출판사별 발행 도서의 발행 년도와 대출건수 알아보기

시스템제어공학과
20210710 맹지우

Content

Introduction

서론_배경 및 목적

Main Body

본론_코드 구현

Conclusion

결론_결과 분석

01

서론

Introduction

배경 및 목적

도서관의 도서 데이터에는 도서의 출판사 정보가 포함되어 있으며, 이를 통해 도서관의 출판사 데이터를 분석할 수 있습니다. 출판사별로 도서의 발행 년도, 수량, 대출 건수 등을 분석함으로써 특정 출판사의 활동성과 인기도, 출판 동향을 파악할 수 있으며, 이를 활용한다면 도서관은 출판사와의 협력 관계 강화, 도서 수장 전략 수립, 독자 서비스 개선 등에 활용할 수도 있습니다.

이 프로젝트에서 진행할 것은 도서관에 있는 도서의 출판사별 데이터 분석을 위해 필요한 데이터 전처리, 시각화, 샘플링 등의 작업할 것이며, 코드를 통해 출판사별 도서의 분포, 발행 년도, 대출 건수 등을 분석하고 시각화 하려고 합니다.

Main Body

02

본론

한글 폰트 설정

```
1  # 한글폰트 설정
2  import matplotlib.pyplot as plt
3  import matplotlib.font_manager as fm
4
5  font_location = 'C:₩₩WINDOWS₩₩Fonts₩₩HancomHoonminjeongeumH.ttf' # For Windows
6  font_name = fm.FontProperties(fname=font_location).get_name()
7  plt.rc('font', family=font_name)
```

matplotlib의 pyplot과 matplotlib의 font_manager 폰트 관련 기능을 제공하는 모듈 불러오기

5번째 줄

사용할 한글 폰트 파일의 경로를 지정

6번째 줄

fm.FontProperties()를 사용하여 폰트 파일을 지정하고, get_name() 함수를 호출하여 폰트의 이름 가져오기

7번째 줄

Plt.rc()를 사용하여 전역적으로 폰트 패밀리 설정

위와 같은 코드를 실행하면 matplotlib에서 사용하는 폰트 패밀리가 설정된 한글 폰트로 변경되어, 이후에 생성하는 그래프나 텍스트 요소들은 설정한 한글 폰트를 사용하여 표시됨

데이터프레임 변환

```
1 import pandas as pd
2
3 ns_book7 = pd.read_csv('ns_book7.csv', low_memory=False)
4 ns_book7.head()
```

	번호	도서명	저자	출판사	발행년도	ISBN	세트 ISBN	부가기호	권	주제분류번호	도서권수	대출건수	등록일자
0	1	인공지능과 흙	김동훈 지음	민음사	2021	9788937444319	NaN	NaN	NaN	NaN	1	0	2021-03-19
1	2	가짜 행복 권하는 사회	김태형 지음	갈매나무	2021	9791190123969	NaN	NaN	NaN	NaN	1	0	2021-03-19
2	3	나도 한 문장 잘 쓰면 바랄 게 없겠네	김선영 지음	블랙피쉬	2021	9788968332982	NaN	NaN	NaN	NaN	1	0	2021-03-19
3	4	예루살렘 해변	이도 게펜 지음, 임재희 옮김	문학세계사	2021	9788970759906	NaN	NaN	NaN	NaN	1	0	2021-03-19
4	5	김성곤의 중국한시기행 : 장강·황하 편	김성곤 지음	김영사	2021	9788934990833	NaN	NaN	NaN	NaN	1	0	2021-03-19

pandas 라이브러리 가져오기

pd.read_csv() 함수를 사용하여 'ns_book7.csv' 파일을 읽어와 데이터프레임으로 변환
low_memory=False는 메모리 사용을 최적화하는 옵션으로, 큰 파일을 읽을 때 사용함. 파일을 읽어와서 ns_book7 변수에 할당
head() 함수를 사용하여 데이터프레임의 상위 5개 행 출력

이 코드를 실행하면 ' ns_book7.csv ' 파일을 읽어와서 데이터프레임으로 변환한 후, 상위 5개 행을 출력하여 데이터프레임의 구조와 내용을 간략히 확인할 수 있음

데이터 변수 저장

```
1 top30_pubs = ns_book7['출판사'].value_counts()[:30]
2 top30_pubs
```

```
문학동네              4410
민음사               3349
김영사               3246
웅진씽크빅             3227
시공사               2685
창비                2469
문학과지성사            2064
위즈덤하우스            1981
학지사               1877
한울                1553
한국학술정보            1496
열린책들              1491
살림출판사             1479
한길사               1460
博英社               1458
커뮤니케이션북스          1445
지식을만드는지식          1390
자음과모음             1364
비룡소               1331
랜덤하우스코리아          1314
넥서스               1310
황금가지              1101
길벗                1094
시그마프레스            1063
현암사               1054
다산북스              1046
집문당               1038
책세상               1037
한국문화사             1028
북이십일 21세기북스      1026
Name: 출판사, dtype: int64
```

top30_pubs = ns_book7['출판사'].value_counts()[:30]

ns_book7 데이터프레임의 "출판사" 열에 대해 value_counts() 함수를 호출하여 각 출판사의 등장 횟수를 세고, 내림차순으로 정렬하여 상위 30개의 값을 top30_pubs 변수에 저장

```
1 top30_pubs_idx = ns_book7['출판사'].isin(top30_pubs.index)
2 top30_pubs_idx
```

```
0          True
1         False
2         False
3         False
4          True
          ...
376765    False
376766    False
376767     True
376768    False
376769    False
Name: 출판사, Length: 376770, dtype: bool
```

top30_pubs_idx = ns_book7['출판사'].isin(top30_pubs.index)

ns_book7 데이터프레임의 "출판사" 열에서 각 행의 값이 top30_pubs.index에 포함되는지 여부를 검사하여 불리언 값을 생성하여서 상위 30개 출판사에 해당하는 행은 True로, 그렇지 않은 행은 False로 채워 짐

데이터 변수 저장

```
1 ns_book8 = ns_book7[top30_pubs_idx].sample(1000, random_state=42)
2 ns_book8.head()
```

	번호	도서명	저자	출판사	발행년도	ISBN	세트 ISBN	부가기호	권	주제분류번호	도서권수	대출건수	등록일자
141760	155786	제갈량 문집	제갈량 지음 ;조영래 옮김	지식을만드는지식	2012	9788966805785	NaN	0	10	808	1	2	2013-04-10
249855	268595	존 레넌을 찾아서	토니 파슨스 지음;이은정 옮김	시공사	2007	9788952750419	NaN	0	NaN	843	1	18	2007-12-14
129347	142802	요리사 & 쇼핑호스트 :생활과학 계열·예체능 계열	와이즈멘토 글 ; 김성희 그림	김영사	2013	9788934959854	9788934959717	7	14	321.55	1	3	2013-12-09
349194	371975	임정섭의 글쓰기 훈련소	임정섭 지음	다산북스	2017	9791130614472	NaN	NaN	NaN	NaN	1	0	1970-01-01
46734	51748	초한지 :이문열의 史記 이야기	지은이: 이문열	민음사	2017	9788937481659	9788937481581	0	7	813.6	1	9	2018-07-02

ns_book7[top30_pubs_idx]

앞에서 만든 불리언 값이 저장된 데이터 변수 top30_pubs_idx를 사용하여 ns_book7 데이터프레임에서 상위 30개 출판사에 해당하는 행들을 선택

.sample(1000, random_state=42)

선택한 행들 중에서 무작위로 1000개의 행을 샘플링. sample() 함수의 첫 번째 매개변수는 샘플링 할 개수를 나타내고, random_state 매개변수는 난수 발생 시드를 설정하여 재현 가능한 샘플링 결과를 얻는데, 여기서는 42로 하여 항상 동일한 샘플링 결과를 얻을 수 있음

도서관에 있는 도서의 발행 년도를 출판사별로 나타냄

```
1 fig, ax = plt.subplots(figsize=(10, 8))
2 ax.scatter(ns_book8['발행년도'], ns_book8['출판사'])
3 ax.set_title('출판사별 발행도서')
4 fig.show()
```

fig, ax = plt.subplots(figsize=(10, 8))

plt.subplots() 함수를 사용하여 객체 생성

fig는 전체 그림, ax는 그림 내에서 각각의 축 의미

figsize=(10, 8) 그림의 크기를 가로 10, 세로 8인치로 설정

ax.scatter(ns_book8['발행년도'], ns_book8['출판사'])

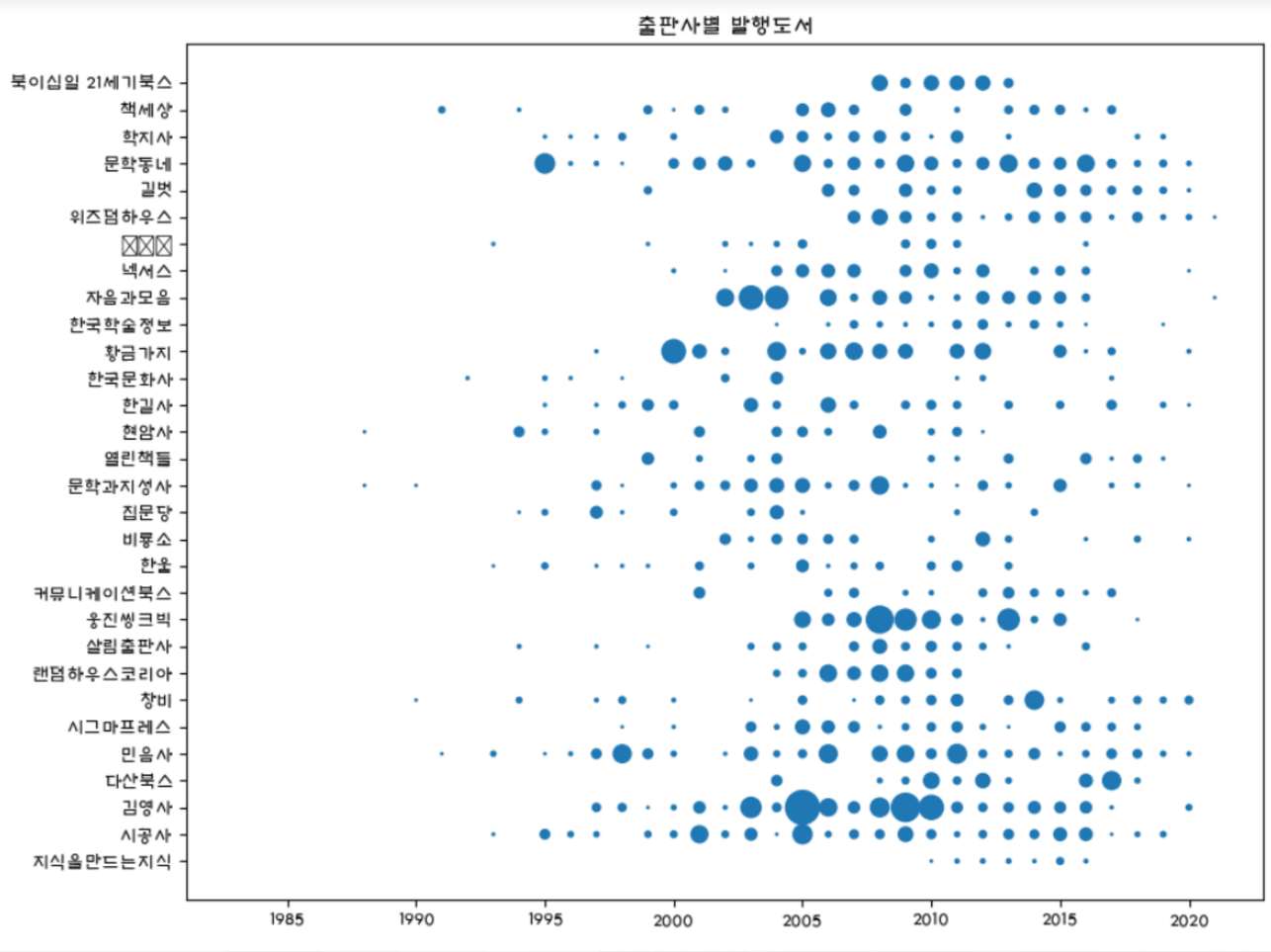
ax.scatter() 산점도 그려주는 함수

ns_book8['발행년도']는 x축 데이터

ns_book8[' 출판사 ']는 y축 데이터

도서관에 있는 도서의 발행 년도를 출판사별로 나타냄 + 대출건수 표현

```
1  fig, ax = plt.subplots(figsize=(10, 8))
2  ax.scatter(ns_book8['발행년도'], ns_book8['출판사'], s=ns_book8['대출건수'])
3  ax.set_title('출판사별 발행도서')
4  fig.show()
```



fig, ax = plt.subplots(figsize=(10, 8))

plt.subplots() 함수를 사용하여 객체 생성

fig는 전체 그림, ax는 그림 내에서 각각의 축 의미

figsize=(10, 8) 그림의 크기를 가로 10, 세로 8인치로 설정

ax.scatter(ns_book8['발행년도'], ns_book8['출판사'],
s=ns_book8['대출건수'])

ax.scatter() 산점도 그려주는 함수

ns_book8['발행년도']는 x축 데이터

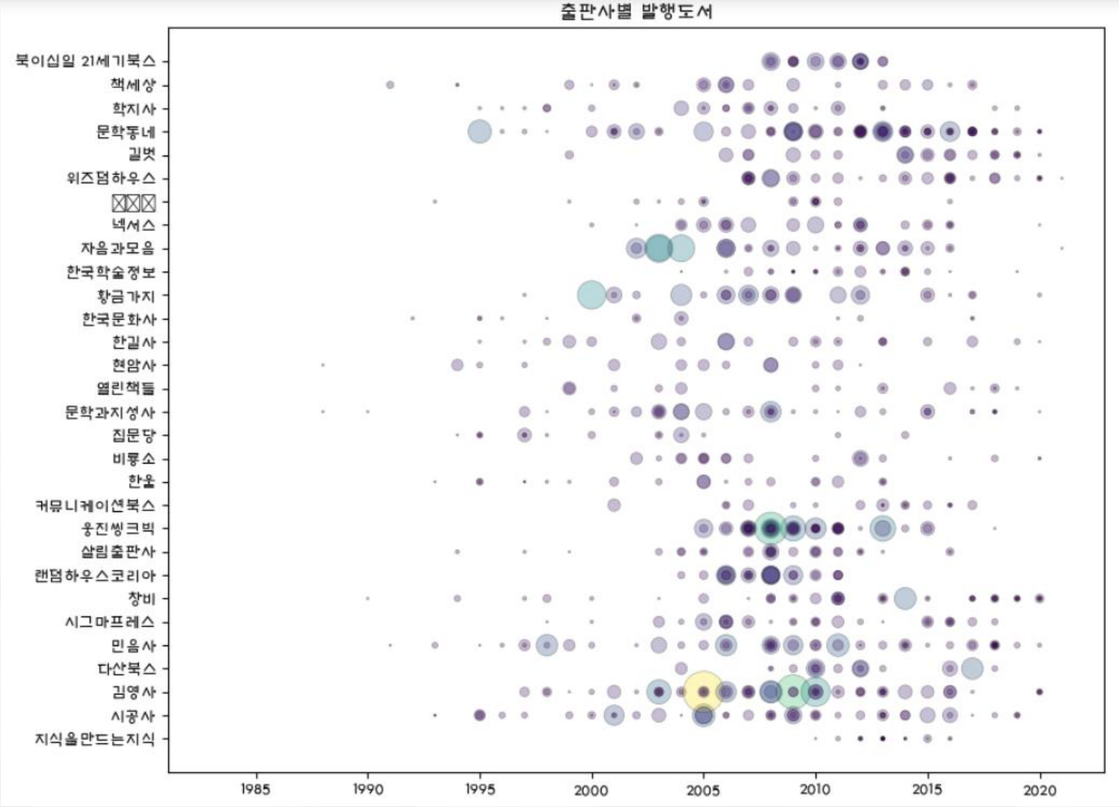
ns_book8[' 출판사 ']는 y축 데이터

ns_book8['대출건수']는 각 점의 크기를 나타내는 데이터로서,

대출건수가 많은 도서일수록 점의 크기가 크게 표현됨

도서관에 있는 도서의 발행 년도를 출판사별로 나타냄 + 대출건수 표현(색상 명도 표현)

```
fig, ax = plt.subplots(figsize=(10, 8))
ax.scatter(ns_book8['발행년도'], ns_book8['출판사'],
           linewidths=0.5, edgecolors='k', alpha=0.3,
           s=ns_book8['대출건수']*2, c=ns_book8['대출건수'])
ax.set_title('출판사별 발행도서')
fig.show()
```



ax.scatter(ns_book8['발행년도'], ns_book8['출판사'],
linewidths=0.5, edgecolors='k', alpha=0.3,
s=ns_book8['대출건수']*2, c=ns_book8['대출건수'])

ns_book8['발행년도']는 x축 데이터

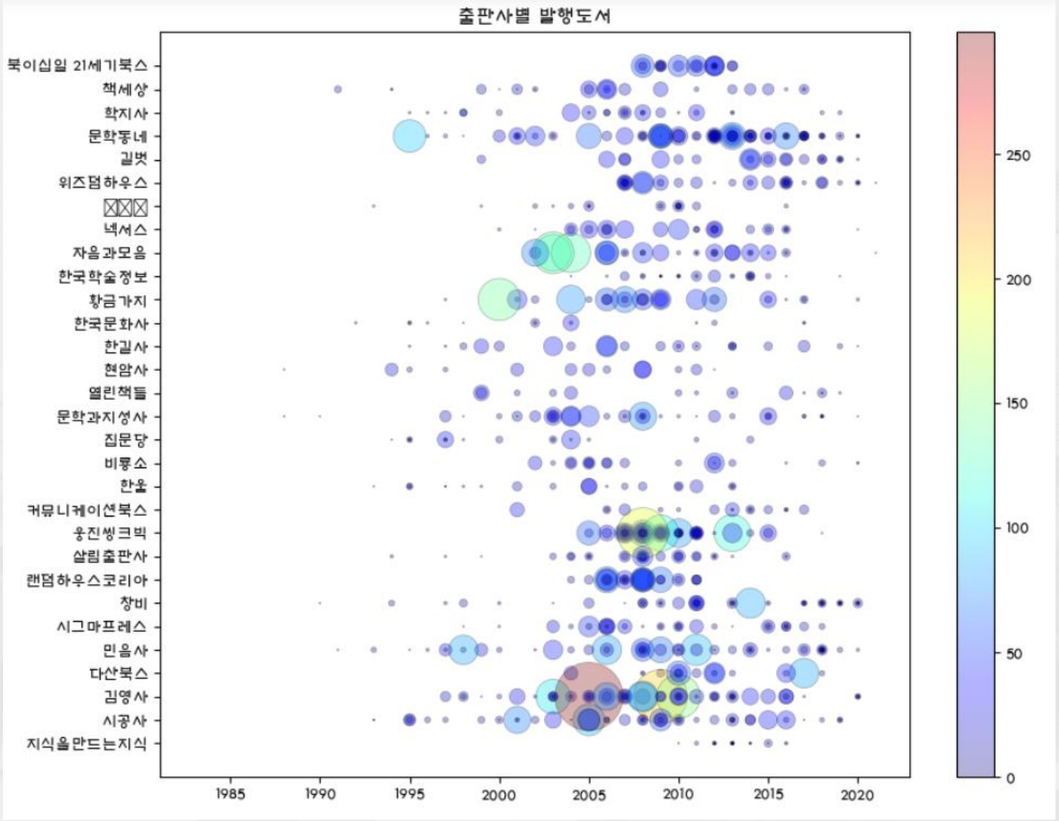
ns_book8['출판사']는 y축 데이터

ns_book8['대출건수']는 각 점의 크기를 나타내는 데이터로서,
대출건수가 많은 도서일수록 점의 크기가 크게 표현됨

c=ns_book8['대출건수']는 점의 색상을 대출건수에 따라 설정하여,
대출건수가 많을수록 색상이 진하게 표현됨

도서관에 있는 도서의 발행 년도를 출판사별로 나타냄 + 대출건수 표현(색상 채도 표현)

```
fig, ax = plt.subplots(figsize=(10, 8))
sc = ax.scatter(ns_book8['발행년도'], ns_book8['출판사'],
                linewidths=0.5, edgecolors='k', alpha=0.3,
                s=ns_book8['대출건수']**1.3, c=ns_book8['대출건수'], cmap='jet')
ax.set_title('출판사별 발행도서')
fig.colorbar(sc)
fig.show()
```



s=ns_book8['대출건수']1.3**

점의 크기를 대출건수의 1.3 제곱으로 설정하여, 대출건수가 많을수록 점의 크기가 더 크게 표현됨

c=ns_book8['대출건수']

점의 색상을 대출건수에 따라 설정

cmap='jet'

색상 맵을 'jet'으로 설정하였으며, 'jet' 색상 맵은 대출건수가 낮을수록 파란색, 높을수록 빨간색으로 표현됨

fig.colorbar(sc)

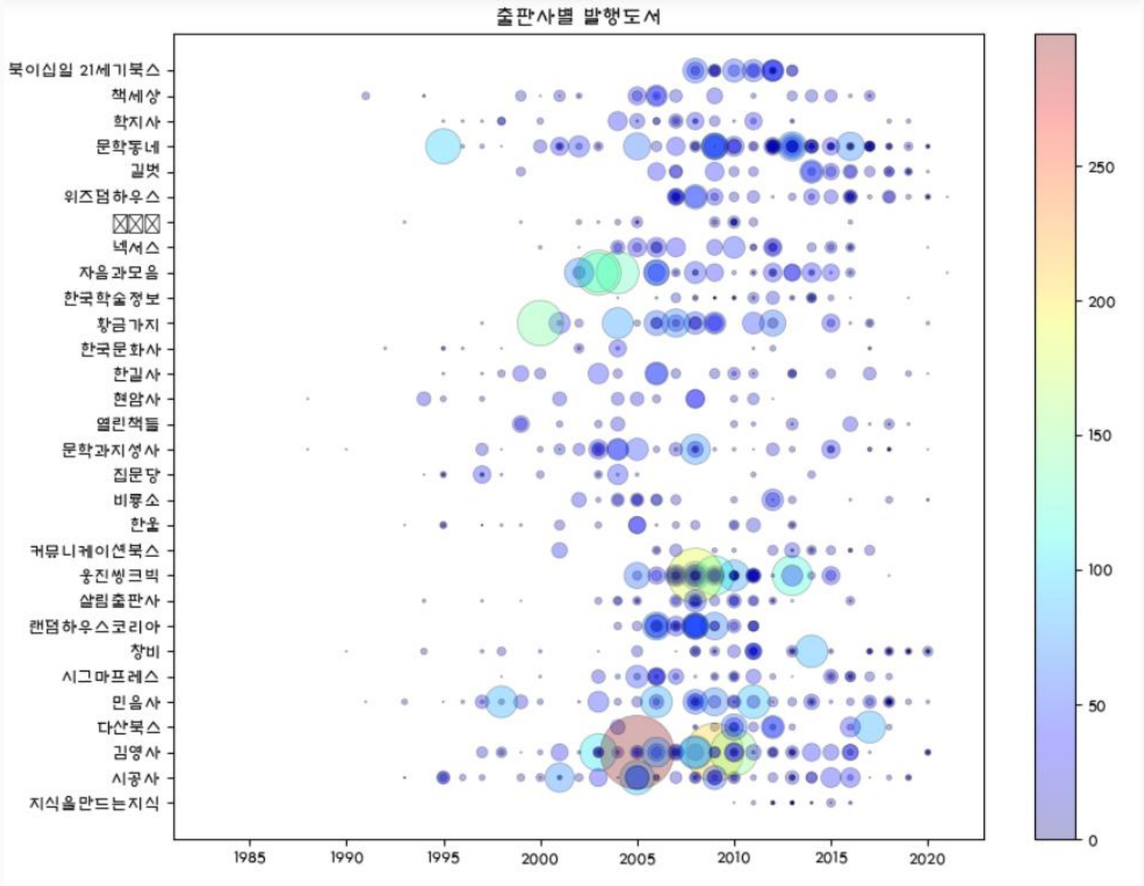
색상 막대기(colorbar)를 추가하여, 산점도 점의 색상과 대출건수 간의 대응 관계를 시각적으로 표현

03

결론

Conclusion

결론



대출건수가 가장 많은 년도를 보면 2005년, 그 다음으로는 2007년에서 2010년 사이인 것을 볼 수 있습니다.

출판사별로 대출건수를 살펴보면 다산북스, 김영사, 시공사 세개의 출판사 도서를 대체적으로 많이 대출한다는 것을 알 수 있습니다.

THANK YOU