

[실습] 스티브 잡스의 스텐퍼드대학교 졸업식 축사 WordCloud 만들기

1. 연설문 텍스트 데이터 구성

- > 'speech_jobs_KOR.txt' 파일을 읽어들여서 데이터 프레임 구성 후 출력 확인
- > 불필요한 문자 제거하기 : 한글을 제외한 문자를 제거
- > 명사만 추출하여 리스트로 구성

2. 단어 길이 데이터 프레임 만들기

- > 각 단어별 길이를 데이터 프레임으로 변환 (단어의 길이 컬럼 추가)
- > 길이가 2 이상인 문자만으로 재구성 : 길이 순서 정렬

3. 단어 빈도수 데이터 프레임 만들기

- > 단어 빈도수 순으로 프레임 구성
- > 빈도수 상위 20 단어 추출
- > 빈도수 상위 20 막대 그래프 그리기

4. 워드 클라우드 만들기

- > [준비] 데이터 프레임을 딕셔너리로 변환
- >> 단어(word) 컬럼을 index로 전환 후 단어와 빈도수를 딕셔너리로 변환
- > 변환된 딕셔너리 10개 item만 출력 확인
- > [준비] 사용할 폰트 지정
- > WordCloud 이미지 출력

5. 직접 만든 이미지로 워드 클라우드 만들기

- > PIL : Python Imaging Library 패키지 사용 이미지 읽어오기
- > 컬러를 원하는 맵 색상으로 설정
- > WordCloud 이미지 출력

In []:

```
In [1]: ##### 1. 연설문 텍스트 데이터 구성
##### > 'speech_jobs_KOR.txt' 파일을 읽어들여서 데이터 프레임 구성 후 출력 확인
```

Out[1]: '
저는 오늘 세계 최고 명문대 중 하나로 꼽히는 이 대학을 졸업하는 여러분들과 함께함을 영광으로 생각합니다.
저는 대학을 졸업하지 못했습니다. 솔직히 말해, 이번이 제가 대학 졸업식을 이렇게 가까이서 보는 것은 처음입니다.
오늘 저는 여러분들에게 제 삶 중에 있었던 3가지 이야기를 하려고 합니다. 그렇게 대단한 이야기는 아니고 그저 3가지 이야기입니다.
첫 번째 이야기는 ‘이어지는 순간들’ 것에 관한 것입니다.
저는 입학한지 6개월만에 리드 대학을 자퇴했지만, 그 후 18개월동안 청강하여 학교에 머물렀습니다.
제가 왜 자퇴했을까요?
이야기는 제가 태어나기 전부터 시작됩니다.
제 생모는 젊은 미혼모 대학생이었는데 저를 낳으면 다른 사람에게 입양을 시키기로 결심했습니다.
그녀는 대학을 졸업한 사람이 저의 양부모가 되기를 간절히 원했습니다. 그래서 저는 태어나자마자 변호사 가정에 입양되기로 모든 계획이 확정되어져 있었습니다.
내가 나타나기 전까지 절 입양시키기로 모든 계획이 확정되어져 있었습니다.
대기자 명단에 있던 양부모님들은 한밤중에 “어떡하죠? 예정에 없던 사내아이가 태어났는데 그래도 입양하실 건가요?”라는 전화를 받았습니다.
그들은 “물론이죠”하고 대답했습니다.
그런데 나중에 알고보니 양어머니는 대졸자도 아니었고, 양아버지는 고등학교도 졸업하지 않았습니다.
친어머니는 입양동의서에 사인하기를 거부했습니다.
생모는 양부모님들이 저를 꼭 대학까지 보내주겠다고 약속한 후 몇개월이 지나서야 마음이 누그러져 받아들였습니다.
17년 후 저는 대학에 가게 되었습니다.
그러나 저는 순진하게도 바로 이곳, 스탠포드의 학비와 맞먹는 값비싼 학교를 선택했습니다.
평범한 노동자였던 부모님이 힘들게 모아뒀던 돈이 모두 제 학비로 들어갔습니다.
6개월 후, 저는 대학 공부에 대하여 그만한 가치를 느낄 수 없었습니다.
제가 진정으로 인생에서 원하는 게 무엇인지, 그리고 대학이 그것을 실현하는데 어떻게 도움이 될 수 있을지 판단할 수 없었습니다.
그런데도 저는 양부모님들이 평'

In [2]: ##### 1. 연설문 텍스트 데이터 구성
> 불필요한 문자 제거하기 : 한글을 제외한 문자를 제거

Out[2]: '
저는 오늘 세계 최고 명문대 중 하나로 꼽히는 이 대학을 졸업하는 여러분들과 함께함을 영광으로 생각합니다. 저는 대학을 졸업하지 못했습니다. 솔직히 말해. 이번이 제가 대학 졸업식을 이렇게 가까이서 보는 것은 처음입니다. 오늘 저는 여러분들에게 제 삶 중에 있었던 가지 이야기를 하려고 합니다. 그렇게 대단한 이야기는 아니고 그저 가지 이야기입니다. 첫 번째 이야기는 이어지는 순간들 것에 관한 것입니다. 저는 입학한지 개월만에 리드 대학을 자퇴했지만 그 후 개월동안 청강하여 학교에 머물렀습니다. 제가 왜 자퇴했을까요. 이야기는 제가 태어나기 전부터 시작됩니다. 제 생모는 젊은 미혼모 대학생이었는데 저를 낳으면 다른 사람에게 입양을 시키기로 결심했습니다. 그녀는 대학을 졸업한 사람이 저의 양부모가 되기를 간절히 원했습니다. 그래서 저는 태어나자마자 변호사 가정에 입양되기로 모든 계획이 확정되어져 있었습니다. 내가 나타나기 전까지 절 입양시키기로 모든 계획이 확정되어져 있었습니다. 대기자 명단에 있던 양부모님들은 한밤중에 어떡하죠. 예정에 없던 사내아이가 태어났는데 그래도 입양하실 건가요 라는 전화를 받았습니다. 그들은 물론이죠 하고 대답했습니다. 그런데 나중에 알고보니 양어머니는 대졸자도 아니었고 양아버지는 고등학교도 졸업하지 않았습니다. 친어머니는 입양동의서에 사인하기를 거부했습니다. 생모는 양부모님들이 저를 꼭 대학까지 보내주겠다고 약속한 후 몇개월이 지나서야 마음이 누그러져 받아들였습니다. 년 후 저는 대학에 가게 되었습니다. 그러나 저는 순진하게도 바로 이곳 스탠포드의 학비와 맞먹는 값비싼 학교를 선택했습니다. 평범한 노동자였던 부모님이 힘들게 모아뒀던 돈이 모두 제 학비로 들어갔습니다. 개월 후 저는 대학 공부에 대하여 그만한 가치를 느낄 수 없었습니다. 저는 제가 진정으로 인생에서 원하는 게 무엇인지 그리고 대학이 그것을 실현하는데 어떻게 도움이 될 수 있을지 판단할 수 없었습니다. 그런데도 저는 양부모님들이 평'

In [3]: ##### 1. 연설문 텍스트 데이터 구성
> 명사만 추출하여 리스트로 구성
형태소분석 인스턴스 hannanum 만들기

연설문에서 명사 추출하기

Out[3]: ['저', '오늘', '세계', '최고', '명문대', '중', '하나', '대학', '졸업', '여러분들']

In [7]: ##### 2. 단어 데이터프레임 만들기
> 각 단어를 데이터프레임으로 변환

```
#### > 단어의 길이 컬럼 추가
```

```
#### > 길이가 2 이상인 문자만으로 재구성 > 길이순 정렬
```

Out[7]:

	word	len
1	오늘	2
2	세계	2
3	최고	2
4	명문대	3
6	하나	2
...
1018	자신	2
1019	그러길	3
1020	졸업	2
1021	출발	2
1022	여러분	3

693 rows × 2 columns

In [6]: #### 3. 단어 빈도수 데이터프레임 만들기
> 단어 빈도수 순으로 프레임 구성

Out[6]:

	word	count
50	그것	17
83	당신	14
286	인생	12
88	대학	11
170	사람	9
...
155	복사	1
154	변환점	1
152	변호사	1
151	벽돌	1
410	히치하이킹	1

411 rows × 2 columns

In [8]:

```
#### 3. 단어 빈도수 데이터프레임 만들기  
#### > 빈도수 상위 20 단어 추출
```

Out[8]:

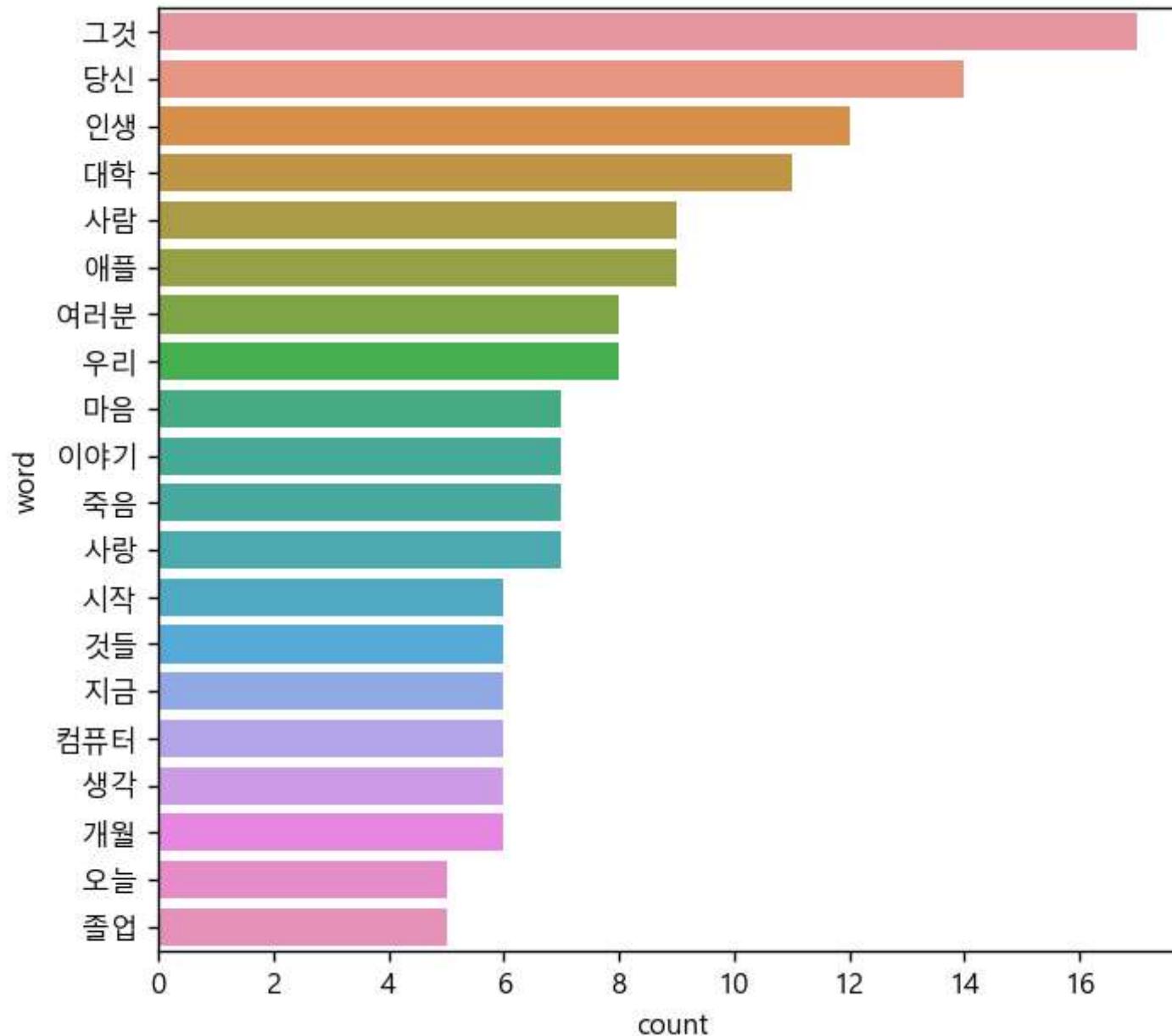
	word	count
50	그것	17
83	당신	14
286	인생	12
88	대학	11
170	사람	9
234	애플	9
244	여러분	8
262	우리	8
109	마음	7
280	이야기	7
321	죽음	7
171	사랑	7
215	시작	6
23	것들	6
329	지금	6
365	컴퓨터	6
181	생각	6
13	개월	6
258	오늘	5
316	졸업	5

In [9]:

```
#### 3. 단어 빈도수 데이터프레임 만들기  
#### > 빈도수 상위 20 막대 그래프 그리기
```

```
C:\Users\ADMIN\anaconda3\lib\site-packages\scipy\__init__.py:155: UserWarning: A NumPy version >=1.18.5 and <1.25.0 is required for
this version of SciPy (detected version 1.25.2)
... warnings.warn(f"A NumPy version >={np_minversion} and <{np_maxversion}"
C:\Users\ADMIN\anaconda3\lib\site-packages\seaborn\_\_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will be
removed in a future version. Use isinstance(dtype, CategoricalDtype) instead
... if pd.api.types.is_categorical_dtype(vector):
C:\Users\ADMIN\anaconda3\lib\site-packages\seaborn\_\_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will be
removed in a future version. Use isinstance(dtype, CategoricalDtype) instead
... if pd.api.types.is_categorical_dtype(vector):
C:\Users\ADMIN\anaconda3\lib\site-packages\seaborn\_\_oldcore.py:1498: FutureWarning: is_categorical_dtype is deprecated and will be
removed in a future version. Use isinstance(dtype, CategoricalDtype) instead
... if pd.api.types.is_categorical_dtype(vector):
<Axes: xlabel='count', ylabel='word'>
```

Out[9]:



```
In [23]: #### 4. 워드 클라우드 만들기  
#### > [준비] 데이터 프레임을 딕셔너리로 변환
```

```
In [24]: ## [확인] 딕셔너리에서 item 10개만 출력
```

```
그것 : 17  
당신 : 14  
인생 : 12  
대학 : 11  
사람 : 9  
애플 : 9  
여러분 : 8  
우리 : 8  
마음 : 7  
이야기 : 7
```

```
In [15]: ##### 4. 워드 클라우드 만들기  
## > [준비] 사용할 폰트 지정
```

```
##### > WordCloud 객체 만들기
```

```
In [16]: ##### 4. 워드 클라우드 만들기  
## > 워드 클라우드 이미지 만들기
```

```
## > WordCloud 이미지 출력
```

```
Out[16]: <matplotlib.image.AxesImage at 0x26d2b53f1f0>
```



```
In [18]: ##### 5. 직접 만든 이미지로 워드 클라우드 만들기  
## > PIL : Python Imaging Library 패키지 사용 이미지 읽어오기  
## > 컬러를 맵 색상으로 설정  
## WordCloud 이미지 출력
```

Out[18]: <matplotlib.image.AxesImage at 0x26d2b5878b0>



In []: