<억압과 검열은 아무 데이터를 남기지 않는가?>

빅데이터학 3차 과제

경제학/빅데이터사이언스 20180590 이윤주

0. 연구 주제

-전세계적으로 억압과 검열의 역사가 존재해왔다. 특정 집단은 그들의 세력을 굳건히 하기 위하여 그들의 이념과 맞지 않는 것들을 억압했다. 반대 이념의 그림, 글 등의 출간을 금지한 것이 대표적인 행동이었다. 그러나 과연 그들이 제거하고 싶어한 문화가 역사 속에서도 제거됐을까? 후대 사람들에게도 숨길 수 있을까? 이 점을 ngram 데이터를 통하여 탐구해보고자 한다.

-1930~40년대 독일은 나치정권이 지배하고 있었고, 샤갈, 고갱, 칸딘스키 등의 미술가를 퇴출시키려 했다. 따라서 독일 미술가 6명의 ngram 빈도수 추이를 독일, 미국 두 개 코퍼스에서 살펴볼 것이다. 이를 위해 GoogleNgrams 함수를 만들고 영어, 독일어 코퍼스에서 미술가 6명의 빈도를 산출한다.

```
import requests
from ast import literal_eval
str_dict="{'a':3,'b':5}"
literal_eval(str_dict)
import pandas as pd
import re, os
import matplotlib.pyplot as plt
```

```
#os.getcwd()='/'.('EU_painters.txt')
painter=pd.read_csv('EU_painters.txt', names=['painter'])
actor=pd.read_csv('US_suppression.txt', names=['actor'])
p_names=[painter['painter'][i] for i in range(0,len(painter))]
a_names=[actor['actor'][j] for j in range(0,len(actor))]
a_names
```

```
['John Howard Lawson',
 'Albert Maltz',
 'Dalton Trumbo',
 'Alvah Bessie',
 'Edward Dmytryk',
 'Herbert Biberman',
 'Lester Cole',
 'Ring Lardner Jr.',
 'Samuel Ornitz',
 'Adrian Scott']
```

```
corpora=dict(eng_2019=26, ger_2019=31)
def GoogleNgrams(A,B=1900, C=1970,D='eng_2019'):
    search=dict(content=A, year_start=B, year_end=C, corpus=corpora[D])
    a=requests.get('http://books.google.com/ngrams/graph', params=search)
    b=literal_eval(re.findall('ngrams.data = (.*?);\\n',a.text)[0])
    c=pd.DataFrame({i['ngram']: i['timeseries'] for i in b},
                   index=range(search['year_start'],
                               search['year_end']+1))
    return(c)
```

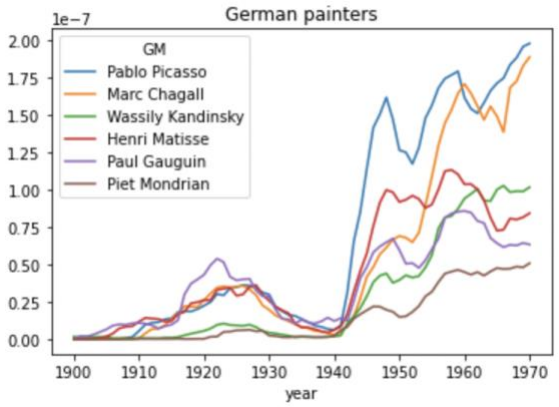
함수를 실행시키면,

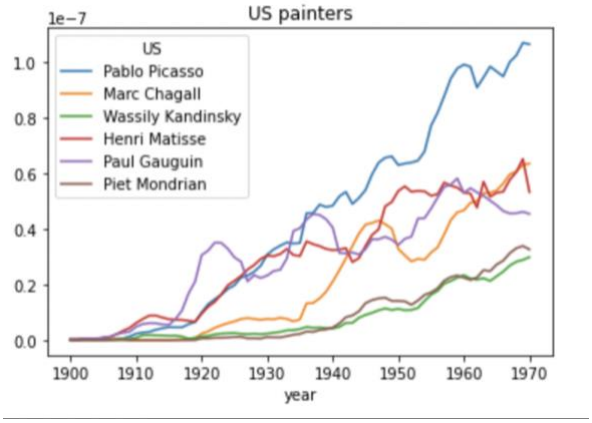
```
gm=GoogleNgrams('Pablo Picasso,Marc Chagall,Wassily Kandinsky,Henri Matisse,Paul Gauguin,Piet Mondrian',D='ger_2019'
us=GoogleNgrams('Pablo Picasso,Marc Chagall,Wassily Kandinsky,Henri Matisse,Paul Gauguin,Piet Mondrian', D='eng_2019

gm.columns.name='GM'
gm.index.name='year'
us.columns.name='US'
us.index.name='year'
gm.plot()
plt.title('German painters')

us.plot()
plt.title('US painters')
```

1930~40년대에 독일 코퍼스에서 눈에 띄는 감소가 있음이 보인다. 반면, 미술에 대한 억압이 없었던 미국에서는 감소 추세 없이 점점 증가하는 모습을 띈다.



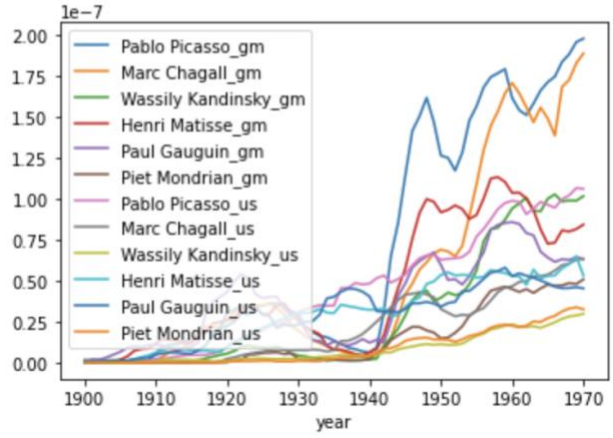


1. merge

두 개의 코퍼스가 따로 출력되어 비교하기에 불편함이 있다. 따라서 merge 함수를 통해 두 코퍼스를 합쳐보았다.

```
painters=pd.merge(gm,us, left_index=True, right_index=True, suffixes=['_gm','_us'])
painters.plot()
```

```
<AxesSubplot:xlabel='year'>
```



한 차트에 표시가 가능해졌지만 12개의 요소가 섞여 어떤 것이 어느 나라 것인지 구분이 힘들다.

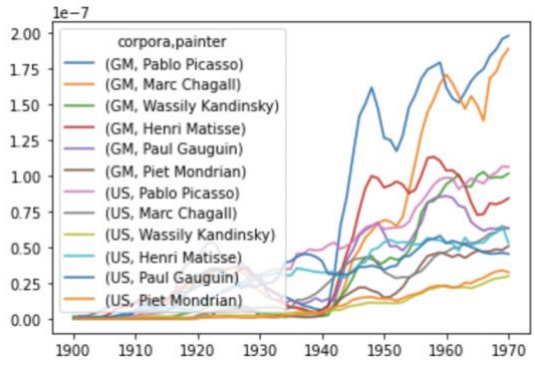
2. hierarchical indexing

나라 별 구분을 주었다. 이제 독일 코퍼스의 화가들, 미국 코퍼스의 화가들을 구분해서 보기 쉬워졌다.

```
painters.columns=pd.MultiIndex.from_product([['GM','US'],['Pablo Picasso','Marc Chagall','Wassily Kandinsky','Henri
                                names=['corpora','painter'])
painters.head()
```

corpora	GM						US					
painter	Pablo Picasso	Marc Chagall	Wassily Kandinsky	Henri Matisse	Paul Gauguin	Piet Mondrian	Pablo Picasso	Marc Chagall	Wassily Kandinsky	Henri Matisse	Paul Gauguin	Piet Mondrian
year												
1900	2.875549e-10	1.479207e-10	1.593368e-09	1.396341e-10	2.626670e-10	0.0	2.453135e-10	1.226279e-10	3.088335e-11	3.036890e-10	2.711678e-10	0.000000e+00
1901	2.300439e-10	1.183366e-10	1.464251e-09	1.117073e-10	1.916148e-09	0.0	1.962508e-10	9.810233e-11	2.470668e-11	2.429512e-10	3.308794e-10	0.000000e+00

painter.plot()으로 차트를 그리면 GM, US로 구분은 되지만 여전히 보기 불편하다.



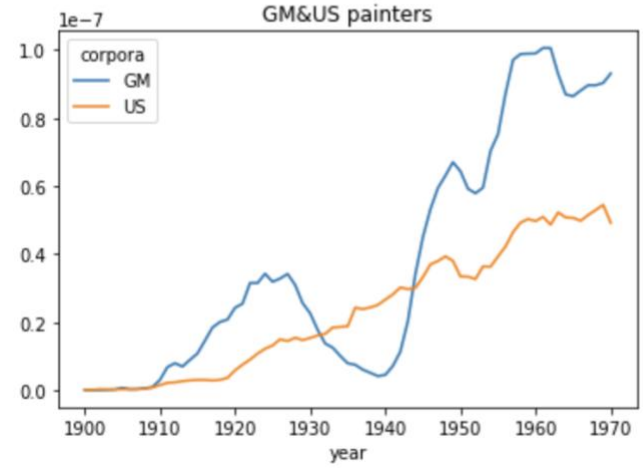
3. groupby

따라서 groupby를 통해 같은 차트에서 독일어 빈도와 영어 빈도의 중앙값만 보기로 한다.

순탄히 올라가는 영어 코퍼스에 비해 독일어 코퍼스는 1930년에 뚜렷한 감소가 있다.

```
painters.groupby(level='corpora',axis=1).median().plot()
plt.title('GM&US painters')
```

```
Text(0.5, 1.0, 'GM&US painters')
```



4. agg

십년 단위로는 어떤 변화가 있는지 보기 위해 십년으로 끊은 인덱스 열을 삽입하고 period라 이름 붙인다. 같은 인덱스 별로 그룹 묶고, agg를 이용하여 중앙값을 계산하면 다음과 같다.

```
painters['period']=[str(i)[:-1]+'0'
                    for i in painters.index]
painters.head()
painters.groupby(by=painters.period, axis=0).agg('median').head()
```

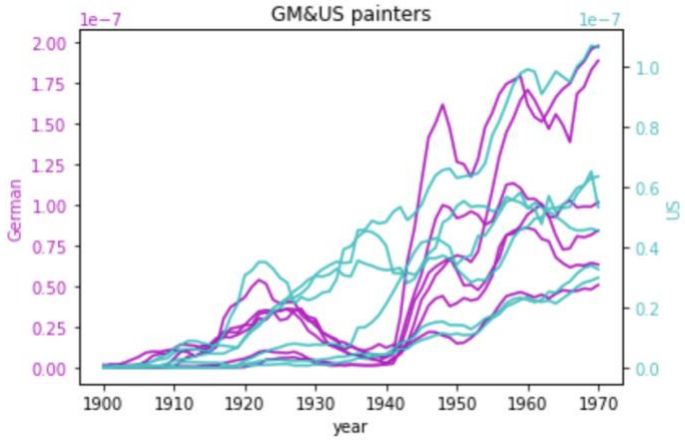
corpora	GM						US					
painter	Pablo Picasso	Marc Chagall	Wassily Kandinsky	Henri Matisse	Paul Gauguin	Piet Mondrian	Pablo Picasso	Marc Chagall	Wassily Kandinsky	Henri Matisse	Paul Gauguin	Piet Mondrian
period												
1900	1.780102e-10	9.156999e-11	1.451503e-09	1.128765e-09	5.629232e-09	0.000000e+00	3.097178e-10	4.484191e-10	3.049400e-11	5.633328e-10	7.945899e-10	3.270095e-11
1910	1.358316e-08	1.137063e-08	9.873081e-10	1.403974e-08	1.060731e-08	1.775584e-10	4.432543e-09	4.002589e-11	1.582483e-09	7.424314e-09	6.103274e-09	1.430488e-11
1920	3.264295e-08	3.394428e-08	9.054490e-09	3.145123e-08	4.107480e-08	5.139714e-09	1.906370e-08	6.414474e-09	2.213583e-09	2.048965e-08	3.013263e-08	7.668478e-10
1930	1.450737e-08	9.086802e-09	1.571462e-09	9.063027e-09	1.306203e-08	1.430501e-09	3.498624e-08	7.811362e-09	3.801588e-09	3.184977e-08	3.971132e-08	2.090768e-09
1940	9.982220e-08	3.454613e-08	2.474889e-08	5.153855e-08	4.458903e-08	1.751519e-08	5.358133e-08	3.915192e-08	8.563703e-09	3.345521e-08	3.439738e-08	1.182609e-08

5. 척도가 다른 두 개의 y축 합치기

```
fig, ax1=plt.subplots()
ax1.set_xlabel('year')
ax1.set_ylabel('German', color='m')
ax1.plot(gm, color='m')
ax1.tick_params(axis='y', labelcolor='m')

ax2=ax1.twinx()
ax2.set_ylabel('US', color='c')
ax2.plot(us, color='c')
ax2.tick_params(axis='y',labelcolor='c')
plt.title('GM&US painters')
```

```
Text(0.5, 1.0, 'GM&US painters')
```



척도가 다른 두 그래프를 같은 기준(더 큰 y값)의 평면에 그리면 y값이 작은 그래프의 추세를 알기 어렵다. 그래서 위 코드로 코퍼스 크기 별로 y축을 다르게 하여 한 그래프에 나타냈다. 이로 인해 코퍼스 크기가 더 작은 독일어 코퍼스의 추세를 더욱 자세히 볼 수 있게 되었다.

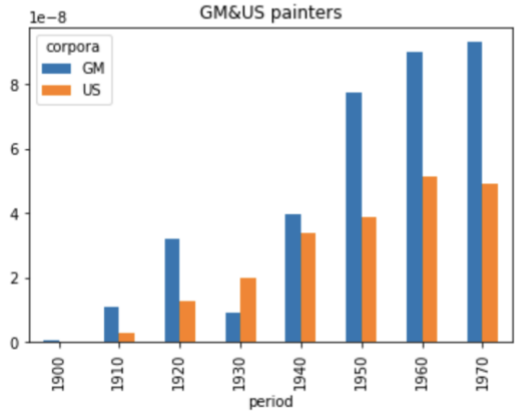
6. 다른 유형의 차트

-bar : 추세의 수치 차이를 확인할 수 있다.

독일어 GM bar를 보면 1930년이 되자 급감했고, 미국 US bar는 감소 추세 없이 꾸준히 증가했다.

```
painters.groupby(painters['period'],axis=0).median().groupby(level='corpora',axis=1).median().plot.bar()
plt.title('GM&US painters')
```

```
Text(0.5, 1.0, 'GM&US painters')
```

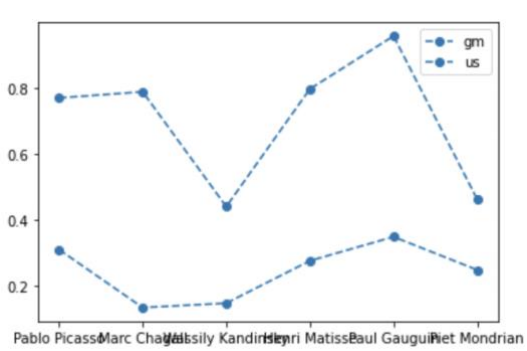


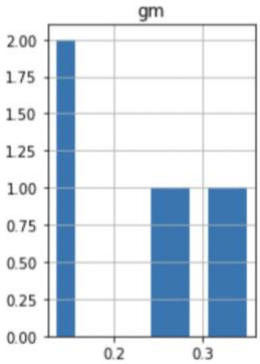
-히스토그램과 "억압지수

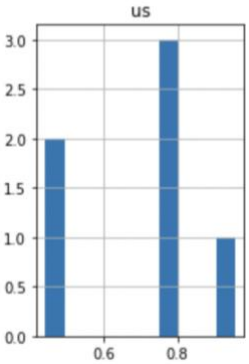
영어, 독일어 코퍼스 별로 '(1933-1945평균빈도)/(1925-1933, 1955-1965) 평균빈도'의 식을 이용하여 직접 억압지수를 계산해보았다. 그리고 억압지수의 빈도수를 히스토그램으로 나타내어 1에 가까운 것이 많은지 확인할 수 있었다. 억압지수=1이면 변화 없음, 억압지수<1이면 억압 받았던 것, 억압지수>1이면 인위적으로 명성이 높아진 것이다. 독일어 코퍼스는 0.5 이하의 빈도가 대부분이고 그 중에서도 0에 가까운 것이 가장 많다. 반면 영어 코퍼스는 1에 가까운 값이 많다.

```
#억압지수
gm1=gm.loc[1933:1945].mean()
gm2=(gm.loc[1925:1933].mean()+gm.loc[1955:1965].mean())/2
suppression=gm1/gm2
us1=us.loc[1933:1945].mean()
us2=(us.loc[1925:1933].mean()+us.loc[1955:1965].mean())/2
suppression_us=us1/us2

suppression,suppression_us
sup=pd.DataFrame({'gm':suppression, 'us':suppression_us})
sup.plot(color='C0',linestyle='--' ,marker='o')
sup.hist()
```







7. 추세 패턴의 인문사회과학적 의미

-독일어 코퍼스에서 1930년도에 예술가들의 언급 빈도가 현저히 낮아졌다가, 1940년이 지나자 다시 증가한 것은 나치 집권 시대에 예술 검열과 억압으로 인해 언급되는 양이 적었다가 집권이 끝난 후 증가한 것이다. 이렇게 Ngram의 특정 분야 직업인의 갑작스러운 감소와 증가는 그 시대의 정치, 사회, 문화적 변화를 반영한다. 시대를 반영하는 경우는 독일 뿐만 아니라 미국에서도 찾아볼 수 있었는데, 미하원 비미활동위원회에게 증언을 거부한 10명의 영화인도 그러했다. 독일어 코퍼스와 영어 코퍼스를 비교해보았더니 이번에는 반대로 영어 코퍼스에서 특정 감소 구간이 나타났다. 블랙리스트에 올랐던 1940년대에는 10인 전부 빈도가 감소했다가 1960년 이후에는 점차 증가한다. 이렇게 ngram의 시대 반영은 독일 뿐만 아니라 전세계에 걸쳐 일어난다고 할 수 있다.

이제 0번에 대한 답을 할 수 있게 되었다. 특정 시대에 숨기고 억압하려고 했던 정보는 영원히 숨겨질 수 있는가? 답은 아니다. 오히려 '숨긴 자국'을 통해 그들의 만행이 드러나게 되었다. 대부분의 사람들이 데이터를 볼 때 빈도가 높은 것에 집중하는 경향이 있지만, 오히려 적은 빈도에 집중했을 때 발견할 수 있는 것이 있다. 구글 엔그램 뷰어 창시자들이 그랬던 것처럼, 하나의 '숨긴 자국', 즉 이상하게 적은 빈도에 집중하면 생각치 못했던 것이 발견되는 것이다. 이는 수학자, 공학자가 아닌 인문사회과학자들의 역할이며, 빅데이터 분석에 없어선 안 될 역할이라고 생각한다.

(미국 배우 10인 코퍼스 분석)

```
#미국
gm_act=GoogleNgrams('John Howard Lawson, Albert Maltz, Dalton Trumbo, Alvah Bessie, Edward Dmytryk, Herbert Biber
                B=1900,C=2000, D='ger_2019')
us_act=GoogleNgrams('John Howard Lawson, Albert Maltz, Dalton Trumbo, Alvah Bessie, Edward Dmytryk, Herbert Biber
                B=1900,C=2000, D='eng_2019')

gm_act.plot(), us_act.plot()
```

```
fig, ax1=plt.subplots()
ax1.set_xlabel('year')
ax1.set_ylabel('US', color='C1')
ax1.plot(us_act, color='C1')
ax1.tick_params(axis='y', labelcolor='C1')

ax2=ax1.twinx()
ax2.set_ylabel('German', color='C0')
ax2.plot(gm_act, color='C0')
ax2.tick_params(axis='y',labelcolor='C0')
```