<빅데이터학 - 3차 과제>

유럽 화가들의 Ngram분포로 분석한 나찌시대 억압의 역사

20180359 김서현
사회학과

lily000101@naver.com

목차

1. 연구 목적

본 연구는 빅데이터학 강의에서 다룬 다양한 주제 중 '데이터가 말하는 억압과 검열의 역사'를 직접 Ngram Viewer의 데이터를 통해 직접 검증해 보기 위해 수행했다. 다양한 억압과 검열의 사례들 중 '독일 나찌시대의 억압과 검열'을 여섯 명의 유럽 화가들의 Ngram 분포를 통해 검증해 내는 것이 본 연구의 목적이다.

Ngram Viewer(이하 Ngram)는 Google이 제공하는 어휘/어구 빈도의 시대적 추세 검색 시스템이다. Ngram은 Google Book Scanning Project의 일환으로 구축된 1800년부터 2019년까지 출간된 문헌의 대규모 코퍼스로 이루어져 있으며 이를 코퍼스를 대상으로 어휘 및 어구(n-gram) 사용 빈도의 시대적 추세 변화를 검색할 수 있다.

Ngram 분석을 위하여 사용한 키워드는 여섯 명의 유럽 화가들의 이름이다. 이 화가들은 Pablo Picasso, Marc Chagall, Wassily Kandinsky, Henri Matisse, Paul Gauguin, Piet Mondrian으로, 이 목록은 빅데이터학 강의에서 제공된 EU.painters.txt 파일을 활용하였다. 이 화가들은 나찌집권 시절 '유대인적'이라고 일컬어지며 퇴출된 아방가르드 미술가들이다. 또한 이 화가들의 목록은 1937년 퇴폐작품으로 압수된 작품의 화가들의 목록과도 일치한다. 코드 및 차트 작성을 위해서 jupyter notebook을 사용했으며, Ngram Viewer의 데이터를 읽어들이기 위해서 빅데이터학 강의에서 제공한 GoogleNgrams 함수를 활용했다.

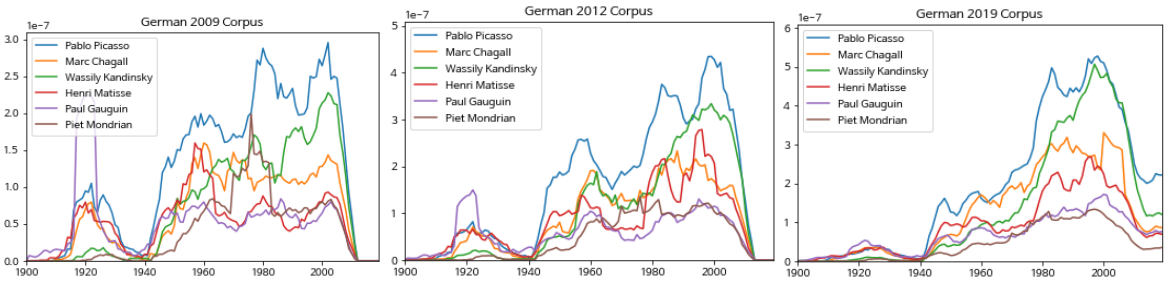
2. Ngram 분석

1) 코퍼스 선정 과정

분석에 이용할 코퍼스를 선정하기 위해 몇 단계의 과정을 거쳤다. 독일어 코퍼스를 사용하기 위해 확인해보니 독일어의 코퍼스는 German 2009, German 2012, German 2019 세 가지가 있었다. 이들 중 적절한 코퍼스를 선택하기 위해 세 코퍼스를 통해 도출된 6인의 화가들의 Ngram 분포를 비교했다.

```
ger09=GoogleNgrams(list_str, B=1900, D='ger_2009')
ger12=GoogleNgrams(list_str, B=1900, D='ger_2012')
ger19=GoogleNgrams(list_str, B=1900, D='ger_2019')
```

사용한 코드[1]



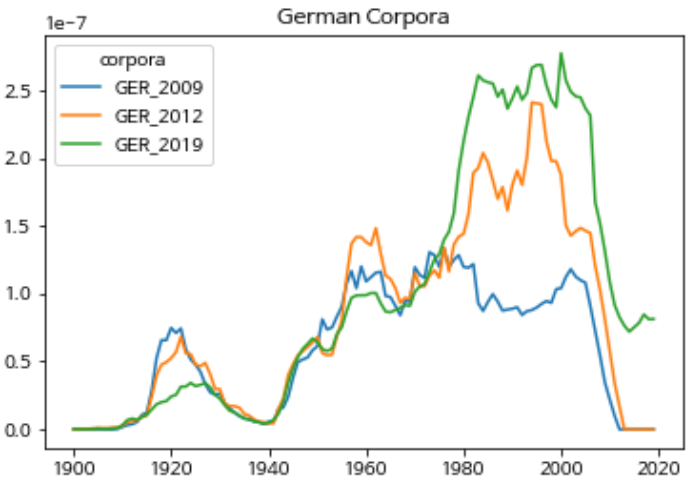
German 2009, German2012, German2019 Ngram 분포

```
In [10]: #세 개의 코퍼스 merge
german_corpus=pd.merge(ger09, ger12, left_index=True, right_index=True)
german_corpus=pd.merge(german_corpus, ger19, left_index=True, right_index=True)

#Hierarchical indexing
german_corpus.columns=pd.MultiIndex.from_product([['GER_2009', 'GER_2012', 'GER_2019'],
                                                  list(painter_list)],
                                                 names=['corpora', 'painters'])

#차트 생성
german_corpus.groupby(level='corpora', axis=1).agg('median').plot()
plt.title('German Corpora')
plt.savefig('E:/서강대학교/6학기/빅데이터학/빅데이터학_3차과제/그래프/german_corpus.png')
```

사용한 코드

[1] list_str은 앞서 읽어온 화가들의 목록을 하나의 문자열('Pablo Picasso, Marc Chagall, Wassily Kandinsky, Henri Matisse, Paul Gauguin, Piet Mondrian')로 바꾼 것이다.

세 개의 코퍼스를 보다 쉽게 비교분석 하기 위해 merge함수를 이용해 코퍼스들을 하나의 Data Frame으로 병합했다. 이후 hierarchical indexing을 사용해 각 층위별로 index를 달아주었다. 이후 groupby 함수를 이용해 코퍼스를 기준으로 grouping 해 준 후 agg 함수를 이용하여 중앙값을 생성하여 차트를 만들었다.



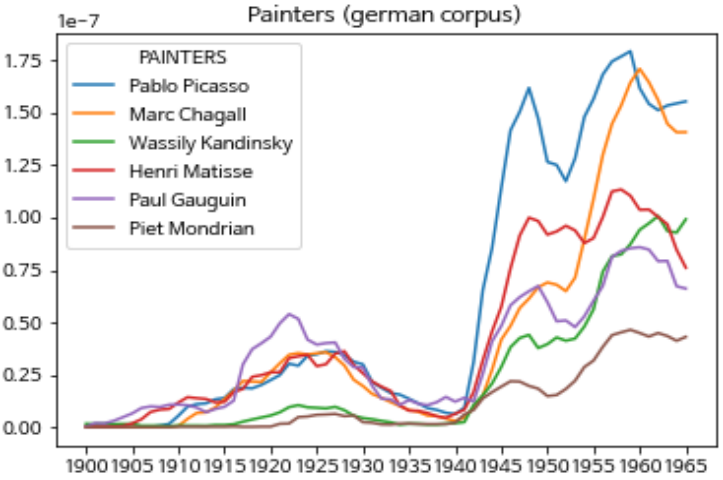
이렇게 생성된 차트를 보면 세 개의 차트 모두 유사한 패턴을 보인다. 특히 German 2012 코퍼스와 German 2019 코퍼스는 거의 일치하는 패턴을 보이고 있다. 따라서 본 연구에서 사용할 코퍼스는 German 2019 로 선정하였다. 위와 같은 방식으로 다른 언어에 대한 코퍼스도 2019 년도의 코퍼스를 선정하였다.

2) 독일어 코퍼스만의 특징

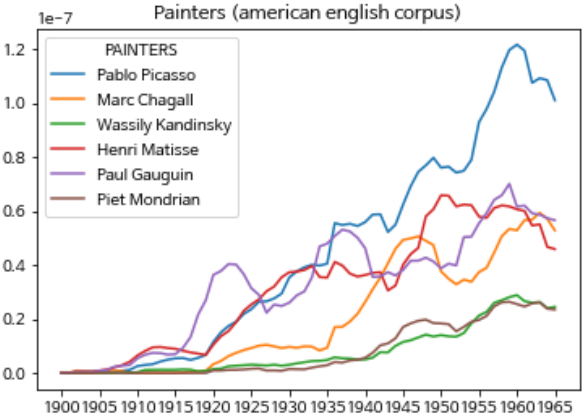
독일어 코퍼스에서만 나타나는 특징을 확인하기 위해 독일어, 미국 영어, 스페인어 코퍼스에서 6인의 화가를 검색했다.

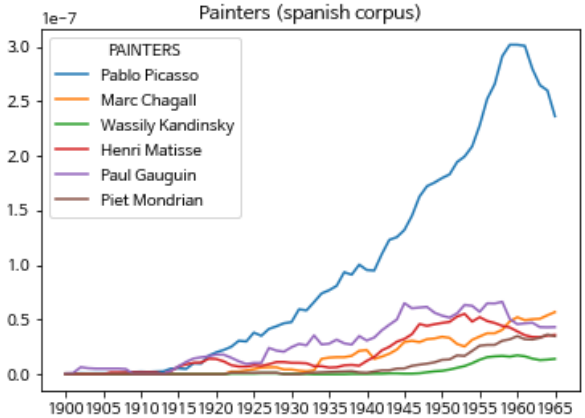
```
#독일어 코퍼스
gerCo=GoogleNgrams(list_str, B=1900, C=1965, D='ger_2019')
gerCo.columns.name='PAINTERS'

#차트 생성
gerCo.plot()
plt.title('Painters (german corpus)')
plt.xticks(range(1900, 1966, 5))
plt.savefig('E:/서강대학교/6학기/빅데이터학/빅데이터학_3차과제/그래프/german_corpus_painters.png')
```

검색 기간은 1900년에서 1965년으로 설정하였다. 이 기간은 독일의 미술활황기인 1920년대, 나찌집권기인 1933년~1945년, 나찌집권기 이후를 모두 포괄하기 위해 설정되었다.



차트를 보면 미술 활황기인 1920년대에 빈도가 증가하다가 1930년대가 되자 급속도로 감소하는 모습을 보인다. 이후 이러한 추세는 1940년대에 들어서자 급상승한다. 이러한 특징이 독일어의 코퍼스에서만 나타나는 것인지 확인하기 위해 미국 영어와 스페인어 코퍼스를 검색해 보았다. 이들의 기간 역시 동일하게 1900년부터 1965년으로 설정했다.





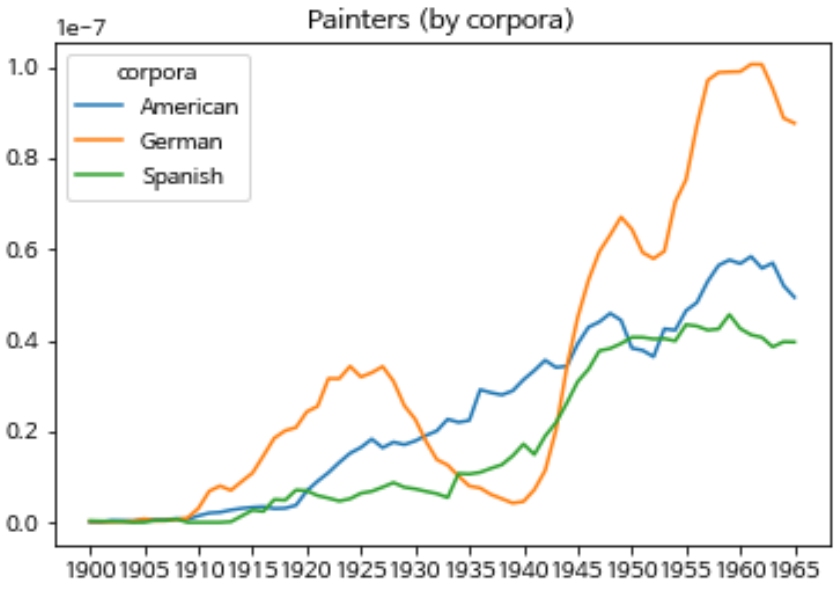
두 차트를 보면 대략 알 수 있듯이 해당 기간, 즉 1933년부터 1945년에 빈도가 급감하는 추세는 미국과 스페인의 코퍼스에서는 찾을 수 없었다. 이를 통해 이러한 추세가 독일어 코퍼스에서만 발견되는 것을 예측할 수 있었다.

3) 독일 – 미국 – 스페인 간의 Ngram 척도 비교

(1) 독일과 미국 간의 비교

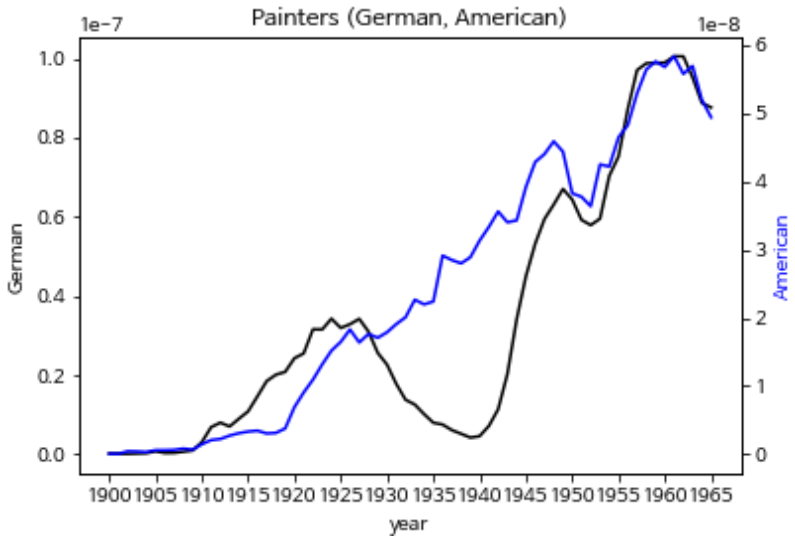
국가별 코퍼스 간의 척도를 비교하기 위해 우선 세 코퍼스를 병합하고 hierarchal indexing을 하여 중앙값으로 차트를 생성했다. 이 때 평균이 아닌 중앙값으로 차트를 생성한 이유는 표본의 수가 여섯 개로 매우 적은 상황에서 중앙값이 지나치게 크거나 작은 값의 영향을 비교적 덜 받기 때문이다. [2]

```
#세 개의 코퍼스 merge
painters_3corpora=pd.merge(gerCo, usCo, left_index=True, right_index=True)
painters_3corpora=pd.merge(painters_3corpora, spaCo, left_index=True, right_index=True)

#Hierarchical Indexing
painters_3corpora.columns=pd.MultiIndex.from_product([['German', 'American','Spanish'],
                                                      list(painter_list)],
                                                     names=['corpora', 'painters'])

#차트 생성
painters_3corpora.groupby(level='corpora', axis=1).median().plot()
plt.title('Painters (by corpora)')
plt.xticks(range(1900, 1966, 5))
plt.savefig('E:/서강대학교/6학기/빅데이터학/빅데이터학_3차과제/그래프/painters_3corpora.png')
```



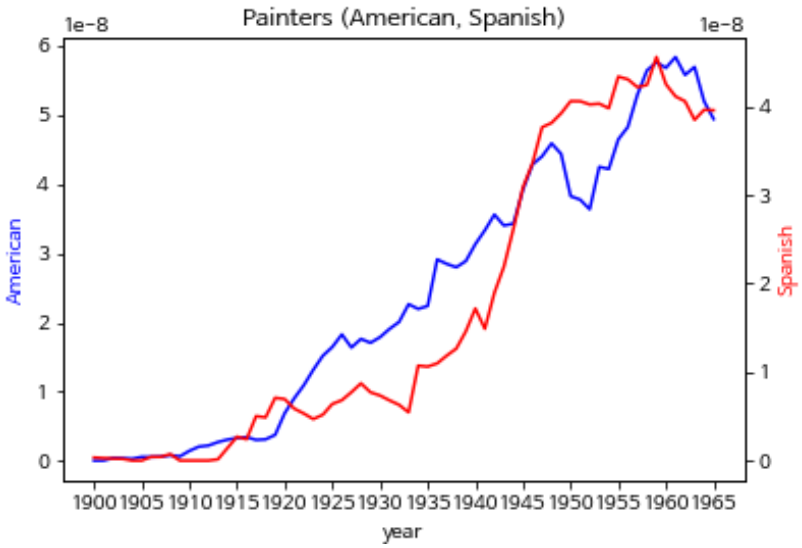
차트를 보면 독일어 코퍼스에서만 나찌집권 시기에 급감하는 추세가 보인다. 이러한 추세를 척도와 함께 확인하기 위해 척도가 다른 코퍼스를 하나의 차트로 병합했다. 독일어 코퍼스에서 화가들의 빈도 중앙값은 최대가 1.1e-7 미만이고, 미국 영어 코퍼스에서는 6.1e-8 미만이며, 스페인어 코퍼스에서 또한 4.1e-8로 세 코퍼스 모두 척도의 차이가 상당하다. 두 차트를 병합하기 위해 각 국가별 코퍼스의 중앙값을 german_median, american_median, spanish_median으로 할당했다. 우선 독일과 미국의 코퍼스를 병합했다.

[2] painter_list는 읽어온 화가 목록을 ndarray 형태로 저장한 것이다.



차트를 보면 1930년부터 1940년 초반까지 미국과 달리 독일의 코퍼스에서만 갑작스레 급감하는 추세를 발견할 수 있다. 또한 1940년대 초·중반에 들어서자 역시 독일어 코퍼스에서만 빈도가 급상승한다.

(2) 미국과 스페인 간의 비교



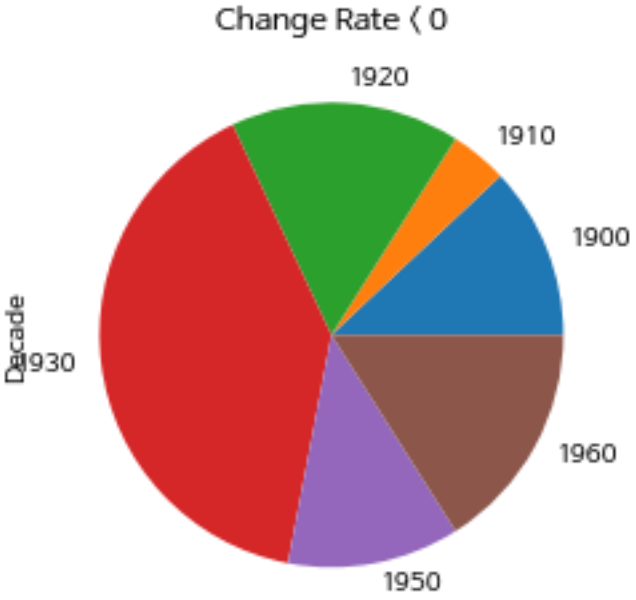
이러한 특징이 독일어 코퍼스가 유일한지, 그리고 미국 영어 코퍼스의 추세가 일반적인 것이 맞는지 확인하기 위해 미국 영어와 스페인어 코퍼스를 병합하였다. 미국 영어와 스페인어 코퍼스의 차트를 보면 두 차트가 거의 동일한 추세로 흘러간다는 것을 확인할 수 있다. 또한 나찌집권 기간의 급감하고, 집권시기 이후의 급상승하는 형태도 찾을 수 없다. 결국 이러한 추세는 독일어 코퍼스에서만 등장하는 특징임을 알 수 있다. 이 특징을 통해 1930년대 실시된 분서운동과 문화 탄압 및 검열의 흔적을 확인할 수 있다. 앞서 언급했듯이 Ngram은 문헌을 통해 구축된 코퍼스를 기반으로 제공되는 시스템이다. 이러한 특징을 고려했을 때 나찌집권 시기에 독일어 코퍼스에서만 빈도가 감소한 것은 나찌의 문화 탄압과 검열에 대한 역사적 사실을 데이터가 잘 반영하고 있음을 알 수 있다.

4) 10년 단위 별 변동률 감소의 빈도

차트로 확인한 감소의 흔적이 맞는지 보다 정확히, 그리고 수치를 통해 알아보기 위해 변동률을 확인해 보았다. 이때 급감하는 구간은 1930년대임을 확인했으므로 10년 단위로 독일어 코퍼스의 변동률을 확인했다. 보다 시각적으로 변화를 확인하기 위해 변동율이 음수로 나타나는 때가 10년 단위 중 어느 때인지를 세어서 10년 단위 별 음수 변동율의 비율을 파이 차트로 그려보았다.

```
#독일어 코퍼스의 변동률
germany_changeRate=pd.DataFrame(corpora.German.pct_change()[1:])
germany_changeRate.columns=['ChangeRate']
germany_changeRate['Decade']=[str(i)[:-1]+'0' for i in germany_changeRate.index]

#변동률 감소 빈도를 나타내는 pie 차트 생성
germany_changeRate[germany_changeRate.ChangeRate<0].Decade.value_counts().sort_index().plot.pie()
plt.title('Change Rate < 0')
plt.savefig('E:/서강대학교/6학기/빅데이터학/빅데이터학_3차과제/그래프/changeRate_minus.png')
```

우선 독일어 코퍼스의 변화율을 데이터 프레임으로 생성한 후 Decade열을 추가하여 10년 단위에 맞게 label을 부여했다. 이후 변동율이 음수, 즉 감소하는 시기의 decade를 센 후 그 결과를 파이 차트로 생성했다. 그 결과 앞서 보았던 차트에서 예측한 것과 같이 1930년대의 감소 빈도가 전체 10년 단위들의 감소 빈도의 거의 반을 차지하고 있음을 확인할 수 있다. 이를 통해 1930년대에 사라진 빈도가 비정상적으로 많았음을 시각적으로 명확하게 알 수 있다. 결국 이 나찌집권 시기에 문화 탄압과 검열이 발생했다는 역사적 사실의 증거를 Ngram을 통해 포착할 수 있었다.



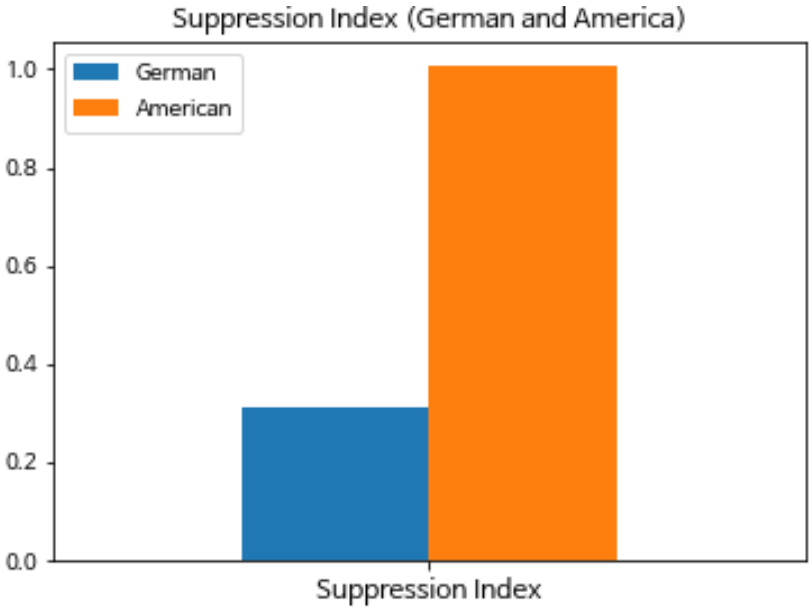
5) 억압 지수

기존의 인문사회과학적 접근에 비해 데이터 및 공학적 접근을 이용하는 것이 유리한 위치를 갖는 이유는 관찰된 패턴을 조작적 정의하여 자동 또는 계량화된 패턴 탐색이 가능하기 때문이다. 이러한 측면에서 나찌의 탄압에 적용해볼 수 있는 조작적 정의는 '억압지수'이다. 억압지수는 특정 기간 전후의 빈도분포 차이의 지수이다. 이러한 억압 지수를 통해 누가 억압을 받았고 누가 인위적 명성을 얻었는지 손쉽게 파악이 가능하며, 계량화하기 어려운 '억압'이라는 개념을 조작적 정의하여 수치화하여 확인할 수 있다. 억압 지수는 (억압된 시기의 평균 빈도) / (억압된 시기 전후 5~10년의 평균 빈도)를 계산하여 구한다. 억압 지수가 1이면 변화가 없는 것, 억압 지수가 1보다 작으면 억압된 것, 억압 지수가 1보다 크면 인위적 명성을 얻은 것으로 간주한다. 따라서 억압 지수가 0에 가까울수로 극심한 억압을 받은 것이고, 억압 지수가 크면 클수록 인위적으로 더 큰 명성을 얻은 것으로 간주할 수 있다.

본 연구에서는 나찌집권 시기의 억압 지수를 구하기 위해 억압된 시기를 1933년부터 1945년으로, 억압 전의 시기를 1923년부터 1933년으로, 억압 후의 시기를 1945년부터 1955년으로 설정했다.

```
#독일 억압지수
german_2333=corpora.German[23:34]
german_3345=corpora.German[33:46]
german_4555=corpora.German[45:56]
german_prepost=german_2333.append(german_4555)
german_sup=german_3345.mean()/german_prepost.mean()

#미국 억압지수
us_2333=corpora.American[23:34]
us_3345=corpora.American[33:46]
us_4555=corpora.American[45:56]
us_prepost=us_2333.append(us_4555)
us_sup=us_3345.mean()/us_prepost.mean()

#억압지수 bar차트 생성
sup=pd.DataFrame({'German':german_sup, 'American':us_sup}, index=['Suppression Index'])
sup.plot.bar()
plt.xticks(rotation=0, fontsize='large')
plt.savefig('E:/서강대학교/6학기/빅데이터학/빅데이터학_3차과제/그래프/g&u_suppression.png')
```

독일과 미국의 억압 지수를 비교하기 위해 우선 각 국가의 억압 지수를 계산했다. 독일어와 미국 영어 코퍼스의 화가 빈도 중앙값을 앞서 설정한 시기에 맞춰 Series로 할당한 후 계산을 했다. 이후 두 수치를 Data Frame으로 병합하여 바 차트를 그려 비교했다.



차트를 보면 미국 영어 코퍼스의 억압 지수는 1에 가깝다. 즉, 미국 영어 코퍼스에서는 전체적으로 6인의 화가들이 억압되지도, 인위적인 명성을 얻지도 않은 것으로 해석할 수 있다. 반면 독일어 코퍼스를 보면 약 0.3으로 매우 0에 가까운 수치이며, 영어의 억압 지수의 1/3 수준이다. 이를 통해 독일어 코퍼스에서는 6인의 화가들이 매우 억압되어왔음을 확인할 수 있다. 이러한 억압 지수를 화가 별로 비교하기 위해 각 화가의 독일어 코퍼스와 미국 영어 코퍼스에서 나타나는 억압 지수를 계산했다.

```
#독일어 코퍼스 화가별 억압지수
picasso_g=gerCo['Pablo Picasso']
g2333_picasso=picasso_g[23:34]
g3345_picasso=picasso_g[33:46]
g4555_picasso=picasso_g[45:56]
gprepost_picasso=g2333_picasso.append(g4555_picasso)
g_picasso_sup=g3345_picasso.mean()/gprepost_picasso.mean()

chagall_g=gerCo['Marc Chagall']
g2333_chagall=chagall_g[23:34]
g3345_chagall=chagall_g[33:46]
g4555_chagall=chagall_g[45:56]
gprepost_chagall=g2333_chagall.append(g4555_chagall)
g_chagall_sup=g3345_chagall.mean()/gprepost_chagall.mean()

kandinsky_g=gerCo['Wassily Kandinsky']
g2333_kandinsky=kandinsky_g[23:34]
g3345_kandinsky=kandinsky_g[33:46]
g4555_kandinsky=kandinsky_g[45:56]
gprepost_kandinsky=g2333_kandinsky.append(g4555_kandinsky)
g_kandinsky_sup=g3345_kandinsky.mean()/gprepost_kandinsky.mean()

matisse_g=gerCo['Henri Matisse']
g2333_matisse=matisse_g[23:34]
g3345_matisse=matisse_g[33:46]
g4555_matisse=matisse_g[45:56]
gprepost_matisse=g2333_matisse.append(g4555_matisse)
g_matisse_sup=g3345_matisse.mean()/gprepost_matisse.mean()

gauguin_g=gerCo['Paul Gauguin']
g2333_gauguin=gauguin_g[23:34]
g3345_gauguin=gauguin_g[33:46]
g4555_gauguin=gauguin_g[45:56]
gprepost_gauguin=g2333_gauguin.append(g4555_gauguin)
g_gauguin_sup=g3345_gauguin.mean()/gprepost_gauguin.mean()

mondrian_g=gerCo['Piet Mondrian']
g2333_mondrian=mondrian_g[23:34]
g3345_mondrian=mondrian_g[33:46]
g4555_mondrian=mondrian_g[45:56]
gprepost_mondrian=g2333_mondrian.append(g4555_mondrian)
g_mondrian_sup=g3345_mondrian.mean()/gprepost_mondrian.mean()
```

```
#미국 영어 코퍼스 화가별 억압지수
picasso_a=usCo['Pablo Picasso']
a2333_picasso=picasso_a[23:34]
a3345_picasso=picasso_a[33:46]
a4555_picasso=picasso_a[45:56]
aprepost_picasso=a2333_picasso.append(a4555_picasso)
a_picasso_sup=a3345_picasso.mean()/aprepost_picasso.mean()

chagall_a=usCo['Marc Chagall']
a2333_chagall=chagall_a[23:34]
a3345_chagall=chagall_a[33:46]
a4555_chagall=chagall_a[45:56]
aprepost_chagall=a2333_chagall.append(a4555_chagall)
a_chagall_sup=a3345_chagall.mean()/aprepost_chagall.mean()

kandinsky_a=usCo['Wassily Kandinsky']
a2333_kandinsky=kandinsky_a[23:34]
a3345_kandinsky=kandinsky_a[33:46]
a4555_kandinsky=kandinsky_a[45:56]
aprepost_kandinsky=a2333_kandinsky.append(a4555_kandinsky)
a_kandinsky_sup=a3345_kandinsky.mean()/aprepost_kandinsky.mean()

matisse_a=usCo['Henri Matisse']
a2333_matisse=matisse_a[23:34]
a3345_matisse=matisse_a[33:46]
a4555_matisse=matisse_a[45:56]
aprepost_matisse=a2333_matisse.append(a4555_matisse)
a_matisse_sup=a3345_matisse.mean()/aprepost_matisse.mean()

gauguin_a=usCo['Paul Gauguin']
a2333_gauguin=gauguin_a[23:34]
a3345_gauguin=gauguin_a[33:46]
a4555_gauguin=gauguin_a[45:56]
aprepost_gauguin=a2333_gauguin.append(a4555_gauguin)
a_gauguin_sup=a3345_gauguin.mean()/aprepost_gauguin.mean()

mondrian_a=usCo['Piet Mondrian']
a2333_mondrian=mondrian_a[23:34]
a3345_mondrian=mondrian_a[33:46]
a4555_mondrian=mondrian_a[45:56]
aprepost_mondrian=a2333_mondrian.append(a4555_mondrian)
a_mondrian_sup=a3345_mondrian.mean()/aprepost_mondrian.mean()
```

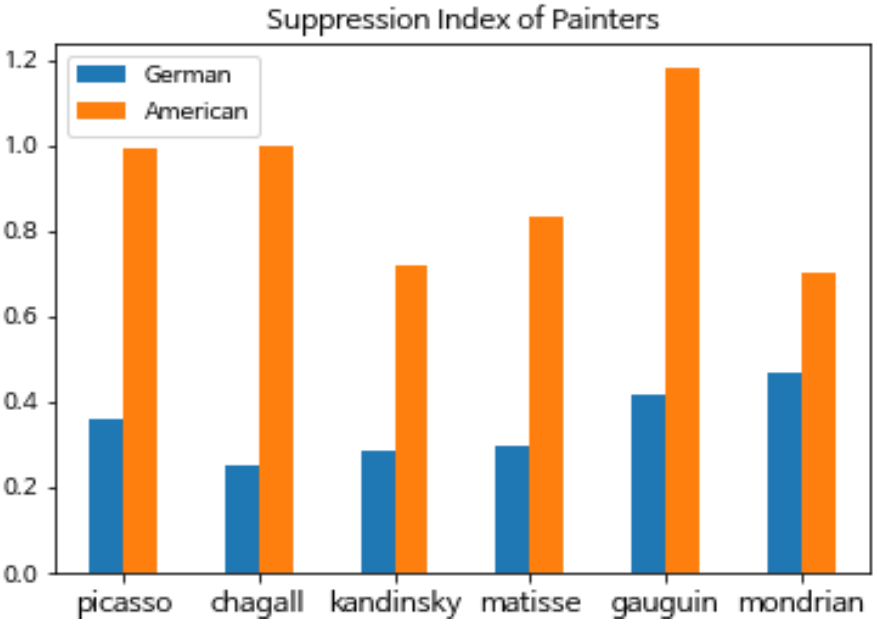
독일어 코퍼스의 화가 별 억압지수와 미국 영어 코퍼스의 화가 별 억압지수를 계산했다.

```
#독일어 코퍼스 화가별 억압지수 Series 생성
gpainters_sup=pd.Series({'picasso':g_picasso_sup,
                         'chagall':g_chagall_sup,
                         'kandinsky':g_kandinsky_sup,
                         'matisse':g_matisse_sup,
                         'gauguin':g_gauguin_sup,
                         'mondrian':g_mondrian_sup})
```

```
#미국 영어 코퍼스 화가별 억압지수 Series 생성
apainters_sup=pd.Series({'picasso':a_picasso_sup,
                         'chagall':a_chagall_sup,
                         'kandinsky':a_kandinsky_sup,
                         'matisse':a_matisse_sup,
                         'gauguin':a_gauguin_sup,
                         'mondrian':a_mondrian_sup})
```

```
#독일어와 미국 영어 코퍼스의 화가별 억압지수 dataFrame 생성
painters_sup=pd.DataFrame(gpainters_sup, columns=['German'])
painters_sup['American']=apainters_sup

#차트 생성
a.plot.bar()
plt.xticks(rotation=0, fontsize='large')
plt.title('Suppression Index of Painters')
plt.savefig('E:/서강대학교/6학기/빅데이터학/빅데이터학_3차과제/그래프/suppression.png')
```

이후 독일어 코퍼스의 화가 별 억압 지수는 gpainters_sup으로, 미국 영어 코퍼스의 화가 별 억압 지수는 apainters_sup으로 할당해 주었다. 이렇게 만들어진 언어별 화가들의 억압 지수 Series를 병합해 하나의 Data Frame을 생성했다. 이후 이 수치를 bar 차트로 만들어서 국가 별로 화가들의 억압 지수가 어떻게 다른지 확인했다.



차트를 보면 거의 대부분의 화가들의 독일어 코퍼스에서의 억압 지수가 미국 영어 코퍼스에서의 억압 지수의 절반이 채 되지 않는다. 또한 미국 영어 코퍼스에서의 억압 지수는 대부분 1에 가까이 있는 반면 독일어 코퍼스에서의 억압 지수는 대부분 0.5 아래에 위치해있다. 이것은 앞서 확인한 국가별 억압 지수와도 동일한 추세이다. 이를 통해 독일어 코퍼스에서는 극심한 억압이 있었다는 것을 알 수 있는 것은 물론, 중앙값이 각 화가들의 척도를 잘 대변했음 또한 알 수 있다.

3. 인문사회과학적 의미

지금까지 본 다양한 차트와 통계 분석을 통해 알 수 있는 추세 패턴의 인문사회과학적 의미는 몇 가지가 있다. 우선 그 무엇보다도 중요한 것은 나찌집권 시기에 '퇴폐적', '유대인적', '비독일적'인 것으로 여겨지던 작품들과 그것을 만들어낸 화가들에 대한 억압이 있었으며, 이러한 억압을 수치를 통해 증명해낼 수 있다는 것이다. 또 데이터를 통한 확인한 차트들은 과거 인문사회과학적 관점으로는 접근할 수 없는 계량화된 차원의 증거이다. 동시에 나찌집권에 대한 인문사회과학적 지식 없이는 이런 차트들을 확인하기도 어려울 뿐더러 확인했다고 하더라도 무의미한 것으로 치부될 수 있다. 결국 이러한 데이터와 인문사회과학은 불가분의 관계인 것이다.

이러한 데이터와 추세 패턴이 갖는 인문사회과학적 의미, 그리고 불가분의 관계는 비단 억압과 검열에만 국한되는 것이 아니다. 이들의 관계는 인문사회과학, 그리고 데이터와 추세 패턴 전반에서 크게 세 가지 의미를 갖는다. 우선 첫째, 인문사회과학적 배경이 없다면 데이터를 보아도 유의미한 결과를 도출할 수 없으며, 더 근본적으로 유의미한 연구를 시작할 수조차 없다. 데이터와 추세 패턴을 보면서 대부분의 사람들이 떠올리는 것들만 발견해낸다면 이는 유의미한 연구라고 할 수 없다. 많은 사람들이 알아채지 못한 부분을 데이터에서 발견해 냈을 때 그것이 비로소 유의미한 연구가 되고, 그러한 연구를 통해 유의미한 결과를 도출해 낼 수 있는 것이다. 그러나 인문사회과학적 배경이 없다면 이렇게 섬세한 관찰을 하기는 것은 어려우며 만약 섬세한 관찰을 했다고 하더라도 그 속에 있는 진정한 의미를 도출해내기 어렵다.

위와 비슷한 맥락에서 둘째, 인문사회과학은 데이터에 유용성을 부여한다. 앞서 화가들의 독일어 코퍼스에서 발견한 특정 시기의 감소 패턴은 '나찌 정권의 탄압과 검열'이라는 인문사회과학적 배경지식이 없다면 그저 낮은 빈도로 치부되어 크게 필요 없는 데이터가 되었을 것이다. 이처럼 별 볼 일 없어 보이는 데이터에도 인문사회과학적 통찰이 가미된다면 비로소 유용한 정보로써 가치를 가질 수 있다.

마지막으로 인문사회과학을 통해 데이터의 잘못된 해석을 피하고 올바른 해석으로 나아갈 수 있다. 앞서 연구한 나찌시대의 독일어 코퍼스 빈도 감소는 이에 대한 예시 중 하나이다. 인문사회과학적 배경지식이 없었다면 1930년대의 급감한 빈도는 단순히 그 화가들에 대한 관심이 줄어들었다고 밖에 해석할 수 없었을 것이다. 그러나 이 시기에 동시에 개최된 두 개의 미술전을 보면 전혀 그렇지 않다는 것을 알 수 있다. 나치가 승인한 예술가의 작품을 전시한 '위대한 독일 미술전'과 탄압받던 화가들의 작품을 전시한

‘퇴폐 미술전’이 그것들이다. 이름에서부터 할 수 있듯이 퇴폐 미술전은 현대 미술의 붕괴를 목적으로 작품들을 모욕적인 슬로건과 함께 볼품없이 전시했으며, 그 규모도 ‘위대한 독일 미술전’과는 비교할 수 없이 작았다. 그러나 ‘퇴폐 미술전’에는 하루 평 1만 7천여 명씩 넉 달 동안 200만명의 관람객이 모이며 ‘위대한 독일 미술전’ 관람객 수의 5배를 기록했다. 이 미술전을 통해 알 수 있듯이 이 시기에 급감한 빈도는 절대 사회적 관심의 부재로 해석될 수 없으며 오히려 사회적 관심을 억누르려는 탄압과 검열의 증거였다. 이러한 예시를 통해 알 수 있듯이 인문사회과학적 배경지식이 있다면 위와 같은 잘못된 해석을 피해 올바른 해석을 도출할 수 있으며 나아가 인문사회과학적 통찰을 통해 더욱 깊고 세밀한 의미 부여가 가능하다.

데이터를 구성하는 대부분의 요소들은 인간 사회와 관련이 있다. 결국 데이터를 다루는 데 있어 무엇보다 중요한 것은 인문사회과학적 배경지식을 기반으로 인간에 대한 관심에서 비롯되는 새롭고 창의적인 질문인 것이다. 이러한 사실을 인지하며, 인문사회과학이라는 정신과 데이터라는 도구가 함께 한다면, 같은 자료를 보더라도 번뜩이는 통찰을 이끌어낼 수 있을 것이다.