



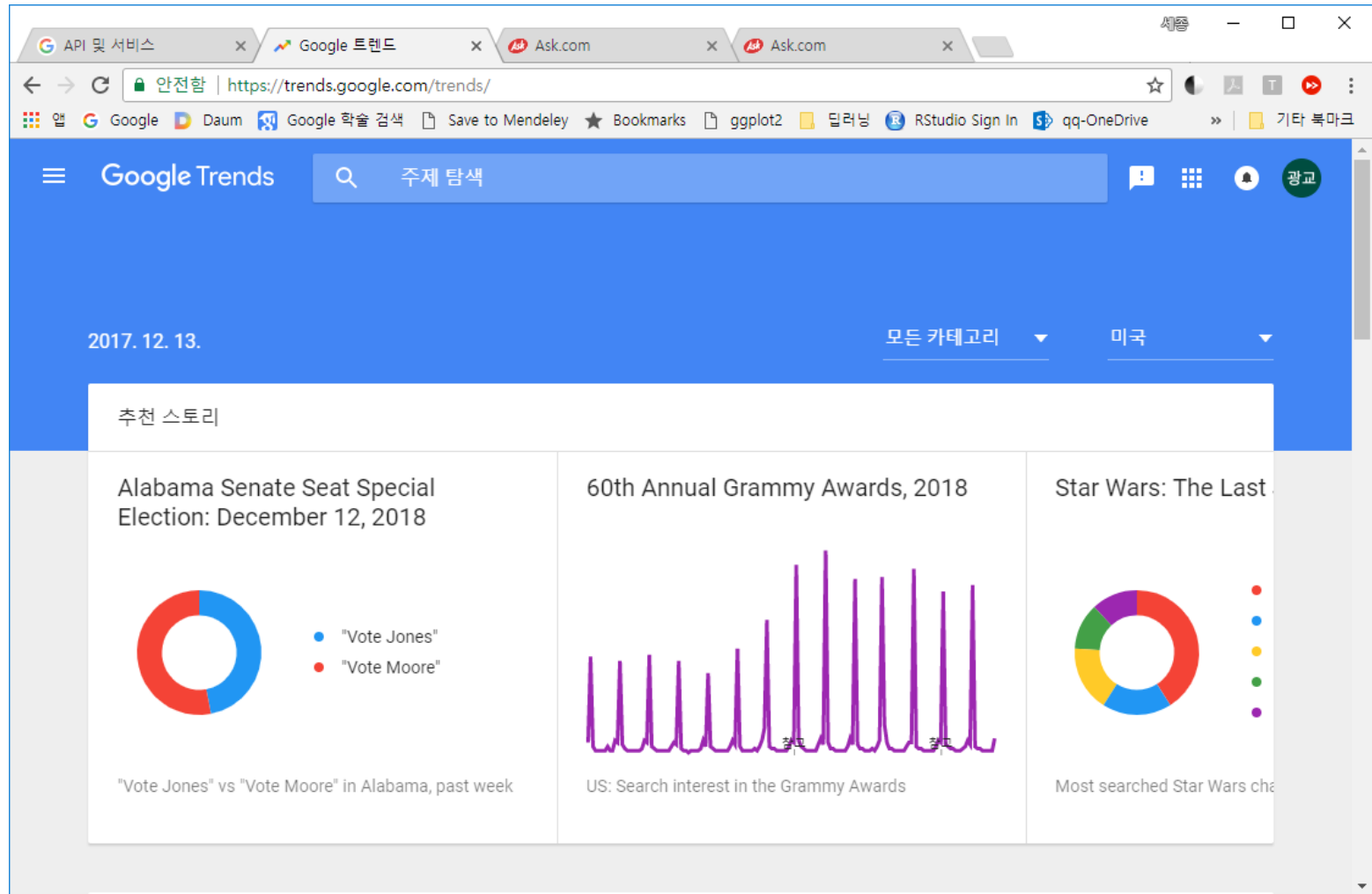
카드 소비통계

카드 소비통계는 전국의 지역별, 업종별로 발생하는 카드 결제 규모를 확인할 수 있는 서비스입니다. 연령별 성별로도 조회할 수 있습니다. 이 카드 소비 데이터는 BC카드에서 제공합니다.



3. 구글 트렌드

- <https://trends.google.com/trends/>



3. 구글 트렌드

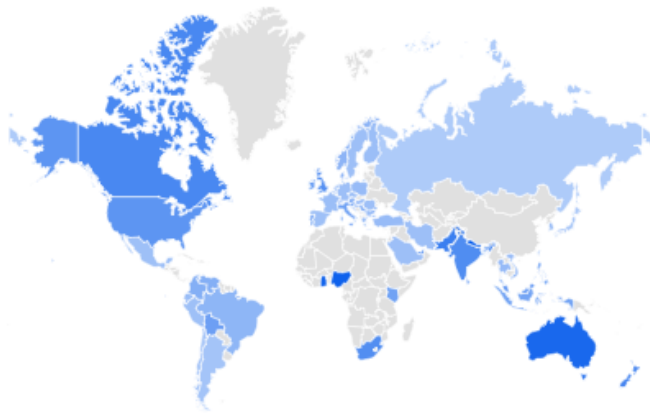


● bit coin

전 세계, 지난 12개월

지역별 관심도 ?

지역 ▼ ⬇ ⏪ ⏩ ↻



1	네팔	100	<div><div></div></div>
2	나이지리아	92	<div><div></div></div>
3	가나	90	<div><div></div></div>
4	오스트레일리아	86	<div><div></div></div>
5	파키스탄	67	<div><div></div></div>

☐ 검색량이 적은 지역 포함

< 1~5/62 개 지역 >



[부록] KoNLP 가 정상적으로 작동하지 않을 때

- (1) Java se 를 설치한다

The screenshot shows a Google search for 'Java se'. The search bar contains 'Java se'. Below the search bar, there are tabs for '전체' (All), '이미지' (Images), '동영상' (Videos), '도서' (Books), '뉴스' (News), and '더보기' (More). The '전체' tab is selected. Below the tabs, it says '검색결과 약 12,000,000개 (0.42초)'. The search results are as follows:

- Java SE - Downloads | Oracle Technology Network | Oracle**
www.oracle.com/technetwork/java/javase/.../index-jsp-138363.ht... ▼ 이 페이지 번역하기
Java SE downloads including: Java Development Kit (JDK), Server Java Runtime Environment (Server JRE), and Java Runtime Environment (JRE).
- Java SE Development Kit 8**
Download JDK 8, a development ...
Java SE Development Kit 8 ...
- Java SE Runtime Environment**
Java SE Runtime Environment 8
Downloads. Do you want to run ...
- Java SE Downloads**
Java SE Downloads Java Platform
(JDK) 8u111 / 8u112. NetBeans ...
- Java SE - Documentation**
Java SE Documentation. Visit the
Java Platform Standard Edition ...
- Java SE 7**
Java SE 7 Archive Downloads. Go to
the Oracle Java Archive page ...
- Oracle Java Archive**
The Oracle Java Archive offers self-
service download access

oracle.com 검색결과 더보기 »

이 슬라이드 이후는
자료로 제공. 설명x

Java SE Development Kit 8 Downloads

Thank you for downloading this release of the Java™ Platform, Standard Edition Development Kit (JDK™). The JDK is a development environment for building applications, applets, and components using the Java programming language.

• (1) Java

The JDK includes tools useful for developing and testing programs written in the Java programming language and running on the Java platform.

See also:

- [Java Developer Newsletter](#): From your Oracle account, select **Subscriptions**, expand **Technology**, and subscribe to **Java**.
- [Java Developer Day](#) hands-on workshops (free) and other events
- [Java Magazine](#)

JDK 8u131 checksum

Java SE Development Kit 8u131

You must accept the [Oracle Binary Code License Agreement for Java SE](#) to download this software.

☒ Accept License Agreement ☐ Decline License Agreement

Product / File Description	File Size	Download
Linux ARM 32 Hard Float ABI	77.87 MB	jdk-8u131-linux-arm32-vfp-hflt.tar.gz
Linux ARM 64 Hard Float ABI	74.81 MB	jdk-8u131-linux-arm64-vfp-hflt.tar.gz
Linux x86	164.66 MB	jdk-8u131-linux-i586.rpm
Linux x86	179.39 MB	jdk-8u131-linux-i586.tar.gz
Linux x64	162.11 MB	jdk-8u131-linux-x64.rpm
Linux x64	176.95 MB	jdk-8u131-linux-x64.tar.gz
Mac OS X	226.57 MB	jdk-8u131-macosx-x64.dmg
Solaris SPARC 64-bit	139.79 MB	jdk-8u131-solaris-sparcv9.tar.Z
Solaris SPARC 64-bit	99.13 MB	jdk-8u131-solaris-sparcv9.tar.gz
Solaris x64	140.51 MB	jdk-8u131-solaris-x64.tar.Z
Solaris x64	96.96 MB	jdk-8u131-solaris-x64.tar.gz
Windows x86	191.22 MB	jdk-8u131-windows-i586.exe
Windows x64	198.03 MB	jdk-8u131-windows-x64.exe

[부록] KoNLP 가 정상적으로 작동하지 않을 때

- (2) 환경변수 값을 설정한다

```
Sys.setenv(JAVA_HOME='C:/Program Files  
/Java/jre1.8.0_121')
```

