CSCI 5408 Data Management, Warehousing, and Analytics

Assignment – 3

Banner ID: - B00843607 Darpan Patel

A. Sentiment Analysis

Total tweets analyzed **3571**.

Search keywords: ["Dalhousie University", "Canada", "University", "Halifax", "Canada Education"]

To clean data I have removed special characters, links, emoticons and converted everything to lower case using regex. After cleaning check for empty values in tweets and news, if present remove that row.

Codes for cleaning were pre-written in assignment 2 so it was reused.

List of positive words is in "positive-words.txt" & list of negative words are in "negative-words.txt" [1]

All positive words extracted from tweets are in "PositiveWords.csv" & all negative words extracted from tweets are in "NegativeWords.csv."

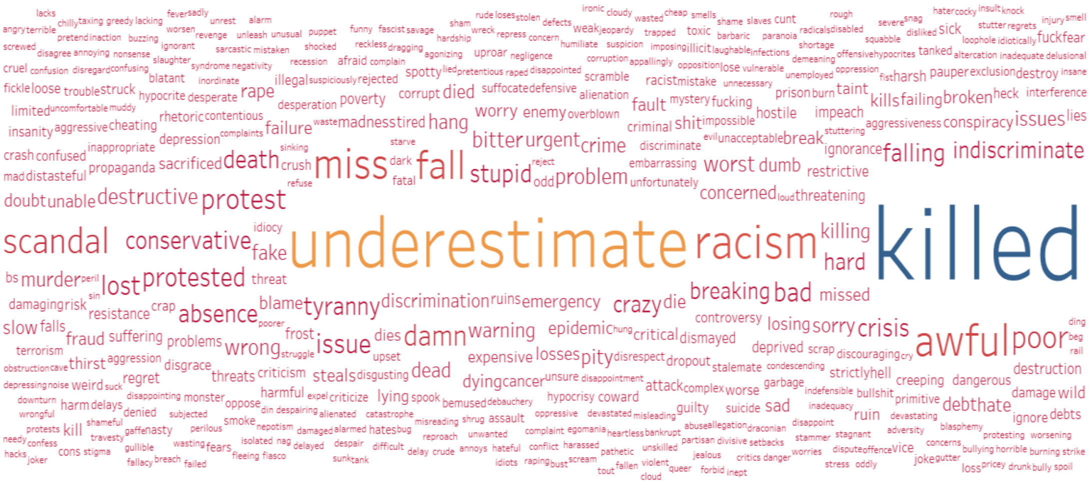


Figure 1: negative_words usage in tweets

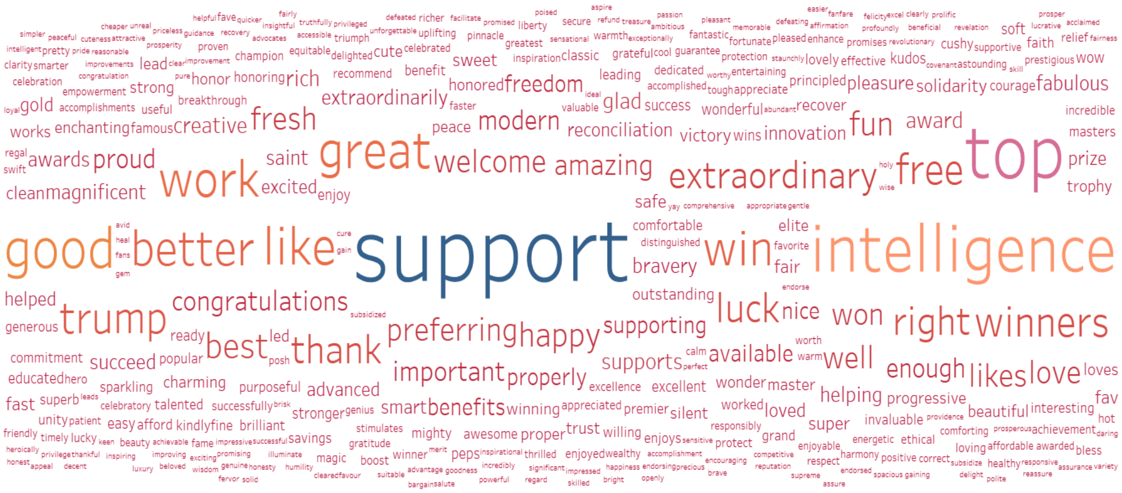
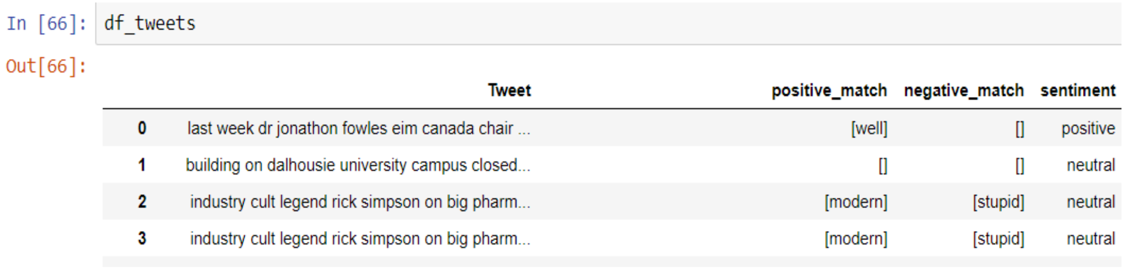


Figure 2: positive_words usage in tweets



```
In [66]: df_tweets
```

Out[66]:

	Tweet	positive_match	negative_match	sentiment
0	last week dr jonathon fowles eim canada chair ...	[well]	[]	positive
1	building on dalhousie university campus closed...	[]	[]	neutral
2	industry cult legend rick simpson on big pharm...	[modern]	[stupid]	neutral
3	industry cult legend rick simpson on big pharm...	[modern]	[stupid]	neutral

Figure 3: sentiment analysis

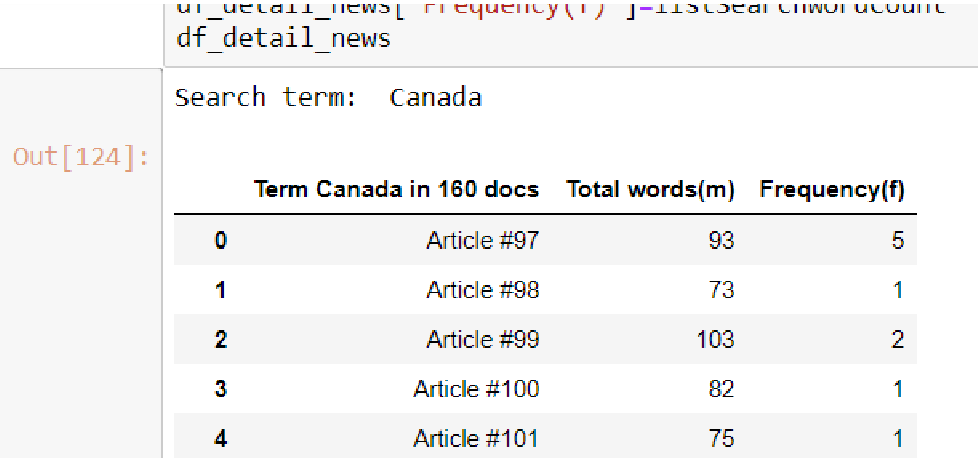
B. Semantic Analysis

Total news analyzed **481**.

from='2019-10-05' to='2019-11-02'

Search keywords: ["Dalhousie University", "Canada", "University", "Halifax", "Canada Education"].

Codes for cleaning were pre-written in assignment 2 so clean .csv for news were used.



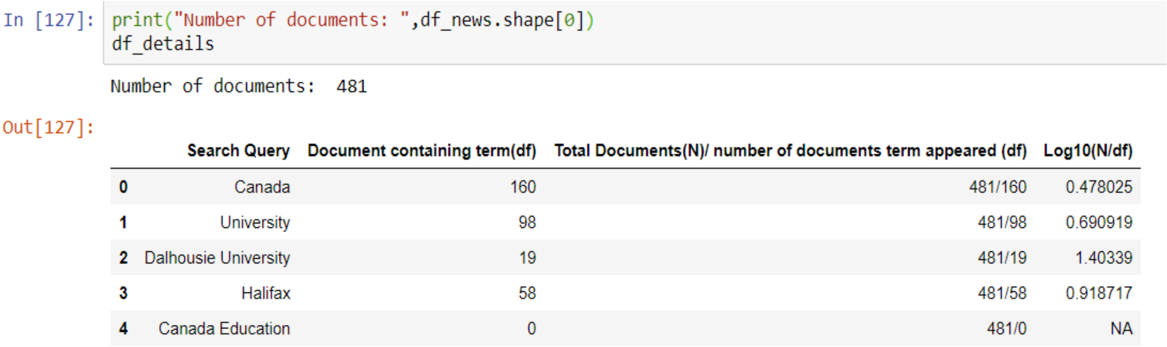
```
df_detail_news['Frequency(f)']=listsearchwordcount
df_detail_news
```

Out[124]:

```
Search term:  Canada
```

	Term Canada in 160 docs	Total words(m)	Frequency(f)
0	Article #97	93	5
1	Article #98	73	1
2	Article #99	103	2
3	Article #100	82	1
4	Article #101	75	1

Figure 4: TF calculation



In [127]:

```
print("Number of documents: ",df_news.shape[0])
df_details
```

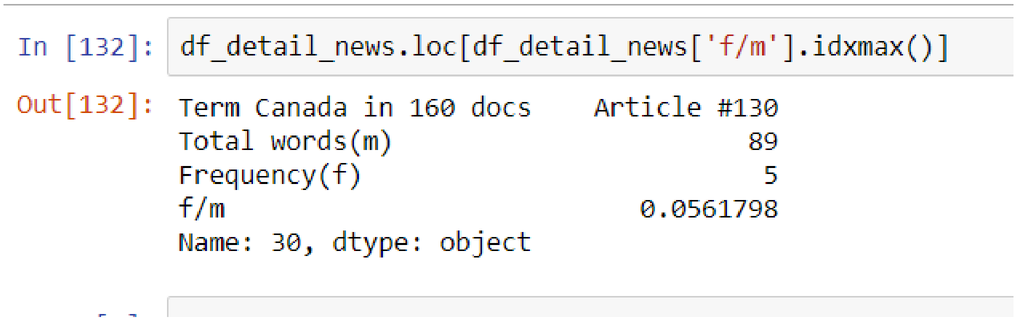
```
Number of documents:  481
```

Out[127]:

	Search Query	Document containing term(df)	Total Documents(N)/ number of documents term appeared (df)	Log10(N/df)
0	Canada	160	481/160	0.478025
1	University	98	481/98	0.690919
2	Dalhousie University	19	481/19	1.40339
3	Halifax	58	481/58	0.918717
4	Canada Education	0	481/0	NA

Figure 5: IDF calculation

Document with maximum f/m value.



In [132]:

```
df_detail_news.loc[df_detail_news['f/m'].idxmax()]
```

Out[132]:

```
Term Canada in 160 docs    Article #130
Total words(m)                       89
Frequency(f)                          5
f/m                           0.0561798
Name: 30, dtype: object
```

Figure 6: max f/m calculation

C. Business Intelligence

The main facts that I have analyzed based on retrieved data in Assignment -1 are:

- Faculty member count in each department
- Number of programs in each department
- Number of departments in each faculty
- Number of buildings in each campus.

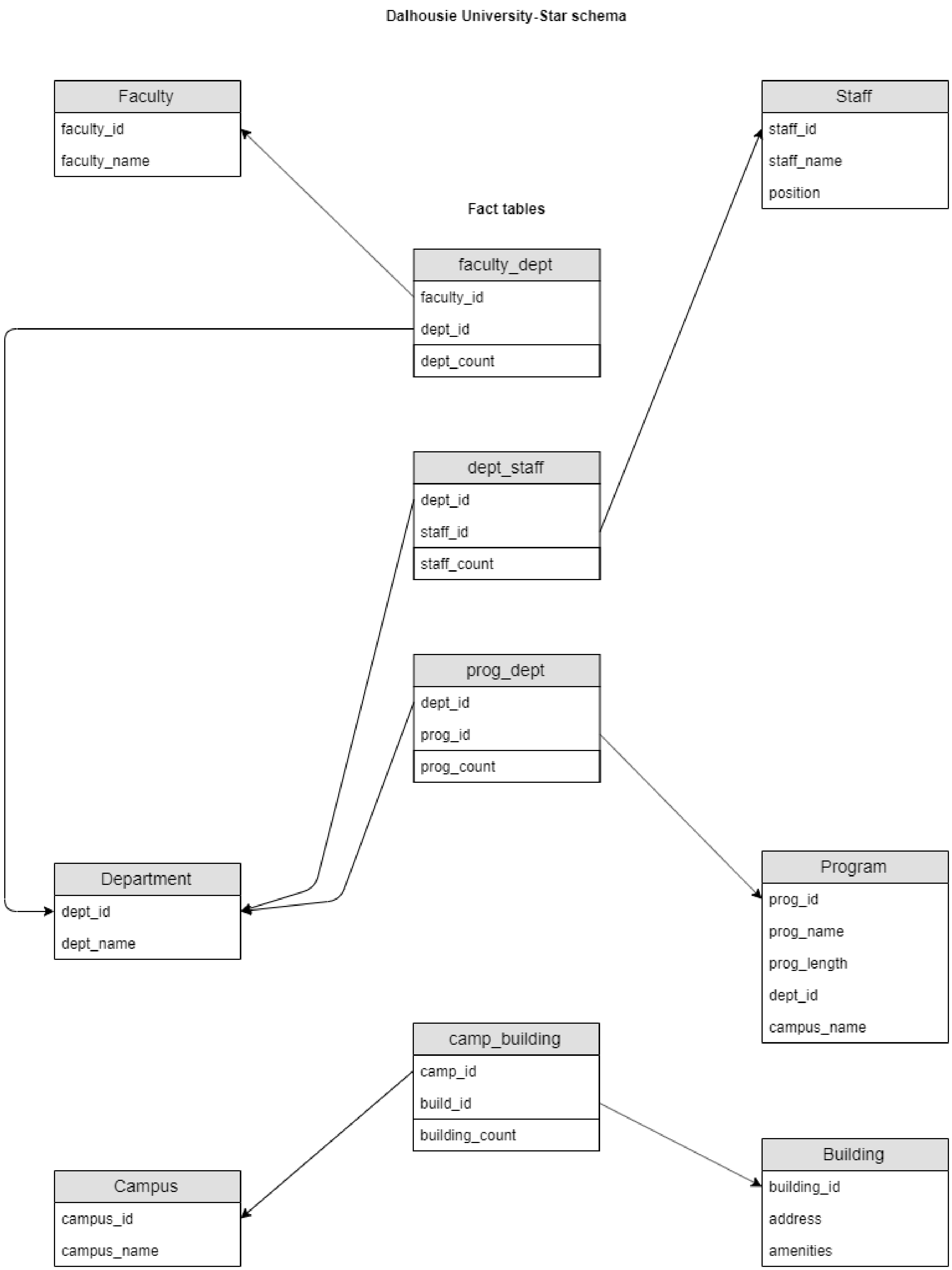
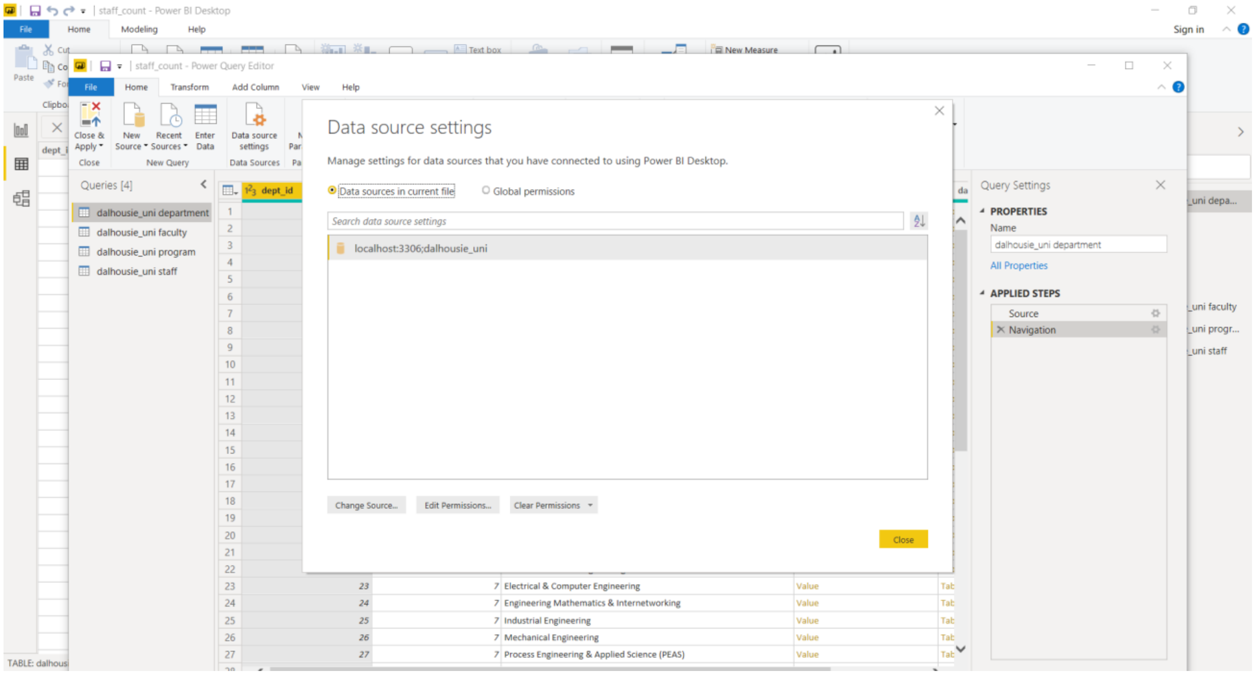


Figure 7: star schema



Data source settings

Manage settings for data sources that you have connected to using Power BI Desktop.

● Data sources in current file ○ Global permissions

Search data source settings

localhost:3306;dalhousie_uni

23	23	7	Electrical & Computer Engineering	Value
24	24	7	Engineering Mathematics & Internetworking	Value
25	25	7	Industrial Engineering	Value
26	26	7	Mechanical Engineering	Value
27	27	7	Process Engineering & Applied Science (PEAS)	Value

Figure 8:Local MySQL database source connected with Power BI

Measure for fact tables are created using DAX query [2]:

- `dept_count = CALCULATE(DISTINCTCOUNT('dalhousie_uni department'[dept_id]),GROUPBY('dalhousie_uni department','dalhousie_uni department'[fac_id]))`
- `program_count = CALCULATE(DISTINCTCOUNT('dalhousie_uni program'[degree]),GROUPBY('dalhousie_uni program','dalhousie_uni program'[dept_name]))`
- `faculty_member_count = CALCULATE(DISTINCTCOUNT('dalhousie_uni staff'[staff_id]),GROUPBY('dalhousie_uni staff','dalhousie_uni staff'[dept_id]))`

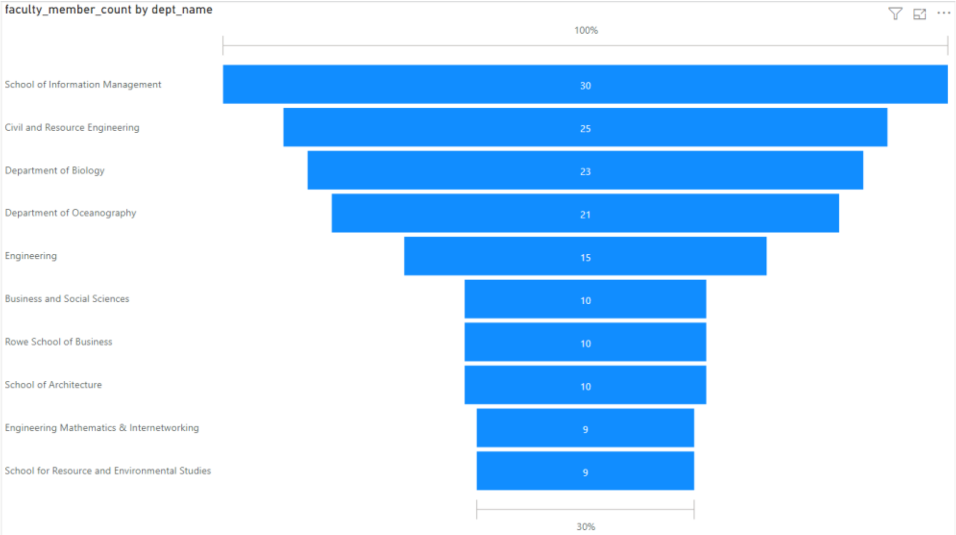


Figure 9: faculty_member_count by department

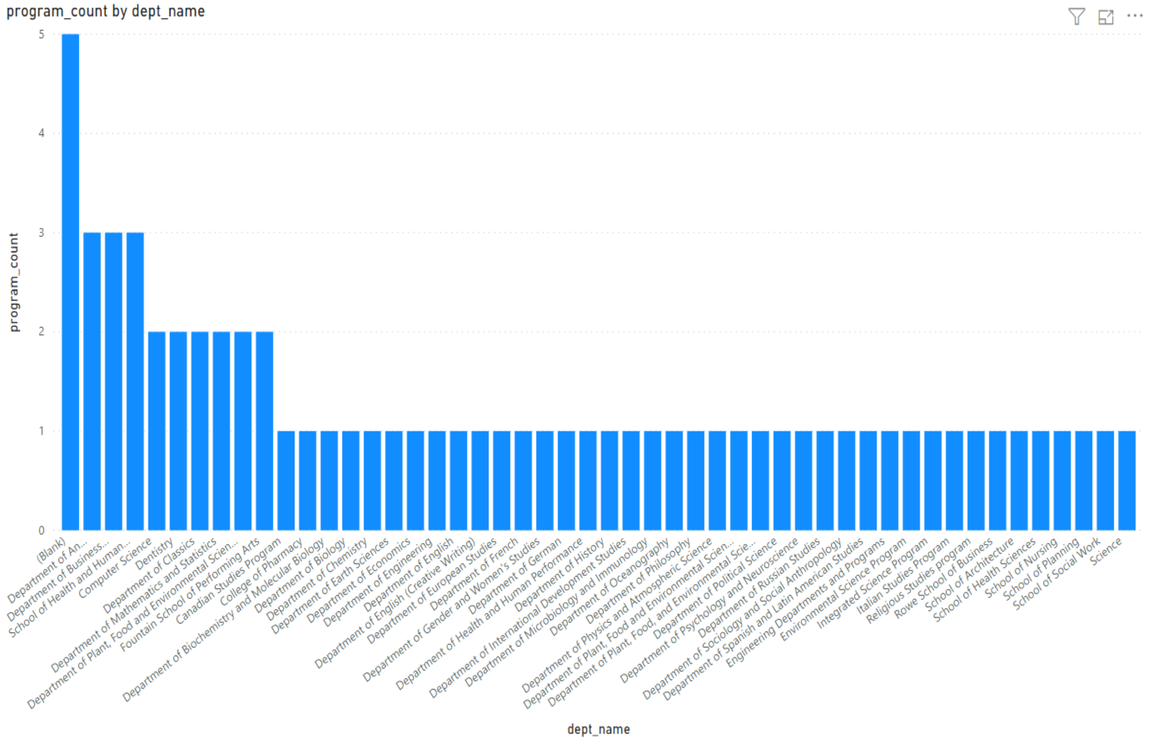


Figure 10: program_count by department

From above visualization we can see, computer science does not have highest number of programs, and number of programs (course) can also be easily visible.

Note: I have scraped data for only undergrad programs in each department in first assignment.

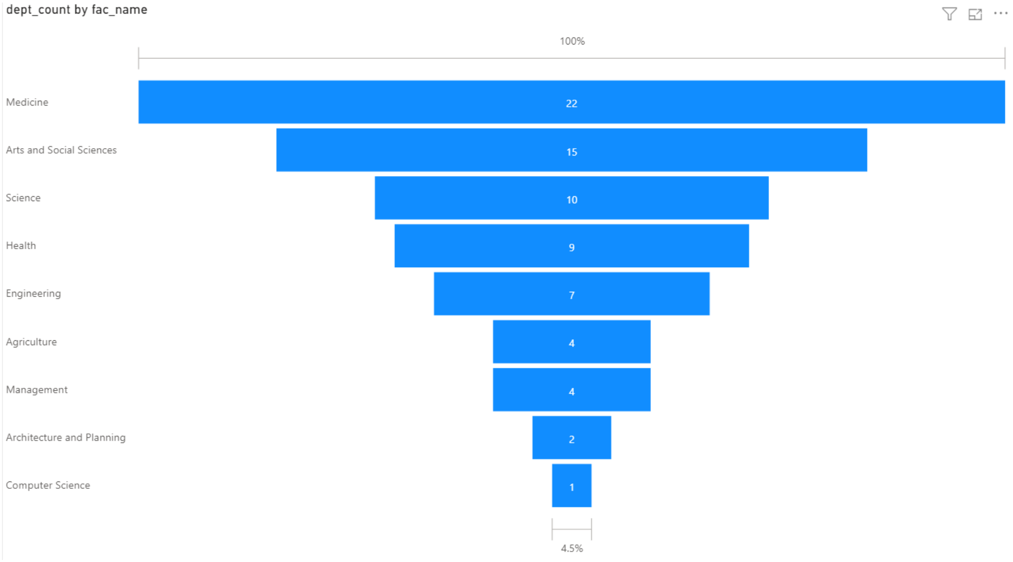


Figure 11: department_count by faculty

References

[1] “Opinion Mining, Sentiment Analysis, and Opinion Spam Detection” [Online] Available: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html [Accessed: 29-Nov-2019].

[2] “DISTINCT COUNT AND GROUP BY” Online Available: https://community.powerbi.com/t5/Desktop/DISTINCT-COUNT-AND-GROUP-BY/td-p/420123 [Accessed: 01-Dec-2019]