Proj 1 - MySQL & Knowledge Discovery

Jiaxin Wu

MAT 259A

Winter 2022

Introduction

There are various fields in Computer Science to study. As a student majoring in Computer Science, I want to know whether the popularity of different fields has an influence on library checkouts. So I write SQL about the checkouts related to fields within the Dewey Class 100. Also, as a programmer, I am interested in analyzing the trend of different programming languages from this data. After that, I compared my results with the statistics collected by the authority to make a deeper explanation.

SQL Query 1

```
SELECT
    YEAR(cout) AS years,
    COUNT(IF(deweyClass >= 000 AND deweyClass < 001, 1, NULL)) AS 'Computer science, information and general works',
    COUNT(IF(deweyClass >= 001 AND deweyClass < 002, 1, NULL)) AS 'Knowledge',
    COUNT(IF(deweyClass >= 003 AND deweyClass < 004, 1, NULL)) AS 'Systems',
    COUNT(IF(deweyClass >= 004 AND deweyClass < 005, 1, NULL)) AS 'Data processing and computer science',
    COUNT(IF(deweyClass >= 005 AND deweyClass < 006, 1, NULL)) AS 'Computer programming, programs and data',
    COUNT(IF(deweyClass >= 006 AND deweyClass < 007, 1, NULL)) AS 'Special computer methods (e.g. AI, multimedia, VR)'
FROM
    spl_2016.outraw
WHERE
    deweyClass < 100 AND YEAR(cout) < 2022
GROUP BY YEAR(cout)
ORDER BY YEAR(cout) DESC
```

I use the query to count the checkout records in different fields of Computer science, information and general works. Then I group the results in the same year together to see the change of popularity over the years.

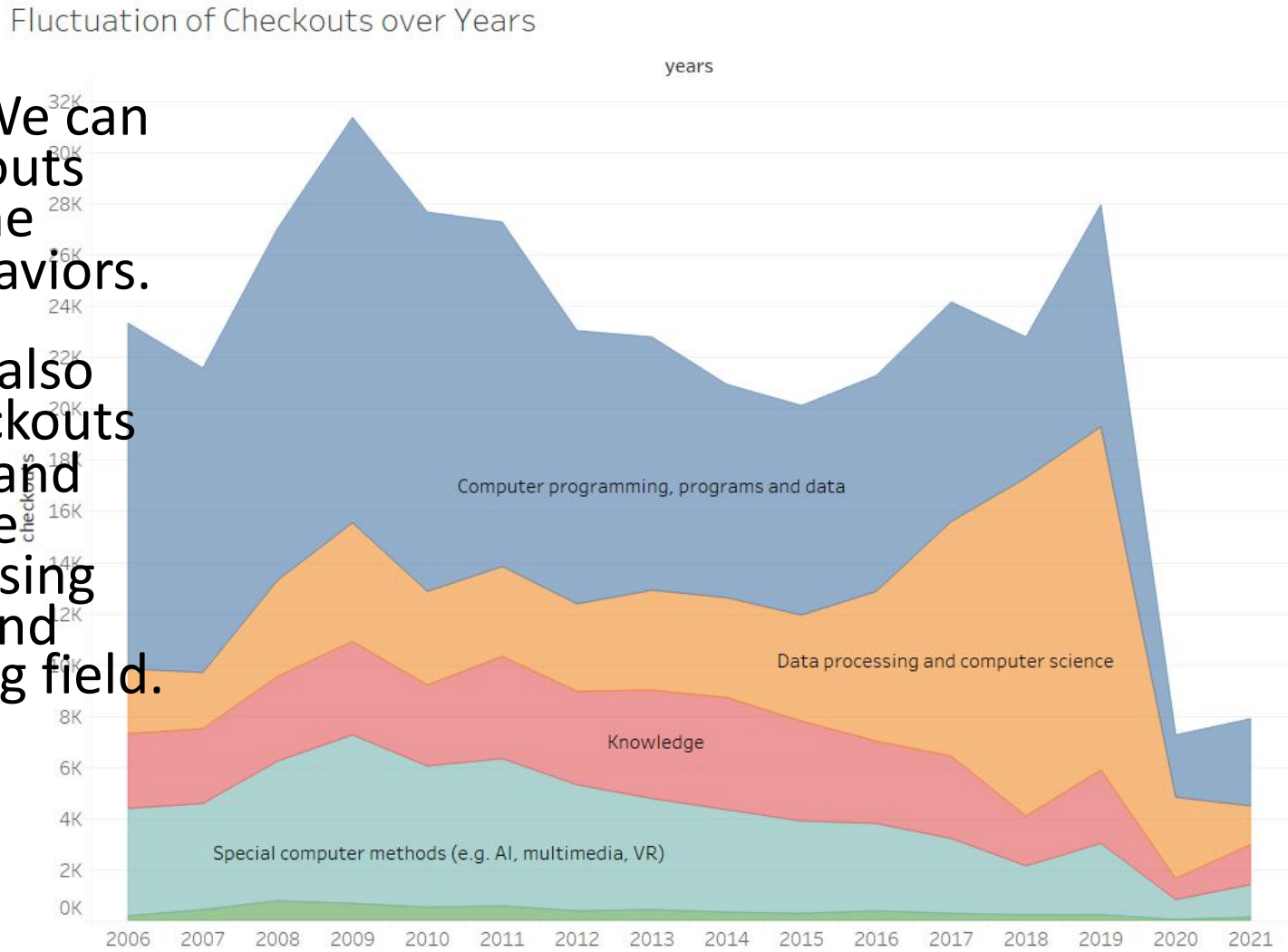
Results 1

After executing the query, I got the results as listed in the table. We can find that the count of the "Computer science, information and general works" category is very high. The reason is that it is a very general category. So a large number of books belong to it.

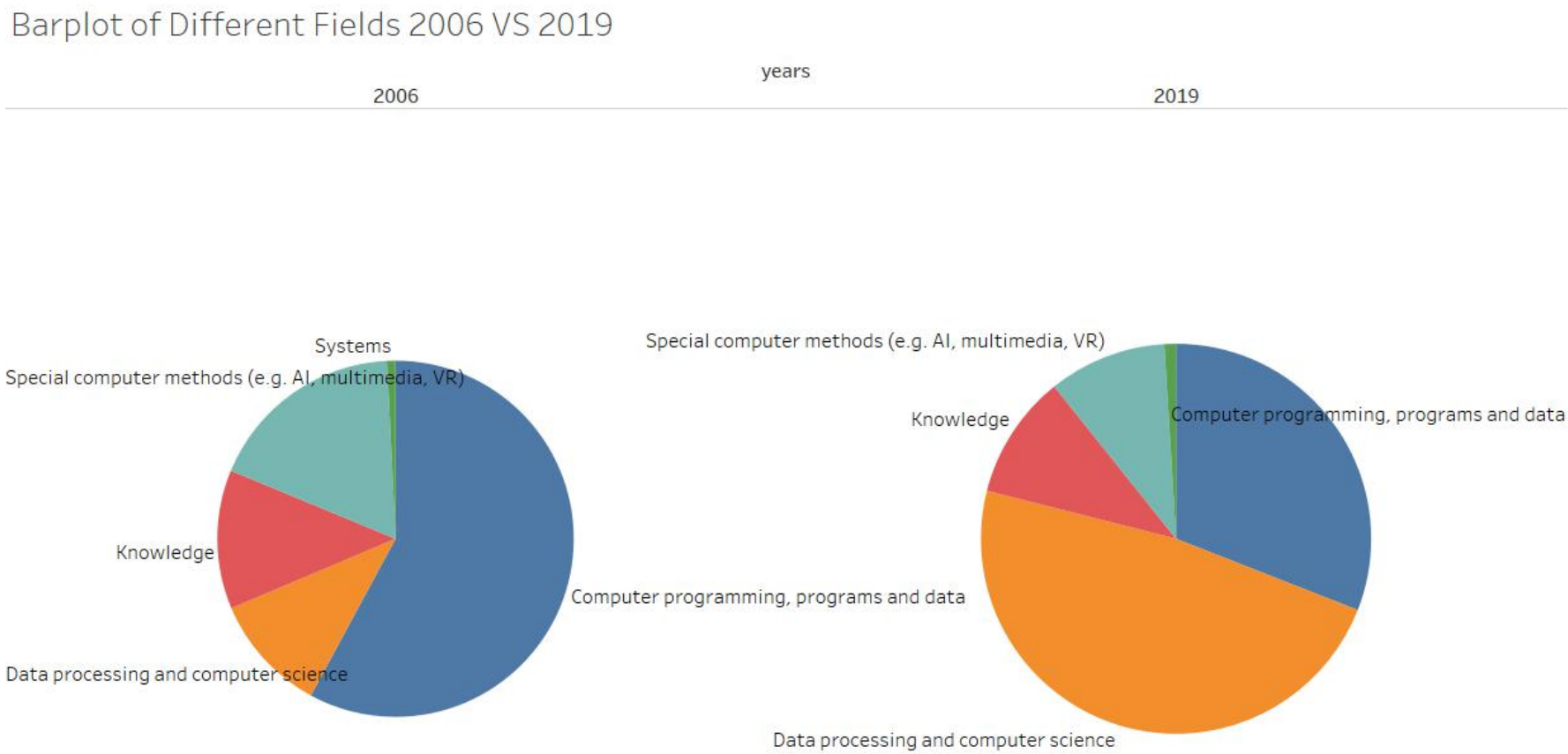
years	Computer science, information and general works	Knowledge	Systems	Data processing and computer science	Computer programming, programs and data	Special computer methods (e.g. AI, multimedia, VR)
2006	3424214	2945	176	2506	13497	4193
2007	3362577	2931	435	2194	11894	4138
2008	4713593	3315	767	3724	13772	5475
2009	5141529	3674	662	4602	15839	6578
2010	4979017	3158	556	3659	14802	5484
2011	4655379	4030	585	3490	13458	5732
2012	4361197	3653	413	3431	10645	4879
2013	4750906	4266	421	3852	9911	4347
2014	4515221	4384	351	3937	8327	3962
2015	4298895	3872	309	4148	8155	3610
2016	4007669	3215	380	5867	8421	3407
2017	3648182	3202	303	9167	8565	2912
2018	2576158	1981	245	13188	5503	1878
2019	3694800	2882	265	13419	8653	2734
2020	1037786	813	65	3194	2423	756
2021	1939875	1554	142	1526	3395	1256

When doing visualization tasks, we delete the data from the “Computer science, information and general works” category to get better results. The visualization results are shown below.

The first figure shows the fluctuation of checkouts in different fields over years. We can observe the change in the ratio of checkouts from different fields. We can even find the influence of COVID-19 on people’s behaviors. The total amount of checkouts dropped dramatically from 2019. Besides, we can also find that as the most basic field, the checkouts of “Computer programming, programs and data” dropped slowly over years. On the contrary, the checkouts of “Data processing and computer science” become more and more over the years since it is a promising field.



The finds mentioned above can also be illustrated by this barplot, which shows the difference of ratio in 2006 and 2019. (Here we exclude 2020 and 2021 due to COVID-19.)



SQL Query 2

In this query, I want to calculate the popularity of different programming languages. So I use the data in the field of “Computer programming, programs and data”. By comparing the title of the books with different keywords, I calculate their trends over years.

```
SELECT
    YEAR(cout) AS years,
    COUNT(CASE
        WHEN (LOWER(title) LIKE '% c %') THEN
        1
    END) AS 'C/C++',
    COUNT(CASE
        WHEN (LOWER(title) LIKE '% sql %') THEN
        1
    END) AS 'SQL',
    COUNT(CASE
        WHEN (LOWER(title) LIKE '%python%') THEN
        1
    END) AS Python,
    COUNT(CASE
        WHEN (LOWER(title) LIKE '%java %') THEN
        1
    END) AS Java,
    COUNT(CASE
        WHEN (LOWER(title) LIKE '%javascript%') THEN
        1
    END) AS JavaScript,
    COUNT(CASE
        WHEN (LOWER(title) LIKE '%php%') THEN
        1
    END) AS PHP,
    COUNT(CASE
        WHEN (LOWER(title) LIKE '%visual basic%' or LOWER(title) LIKE '% vb %') THEN
        1
    END) AS VB,
    COUNT(CASE
        WHEN (LOWER(title) LIKE '%assembly%') THEN
        1
    END) AS Assembly
FROM
    spl_2016.outraw
WHERE
    YEAR(cout) < 2022 AND deweyClass >= 005
        AND deweyClass < 006
GROUP BY YEAR(cout)
```

Results 2

The data collected is showed in this table. Though not evident, we can still find that the record of 2020 and 2021 is influenced by COVID-19 and less people borrow books at that time.

years	C/C++	SQL	Python	Java	JavaScript	PHP	VB	Assembly
2006	289	196	56	372	211	306	281	12
2007	196	169	37	233	177	338	212	11
2008	266	205	101	244	201	398	195	13
2009	431	154	272	229	323	523	149	14
2010	321	103	227	135	360	404	81	4
2011	296	103	246	205	408	306	50	2
2012	161	59	210	201	551	164	19	3
2013	107	98	233	230	470	102	14	1
2014	74	84	257	217	487	70	12	0
2015	78	80	456	277	615	84	10	0
2016	75	108	694	247	533	80	8	0
2017	61	105	876	222	509	57	1	0
2018	39	85	704	163	268	41	0	0
2019	60	155	1099	223	358	39	0	2
2020	22	44	342	59	71	7	0	1
2021	21	53	506	77	123	16	0	0

We can plot the data as follows and compare it with the trend of different programming languages collected by the authority. We can find that fewer and fewer people use VB and Assembly nowadays, so their checkouts drop gradually. As a promising language, Python is used by more and more people in various fields. As a result, the checkouts related to Python experienced a dramatic lift in the last ten years. As to programming languages like Java, JavaScript, and C/C++, their popularity remain similar over the years.