

문제해결을 위한 데이터 분석 및 시각화

텍스트 분석

워드 클라우드

한성대학교 노은희 교수

“미래로 향하는 새로운 이정표”



오늘의 학습

학습 내용

- 텍스트마이닝 개요
- 형태소 분석기(koNLPy)
- word cloud
- 워드클라우드 코딩없이 만들기

텍스트 마이닝 (Text Mining)

- 비/반정형 텍스트 데이터에서 자연어처리(Natural Language Processing)기술에 기반하여 유용한 정보를 추출, 가공하는 것을 목적으로 하는 기술


■ 텍스트 마이닝의 주요 기술

- 자연어 처리 (파싱, 형태소 분석, 품사 태깅, 관계 추출, 의미 추출)
- 언어모델링 (언어 감지, 규칙기반 개체명과 상용어 인식)
- 기계 학습 알고리즘 (반복훈련을 통해 습득한 정보 사용능력을 개선)
- 마이닝 기술 (각종 통계적 기법을 활용한 정보 분류 및 분석 기술)

■ 텍스트 마이닝의 3단계

1. 문서 수집 (Crawling)
2. 형태소 분석 (Konlpy)
3. 시각화 (Word Cloud)

* **형태소** : 의미를 가진 가장 작은 말의 단위. 더 나누면 뜻을 잃어버림.



자연어 처리-KoNLPy 모듈 설치

한국어 정보처리를 위한 파이썬 패키지

1. Java 설치
2. Java 환경변수 설정하기



KoNLPy: 파이썬 한국어 NLP

build passing docs passing

KoNLPy("코엔엘파이"라고 읽습니다)는 한국어 정보처리를 위한 파이썬 패키지입니다. 설치 방법은 [이 곳을](#) 참고해주세요.

NLP를 처음 시작하시는 분들은 [시작하기](#) 에서 가볍게 기본 지식을 습득할 수 있으며, KoNLPy의 사용법 가이드는 [사용하기](#), 각 모듈의 상세사항은 [API](#) 문서에서 보실 수 있습니다.

```
>>> from konlpy.tag import Kkma
>>> from konlpy.utils import pprint
>>> kkma = Kkma()
>>> pprint(kkma.sentences(u'네, 안녕하세요. 반갑습니다.'))
[네, 안녕하세요...,
 반갑습니다.]
>>> pprint(kkma.nouns(u'질문이나 건의사항은 깃헙 이슈 트래커에 남겨주세요.'))
[질문,
 건의,
 건의사항,
 사항,
 깃헙,
 이슈,
 트래커]
>>> pprint(kkma.pos(u'오류보고는 실행환경, 에러메세지와함께 설명을 최대한상세히!^^'))
[(오류, NNG),
 (보고, NNG),
```

KoNLPy is a Python package for Korean natural language processing.

Table of Contents

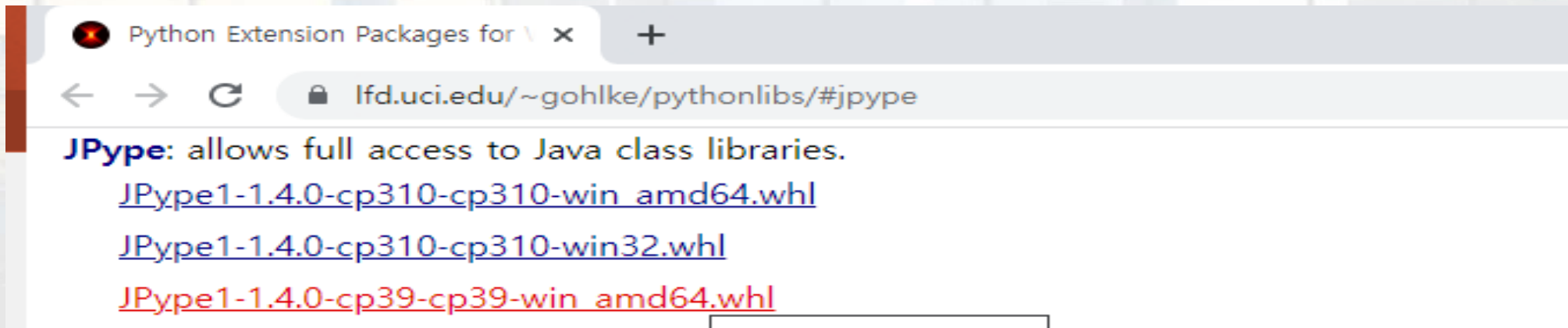
KoNLPy: 파이썬 한국어 NLP

- 거인의 어깨 위에 서기
- 라이선스
- 참여하기
- 시작하기
- 사용하기

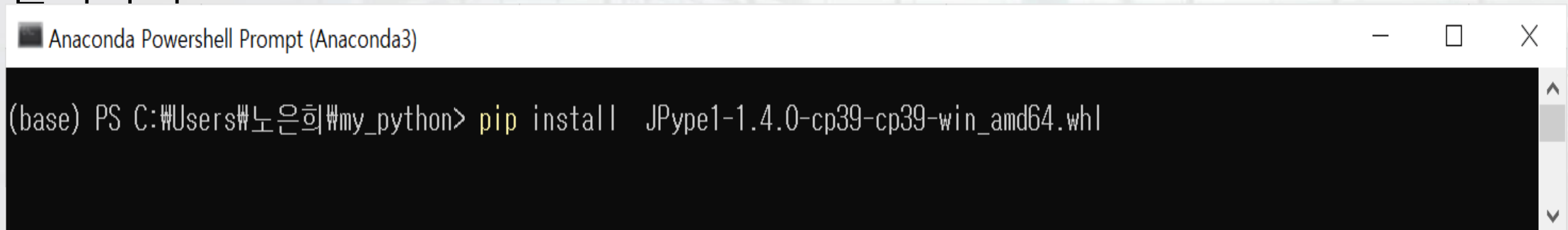
Fork me on GitHub

v: latest

3. JPype 다운로드 받아 설치하기(JPype는 자바 라이브러리를 파이썬으로 사용할 수 있도록 해줌) 다운받기



설치하기



4. koNLPy 설치하기

`pip install konlpy`



형태소 분석기

형태소 분석기의 각 장점과 단점에 따라 프로젝트에서 사용할 형태소 분석기를 사용해야 한다.
Konlpy(코모란, 한나눔, Kkma, mecab, Okt)에서 선택하여 사용 하는 것을 추천한다. Konlpy 안에서 각 형태소의 장단점


	코모란	한나눔	Kkma	Mecab	Okt(구:witter)
장점	자소 분리 가능, 오타자 분석 가능, 고유명사 분석 가능	로딩시간이 빠른편	띄어쓰기 오류에 덜 민감	새로운 사전 추가 가능, (twitter 공개전) 띄어쓰기에서 가장 좋은 성능, 속도, 정확도	띄어쓰기 성능이 가장 좋다, stemming 가능, 이모티콘, 해쉬태그 같은 인터넷 텍스트에 강함, 비속어, 비표준어도 분석 가능
한계점	로딩 속도가 길다, 띄어쓰기 없는 문장분석에 취약하다	띄어쓰기 없는 문장 분석에 매우 취약, 정제된 언어가 사용되지 않는 문서에 대한 형태소 분석 정확도가 높지 않은 문제점	분석 시간 오래 걸림, 정제된 언어가 사용되지 않는 문서에 대한 형태소 분석 정확도가 높지 않은 문제점	미등록어 처리 문제, 동음의어 처리 문제	미등록어 처리 문제, 동음의어 처리문제, 분석 범주 적은편

<출처><https://velog.io/@metterian/한국어-형태소-분석기POS-분석-3편.-형태소-분석기-비교>

형태소(形態素, 영어: morpheme)는 언어학에서 (일반적인 정의를 따르면) 일정한 의미가 있는 가장 작은 말의 단위



wordcloud



워드 클라우드(Word Cloud)

워드 클라우드(Word Cloud)는 단어들을 분석해 중요도 빈도수 인기도 등을 고려해 단어들을 시각적으로 배치하는 것

중요도를 나타내기 위해 글자 크기, 굵기, 색 등의 형태를 바꿔 나타낸다.

- `font_path` : 한글 폰트의 경로
- `background_color` : 배경 색 지정
- `width` : 가로폭 지정
- `height` : 세로폭 지정
- `max_words` : 이미지에 넣을 최대 word 수를 지정
- `max_font_size` : 이미지에 넣을 최대 폰트 크기를 지정

wordcloud.WordCloud

wordcloud

1.8.1

Search docs

USER DOCUMENTATION

API Reference

wordcloud.WordCloud

wordcloud.ImageColorGenerator

wordcloud.random_color_func

wordcloud.get_single_color_func

Command Line Interface

Gallery of Examples

Changelog

» API Reference » wordcloud.WordCloud

View page source

wordcloud.WordCloud

```
class wordcloud.WordCloud(font_path=None, width=400, height=200, margin=2, ranks_only=None, prefer_horizontal=0.9, mask=None, scale=1, color_func=None, max_words=200, min_font_size=4, stopwords=None, random_state=None, background_color='black', max_font_size=None, font_step=1, mode='RGB', relative_scaling='auto', regex=None, collocations=True, colormap=None, normalize_plurals=True, contour_width=0, contour_color='black', repeat=False, include_numbers=False, min_word_length=0, collocation_threshold=30)
```

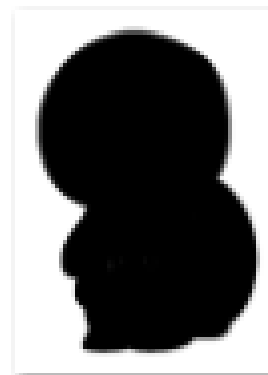
[source]

Word cloud object for generating and drawing.

Parameters:

font_path	: string
Font path to the font that will be used (OTF or TTF). Defaults to DroidSansMono path on a Linux machine. If you are on another OS or don't have this font, you need to adjust this path.	
width	: int (default=400)

실습



hsu_mask.png

사용 라이브러리

from wordcloud import WordCloud

- 워드클라우드를 생성에 필요한 기본 모듈

import matplotlib.pyplot as plt

- 생성한 워드클라우드 데이터를 시각화하여 그리기위한 모듈

from collections import Counter

- 텍스트를 추출하고, 빈도 수를 추출하기 위해 기본적으로 워드클라우드는 단어의 출현 빈도가 클수록 더 크게 그려집니다.

from konlpy.tag import Okt

- 한국어를 처리하는 대표적인 형태소 분석 패키지.

Okt, Kkma 등 여러가지 패키지들이 존재하는데 형태소 분석기마다 명사, 명사 등의 형태소를 조금씩 다르게 처리하므로 다양하게 사용해본 후, 가지고 있는 문서 특성에 적합한 형태소 분석기를 사용하는 것이 좋습니다.

from PIL import Image

- 워드클라우드를 원하는 형태로 그리기 위해 그림을 불러오는 패키지

import numpy as np

- 불러온 그림을 배열로 나타내어 쉽게 처리할 수 있도록 도와주는 패키지

[전체 소스코드]

```
1 | # 필요 라이브러리 불러오기
2 | from wordcloud import WordCloud # 워드클라우드를 생성에 필요한 기본 모듈
3 | import matplotlib.pyplot as plt # 생성한 워드클라우드 데이터를 시각화
4 | from collections import Counter # 텍스트를 추출하고, 빈도 수를 추출
5 | from konlpy.tag import Okt # 한국어를 처리하는 대표적인 형태소 분석 패키지
6 | from PIL import Image # 워드클라우드를 원하는 형태로 그리기 위해 그림을 불러오는 패키지
7 | import numpy as np # 불러온 그림을 배열로 나타내어 쉽게 처리할 수 있도록 도와주는 패키지
8 |
9 | # 텍스트 파일 불러오기
10 | text = open('대한민국헌법.txt', 'r', encoding='UTF-8').read()
11 |
12 | # Okt 형태소 분석기 객체 생성과 명사만 추출
13 | okt = Okt() # Okt 형태소 분석기 객체를 생성
14 | nouns = okt.nouns(text) # 명사만 추출
15 |
16 | # 추출된 단어 중 단어의 길이가 1개인 것은 제외
17 | words = []
18 | for n in nouns:
19 |     if len(n) > 1: # 단어의 길이가 1개인 것은 제외
20 |         words.append(n)
21 |
```

[전체 소스코드]

```
22 # 단어 빈도수 구하기
23 c = Counter(words) # 위에서 얻은 words를 처리하여 단어별 빈도수 형태의 딕셔너리 데이터를 구함
24
25 # 이미지로 사용될 이미지 불러오기
26 img = Image.open('image/book_mask.png') #image.open 함수를 통해 이미지를 불러오기
27 img_array = np.array(img)
28
29 # 워드 클라우드 만들기
30 wc = WordCloud(font_path = 'C:/Windows/Fonts/malgun.ttf',
31                background_color="black",
32                width=400,
33                height=400,
34                scale=2.0,
35                max_font_size=250,
36                mask=img_array) #mask에 넣기
37
38 gen = wc.generate_from_frequencies(c)
39 plt.axis('off')
40 plt.imshow(gen)
41 plt.show()
```

이미지맵(imshow)의 원리

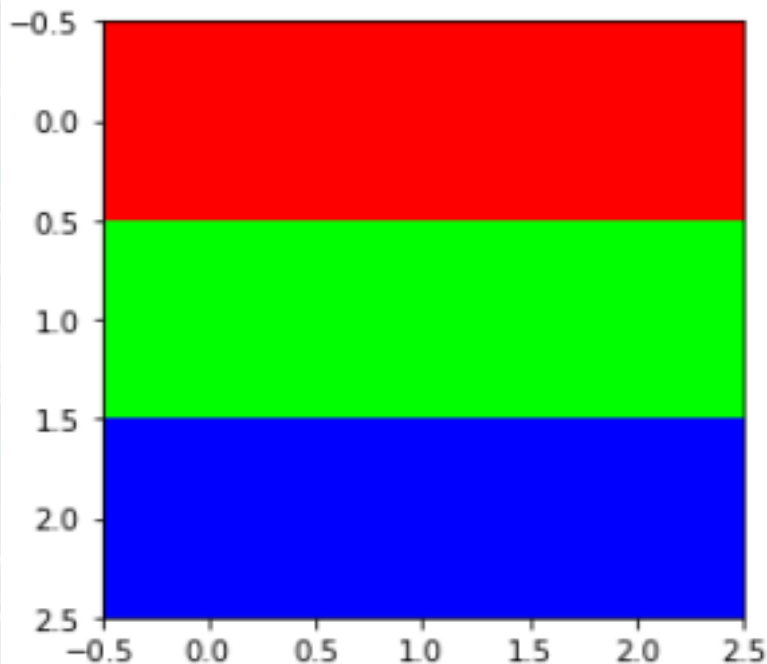
이미지맵(imshow)의 원리


imshow는 원하는 사이즈의 픽셀을 원하는 색으로 채워서 만든 그림
원하는 크기의 행렬을 만들어서 각 칸을 원하는 색으로 채우는 것
각 칸을 채우는 방법은 colormap, RGB, RGBA

예>RGB

RGB는 행렬의 각 원소로 [R,G,B] 값을 입력. 값은 0-1 사이 실수 혹은 0-255 사이 정수로 입력

```
1 import numpy as np
2 from matplotlib import pyplot as plt
3
4 m=np.array(
5
6 [
7
8     [[255,0,0],[255,0,0],[255,0,0]],
9
10    [[0,255,0],[0,255,0],[0,255,0]],
11
12    [[0,0,255],[0,0,255],[0,0,255]],
13 ])
14
15 plt.imshow(m)
16 plt.show()
```

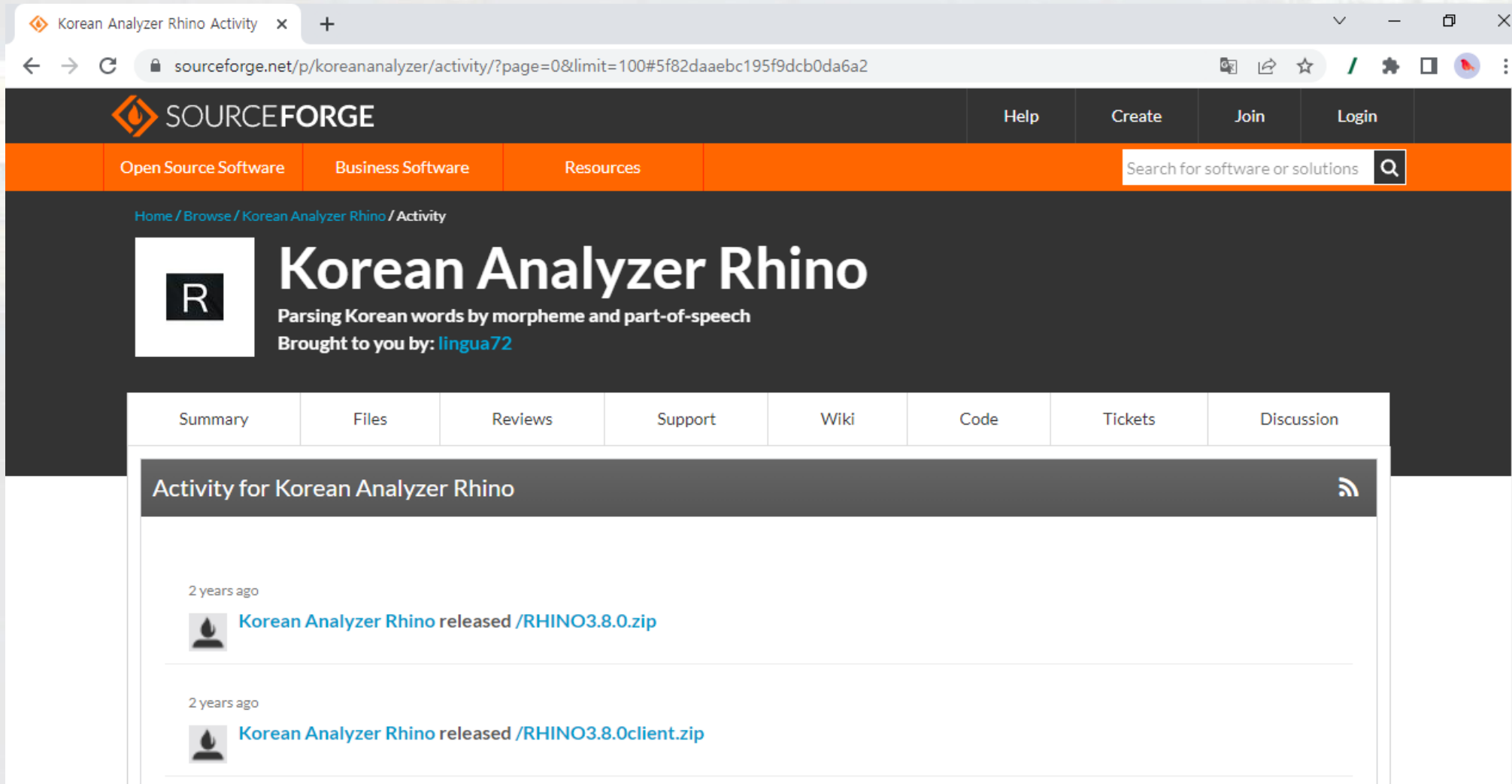






형태소 분석 프로그램 사용

Real Hangu INput Object

<https://sourceforge.net/p/koreananalyzer/activity/?page=0&limit=100#5f82daaebc195f9dcb0da6a2>



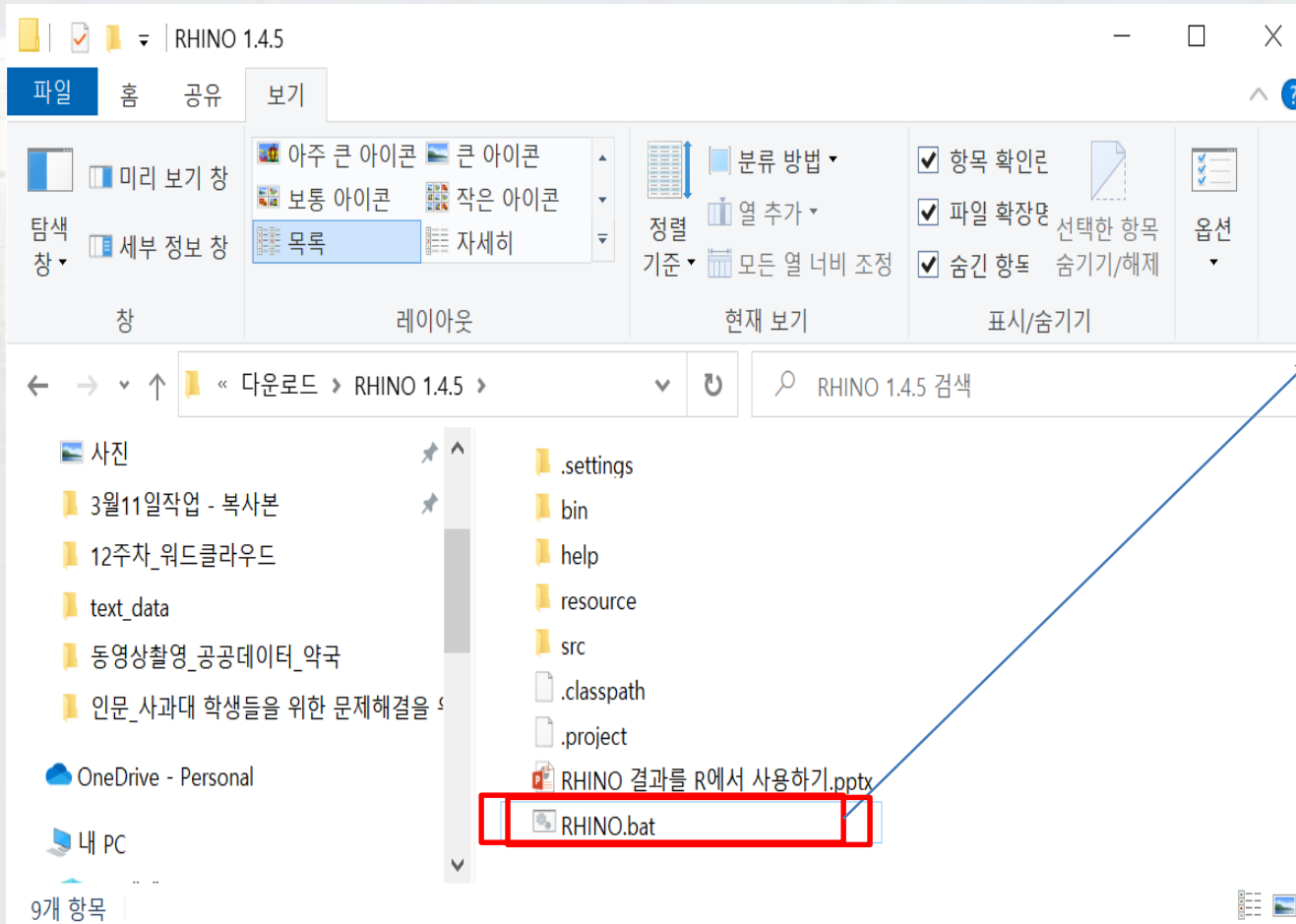
The screenshot shows a web browser window displaying the SourceForge activity page for the project "Korean Analyzer Rhino". The browser's address bar shows the URL: sourceforge.net/p/koreananalyzer/activity/?page=0&limit=100#5f82daaebc195f9dcb0da6a2. The SourceForge logo is in the top left, and navigation links for "Help", "Create", "Join", and "Login" are in the top right. Below the logo, there are tabs for "Open Source Software", "Business Software", and "Resources", along with a search bar. The main content area has a breadcrumb trail: "Home / Browse / Korean Analyzer Rhino / Activity". The project name "Korean Analyzer Rhino" is prominently displayed, followed by the description "Parsing Korean words by morpheme and part-of-speech" and "Brought to you by: [lingua72](#)". A horizontal menu contains links for "Summary", "Files", "Reviews", "Support", "Wiki", "Code", "Tickets", and "Discussion". The "Activity for Korean Analyzer Rhino" section shows two recent releases, both dated "2 years ago":

-  [Korean Analyzer Rhino released /RHINO3.8.0.zip](#)
-  [Korean Analyzer Rhino released /RHINO3.8.0client.zip](#)

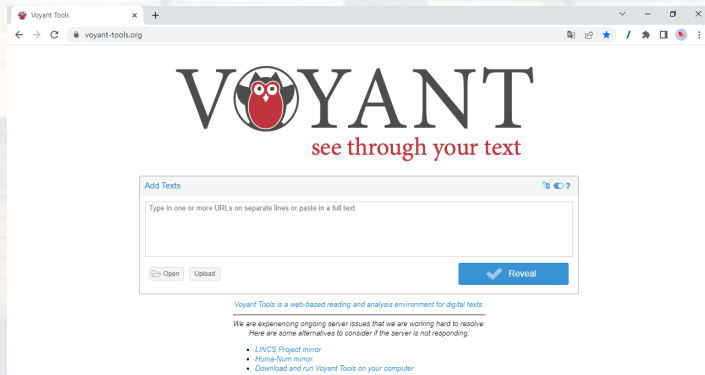
한글 형태소 분석기 RHINO

한글 형태소 분석기 RHINO

Korean Morphological Analyzer, RHINO



VOYANT사용하여 워드클라우드 만들기



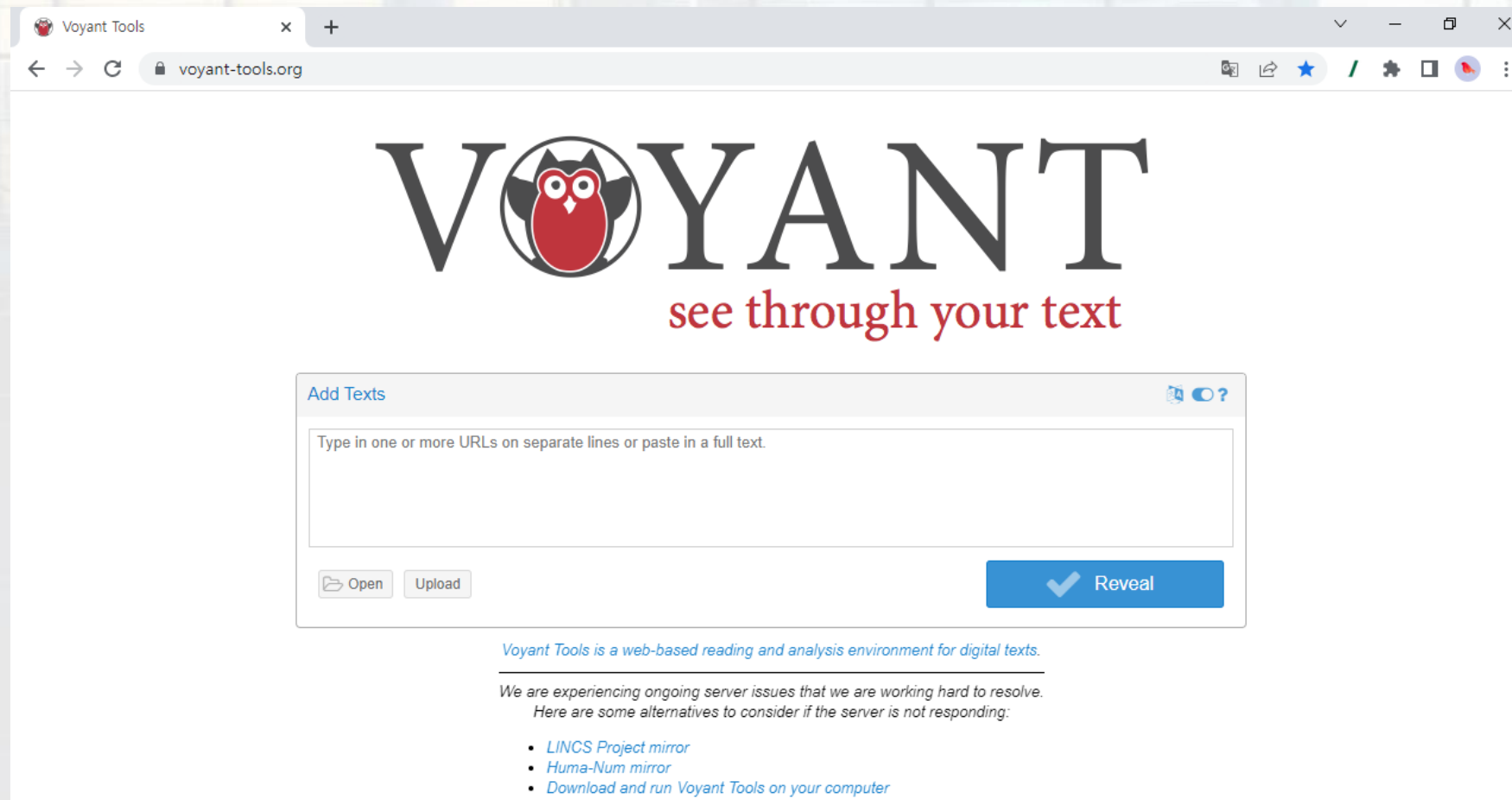
워드클라우드 생성기를 이용하여 워드클라우드 만들기

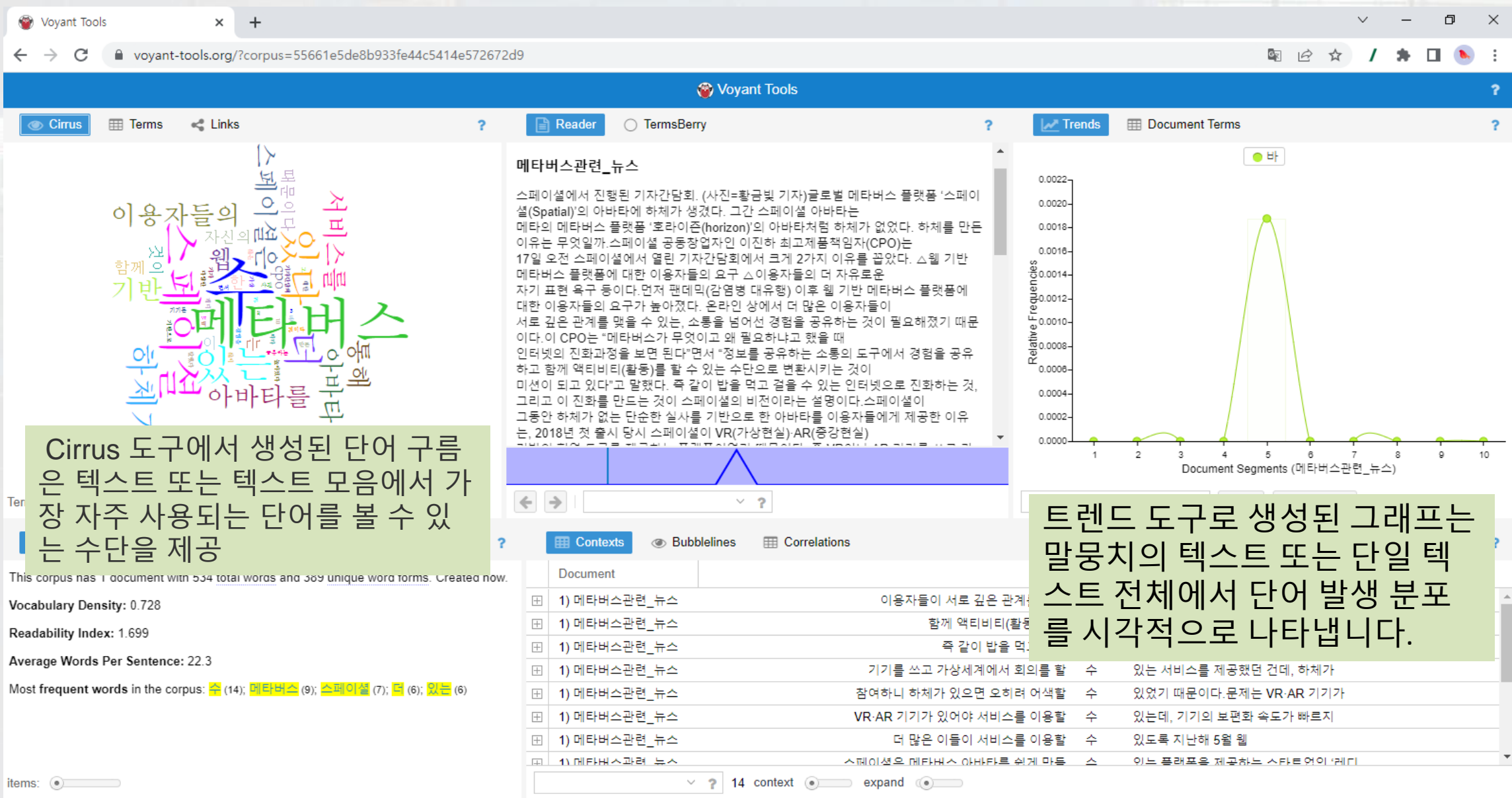


보이언트 툴(Voyant Tools)

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

Voyant Tools 는 텍스트 분석을 수행하기 위한 오픈 소스 웹 기반 응용 프로그램
디지털 인문학 분야의 학자 뿐 아니라 학생과 일반 대중이 텍스트 또는 말뭉치를 학술적으로 읽고
해석할 수 있도록 지원





<http://wordcloud.kr/>

jeju.com

저장&공유

영기

<http://wordcloud.kr/>



워크생성기

[갤러리](#)

무료폰트

영어버전

일본어 버전

글자색 rainbow

[폰트](#)
[나눔고딕](#)

I 폰트미리보기

배경색



마스크

크기

직접입력

390px

00px

단어수

9507H

키워드

텍스트

기자, 간담회, 사진, 갈금빛, 기자, 글로벌, 메타버스, 플랫폼, 의, 아바타, 하체, 그간, 아바타, 메타, 메타버스, 플랫폼, 호라이즌, 아바타, 하체, 하체, 이유, 무엇, 일, 공동, 창업자, 이진, 최고, 제품, 책임자, 일, 오전, ", 열린, 기자, 간담회, 이유, 웹, 기반, 메타버스, 플랫폼, 이용자, 요구, 이용자, 자유로, 자기 표현, 어떤, 편, 각자, 변, 이해, 이해, 기반, 메타버스, 플랫폼, 이

⚙️ 워드클라우드 만들기

저장&공유

Google에 의해 종
료된 광고입니다.

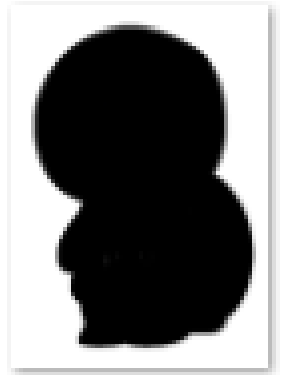


Media in Translation

기자, 간담회, 사진, 황금빛, 기자, 글로벌, 메타버스", 플랫폼, 의, 아바타, 하체, 그간, 아바타, 메타, 메타버스", 플랫폼, 호라이즌, 아바타, 하체, 하체, 이유, 무엇, 일, 공동, 창업자, 이진, 최고, 제품, 책임자, 일, 오전, ", 열린, 기자, 간담회, 이유, 웹, 기반, 메타버스", 플랫폼, 이용자, 요구, 이용자, 자유로, 자기, 표현, 욕구, 팬, 믹, 감염, 병, 유행, 이후, 웹, 기반, 메타버스", 플랫폼, 이용자, 요구, 온라인, 이용자, 관계, 소통, 경험, 필요, 해, 메타버스", 무엇, 필요, 때, 인터넷, 진화, 과정, 정보, 소통, 도구, 경험, 액, 티비, 티, 활동, ", 수단, 변환, 미션, 밥, 인터넷, 진화, ", 비전, 설명, ", 그동안, 하체, 실사, 기반, 아바타, 이용자, 이유, 출시, 당시, 가상현실, 증강, 현실, 기반, 협업, 도구, 플랫폼, 기기, 가상, 세계, 회의, 서비", 하체, 기기, 상체, 시야, 사람, 실제, 의자, 회의, 하체, 문제, 기기, 서비", 이용, 기기, 보편, 속도, 서비", 이용, 지난해, 웹, 버전, 기존, 접근, 자신, 개성, 이용자, 요구, ", 문화, 예술, 콘텐츠, 중심, 메타버스", 플랫폼, 창작자, 자신, 작품, 차원, 가상, 갤러리, 용도, 주제, 공간, 이벤트, 파티, 창작자, 입장, 자신, 아바타, 메타버스", 욕구, 결국, ", 메타버스", 아바타, 플랫폼, "타트, 레디, 플레이어, 미, 파트너십, 바탕, 전신, 아바타, 지원, 기능, 이날, 이번, 파트너십, 계기, 이용자, 프로필, 피, ", 신체, 모양, 복장, 헤어, "타일, 가운데, 아바타, 얼굴, 기존, 실제, 사진, 기반, 옵션, 레디, 플레이어, 미, 일러"트, 기반, 캐릭터, 옵션, 아바타, 자신, 신체, 문화, 정체, 자유, 지속, 방침, 문화, 전통, 의상, 향후, 아바타, 아이템, 창작자, 기능, 예정, ", 아바타, 기능, 사진, 황금빛, 기자, ", 웹, 기반, 메타버스", 플랫폼, 서비", 본격, 이상, 가상, 공간, ", 이용자, 지난해, 때, 배, 가량, 인간, 방식, 보편, 다양, 플랫폼, 이용자, 커뮤니티, 저변, 이날, 기자, 간담회, 참석, 메타버스", 체험, 기자, 이동, 방법, 자동, 화질, 간담회, 참석자, 공간, 때, 속도, ", 이, 화질, 속도, 경우, 컴퓨터, 와이파이, 상황, 조금, ", 이용자, 중심, 서비", 개선, 지속

마무리

- 텍스트마이닝 개요
- 형태소 분석기(koNLPy)
- word cloud
- 워드클라우드 코딩없이 만들기



hsu_mask.png