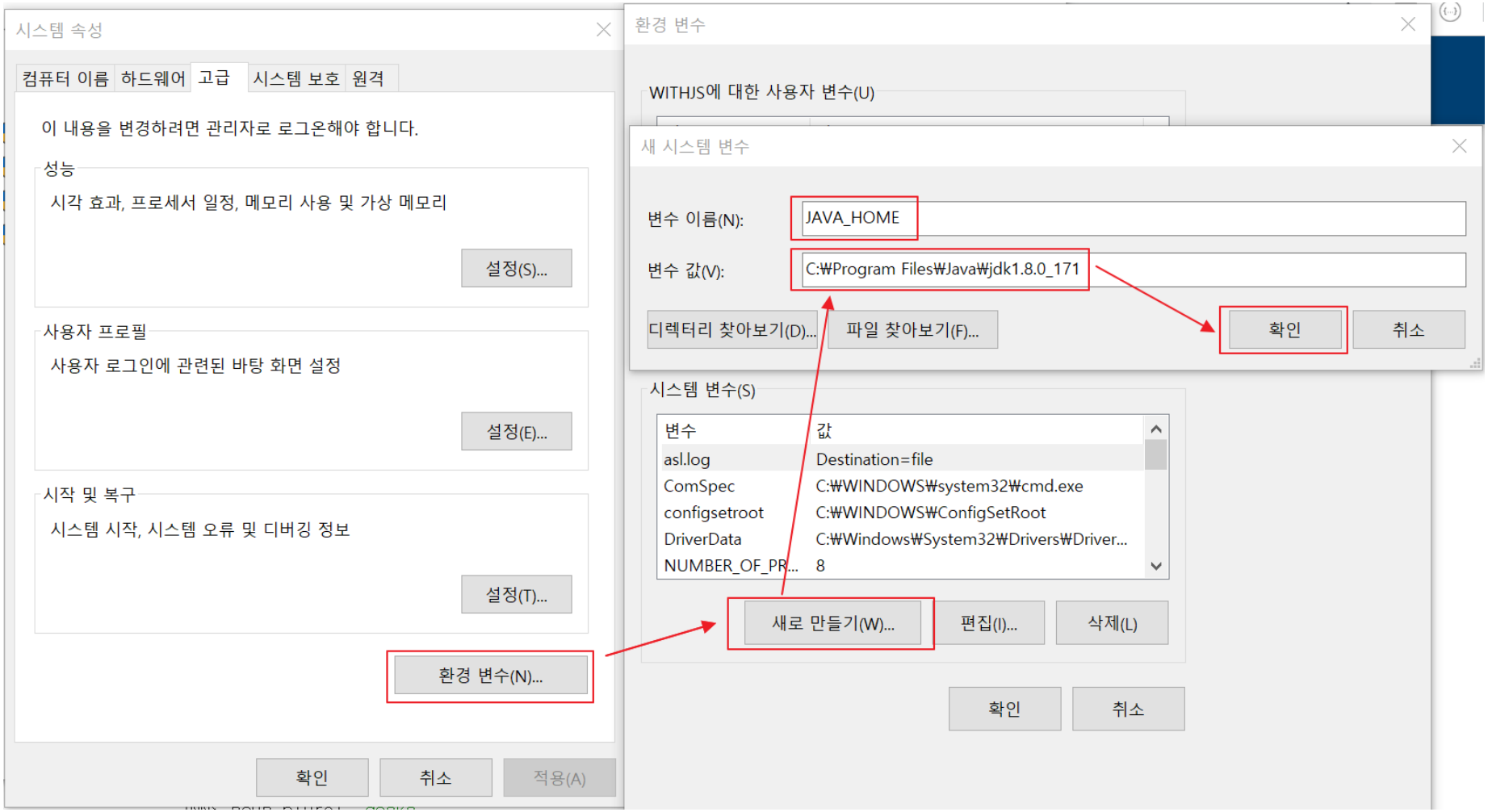
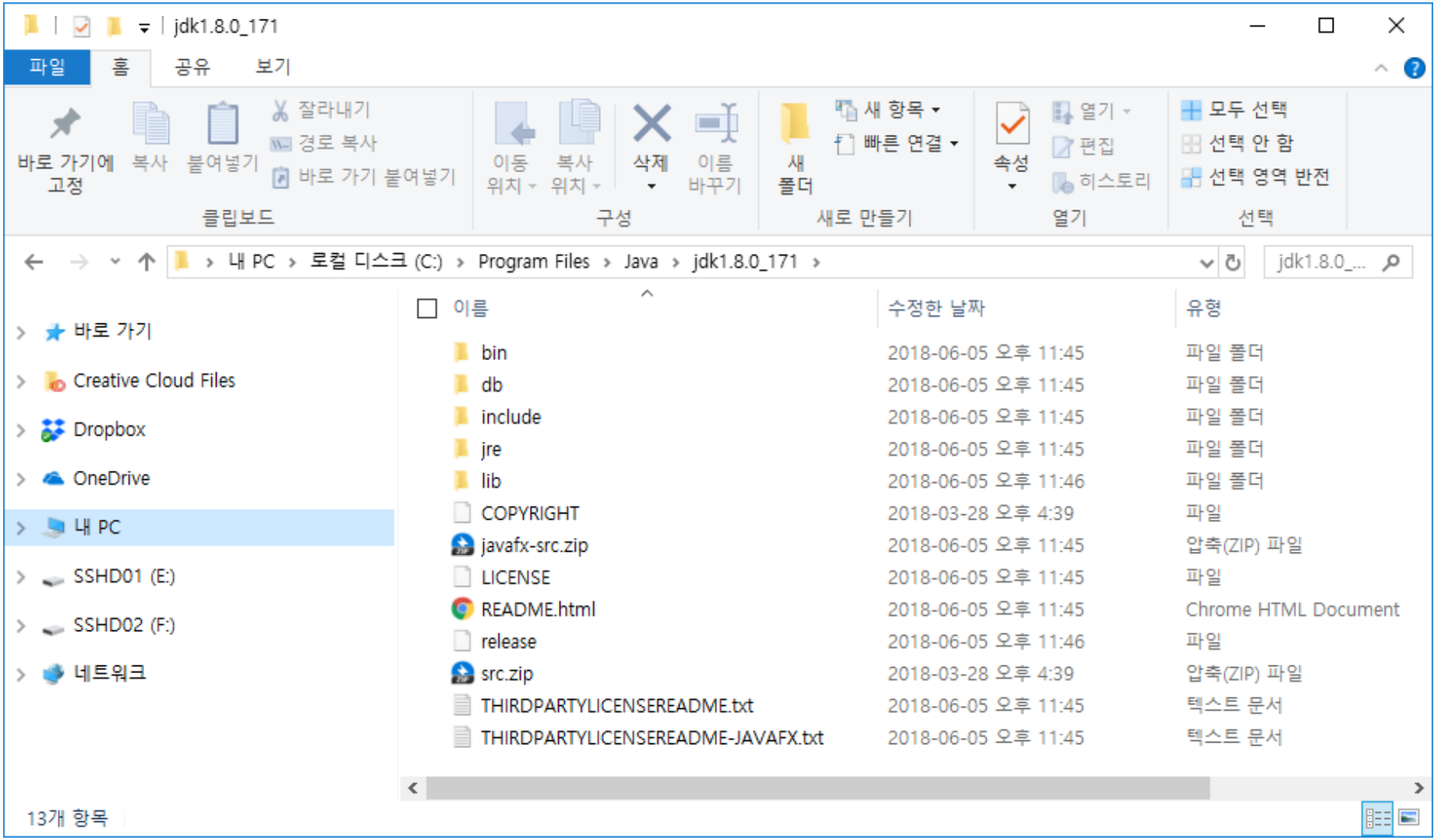
KoNLPY(Korean NLP in Python)

- Lucy Park 분이 개발
- 26회 한글 및 한국어 정보처리 학술대회 논문집(2014년)에 Lucy Park 님이 KoNLPy를 발표

01. 설치하기

- JDK 홈페이지에서 JDK 다운로드 후, 설치
- JDK 설치 후, JAVA_HOME 설정을 수행.
  - 원도우 유저 [제어판] - [시스템] - [고급시스템 설정]- [환경변수] - [새로만들기] - [JAVA_HOME추가]
  - MAC 유저 export JAVA_HOME=$(/usr/libexec/java_home)
- Jpype1-py3 설치
  - (JDK 버전, Jupyter의 설치 경로에 한글이 있는 경우, PC 이름에 한글이 있는 경우 에러 발생이 있음)
  - pip install --upgrade pip # 인스톨을 위해 pip 업그레이드가 필요할 수 있음.
  - pip install Jpype1-py3
- 터미널에서 konlp 설치
  - pip install konlpy
- pip install wordcloud


시스템 속성
컴퓨터 이름
하드웨어
고급
시스템 보호
원격
이 내용을 변경하려면 관리자로 로그온해야 합니다.
성능
시각 효과, 프로세서 일정, 메모리 사용 및 가상 메모리
설정(S)...
사용자 프로필
사용자 로그인에 관련된 바탕 화면 설정
설정(E)...
시작 및 복구
시스템 시작, 시스템 오류 및 디버깅 정보
설정(T)...
환경 변수(N)...
확인
취소
적용(A)
환경 변수
WITHJS에 대한 사용자 변수(U)
새 시스템 변수
변수 이름(N):
JAVA_HOME
변수 값(V):
C:₩Program Files₩Java₩jdk1.8.0_171
디렉터리 찾아보기(D)...
파일 찾아보기(F)...
확인
취소
시스템 변수(S)
변수
값
asl.log
Destination=file
ComSpec
C:₩WINDOWS₩system32₩cmd.exe
configsetroot
C:₩WINDOWS₩ConfigSetRoot
DriverData
C:₩Windows₩System32₩Drivers₩Driver...
NUMBER_OF_PR...
8
새로 만들기(W)...
편집(I)...
삭제(L)
확인
취소



KoNLPy 시작하기

- 공식 웹 사이트 : http://konlpy.org/en/latest/ (http://konlpy.org/en/latest/)
- 꼬꼬마, 한나눔 등의 엔진 사용이 가능
  - http://kkma.snu.ac.kr/ (http://kkma.snu.ac.kr/) (꼬꼬마-서울대 이상구 교수 연구팀)
  - http://semanticweb.kaist.ac.kr/home/index.php/HanNanum (http://semanticweb.kaist.ac.kr/home/index.php/HanNanum) KAIST(1999) 최기선 교수 연구팀

In [9]:
```
from konlpy.tag import Kkma
kkma = Kkma()

kkma.sentences("안녕하세요! 오늘은 한글어 분석을 시작합니다")
```

Out[9]:
```
['안녕하세요!', '오늘은 한글어 분석을 시작합니다']
```

- 꼬꼬마 모듈을 이용하여 문장 분석을 했다. 두 개의 문장으로 구분함.

명사(nouns) 분석

In [10]:
```
kkma.nouns("안녕하세요! 오늘은 한글어 분석을 시작합니다")
```

Out[10]:
```
['안녕', '오늘', '한글', '분석', '시작']
```

형태소 분석

- 형태소 품사(Part Of Speech, POS) 태그표 : http://kkma.snu.ac.kr/documents/index.jsp?doc=postag (http://kkma.snu.ac.kr/documents/index.jsp?doc=postag)

In [11]:
```
kkma.pos("안녕하세요! 오늘은 한글어 분석을 시작합니다")
```

Out[11]:
```
[('안녕', 'NNG'),
 ('하', 'XSV'),
 ('세요', 'EFN'),
 ('!', 'SF'),
 ('오늘', 'NNG'),
 ('은', 'JX'),
 ('한글', 'NNG'),
 ('어', 'XSN'),
 ('분석', 'NNG'),
 ('을', 'JKO'),
 ('시작', 'NNG'),
 ('하', 'XSV'),
 ('ㅂ니다', 'EFN')]
```

한나눔 한글 엔진 사용해 보기

In [8]:
```
from konlpy.tag import Hannanum
hannanum = Hannanum()
```

In [12]:
```
hannanum.nouns("안녕하세요! 오늘은 한글어 분석을 시작합니다")
```

Out[12]:
```
['안녕', '오늘', '한글어', '분석', '시작']
```

wordCloud 사용해 보기

참조 URL : [https://github.com/amueller/word_cloud (https://github.com/amueller/word_cloud)](https://github.com/amueller/word_cloud)

wordcloud 라이브러리는 MIT 라이선스가 있지만 Google에서 True Type 글꼴 인 DroidSansMono.ttf가 포함되어 있습니다. 즉, Apache 라이선스입니다. 글꼴은 절대로 완전한 것이 아니며 WordCloud 개체를 만들 때 font_path 변수를 설정하여 다른 글꼴을 사용할 수 있습니다.

In [13]:
```
from wordcloud import WordCloud, STOPWORDS

import numpy as np
from PIL import Image
```

In [14]:
```
text = open("data/alice.txt").read()
alice_mask = np.array(Image.open("img/alice_color.png"))

stopwords = set(STOPWORDS)  # 불용어 처리
stopwords.add("said")       # 불용어 처리
```

In [16]:
```
import matplotlib.pyplot as plt
import platform
```

한글 폰트 글꼴 설정

In [21]:
```
from matplotlib import font_manager, rc
path = "C:/Windows/Fonts/malgun.ttf"
if platform.system() == "Windows":
    font_name = font_manager.FontProperties(fname=path).get_name()
    rc('font', family=font_name)
elif platform.system()=="Darwin":
    rc('font', family='AppleGothic')
else:
    print("Unknown System")
```

In [22]:
```
wc = WordCloud(background_color="white", max_words=2000, mask=alice_mask,
               stopwords=stopwords, contour_width=3, contour_color='steelblue')

# generate word cloud
wc.generate(text)
wc.words_
```

```
 'eye': 0.09863013698630137,
 'Oh': 0.0958904109589041,
 'came': 0.0958904109589041,
 'last': 0.09315068493150686,
 'nothing': 0.09315068493150686,
 'tell': 0.09315068493150686,
 'day': 0.09041095890410959,
 'large': 0.09041095890410959,
 'great': 0.09041095890410959,
 'hand': 0.09041095890410959,
 'found': 0.08767123287671233,
 'long': 0.08767123287671233,
 'door': 0.08767123287671233,
 'looking': 0.08767123287671233,
 'word': 0.08493150684931507,
 'March Hare': 0.08493150684931507,
 'made': 0.0821917808219178,
 'heard': 0.0821917808219178,
 'look': 0.07945205479452055,
```

앨리스 그림 확인

In [25]:
```
plt.figure(figsize=(15,8))
plt.imshow(alice_mask, cmap=plt.cm.gray, interpolation='bilinear')
plt.axis('off')
plt.show()
```

In [30]:
```
plt.figure(figsize=(15,8))
plt.imshow(wc, interpolation='bilinear')
plt.axis('off')
plt.show()
```

In [ ]:
```

```