INFORMATION RETRIEVAL - COMP 479/6791

Prof: Dr. Sabine Bergler

Final Project Report

Fall 2015

Submitted by: Samer Ayoub (26750265)

Meet Anadkat (27122829)

Souleymane Guindo (21390708)

We acknowledge that this project is our complete team work, and complies with the expectation of originally stated by the ENCS code of conduct.

Project Overview

The goal behind this project is to perform sentiment analysis on the crawled documents and perform clusterization to differentiate these documents based on their sentiment values.

Design

Firstly, we used Websphinx tool to crawl each of the department from encs website of Concordia University separately and used the parameters mentioned in the project description. After crawling the required department documents they are passed to the indexer that we designed in the previous assignment to parse documents and generate inverted index (tf-idf).
Next step includes performing sentiment analysis on these crawled data. To achieve this, we used aFinn sentiment dictionary that contains sentiment score as integer value for each english word. Next, assigning these scores to their respective terms in each documents and calculating the score for each document by adding the score for each term and later dividing the sum with total number of tokens in their respective document. Note that this process will be done for each document inside each department.
Finally, summing up the score for each document gives the overall score for each department.

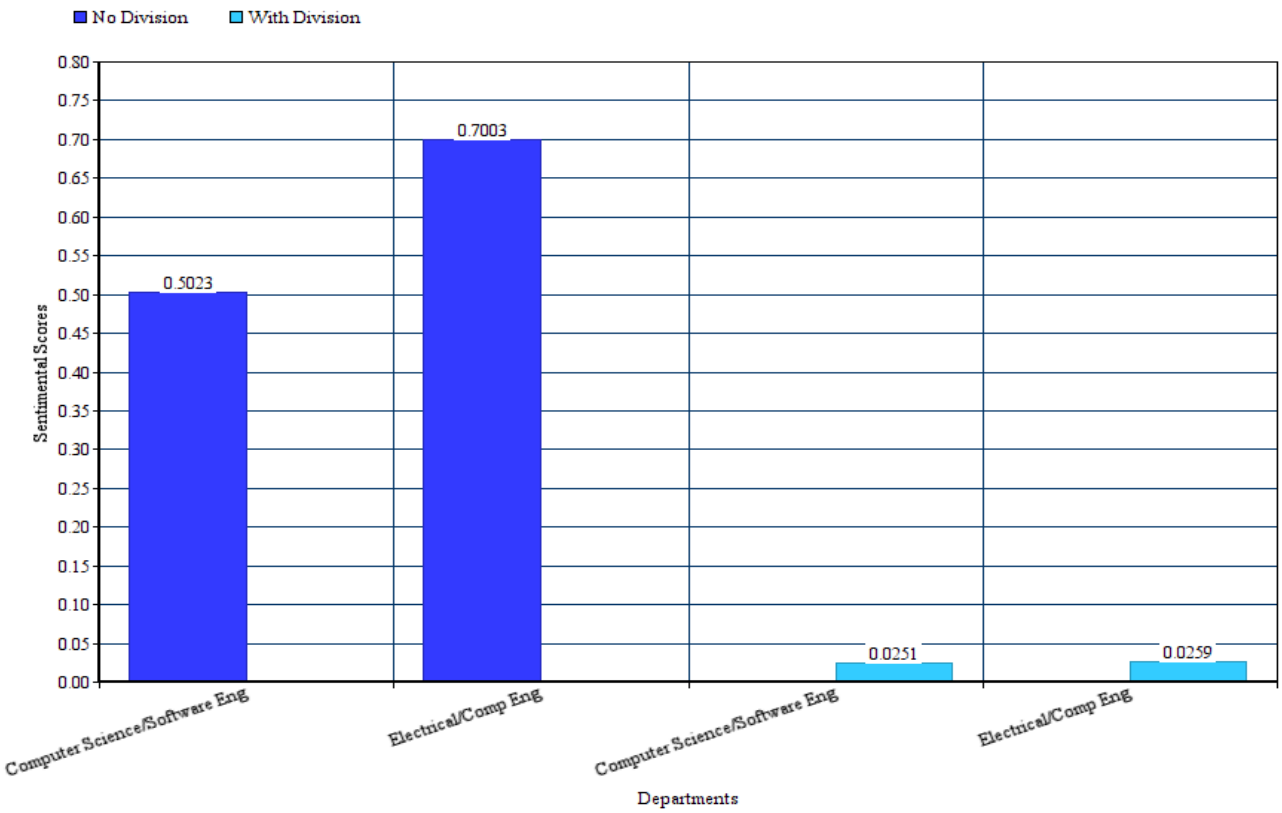
Which is the most positive Department in ENCS at Concordia?

Based on the calculation performed by Afinn on the numerous parsed documents, we noticed all departments are close to one another in terms of sentimental scores. The most positive department was noticed to be **Building, Civil, & Environmental Engineering** with a sentimental score of 0.8106 when not dividing by total document count.

However, the most positive department was **Information Systems Engineering** when the sentimental score was divided by the with a score of 0.0313.

Is Computer Science and Software Engineering more positive or less positive than Electrical and Computer Engineering?

Comparing the different sentimental scores together, although each department had some similar documents suchs about.html co-op.html and many others. Some unique files made them have a more positive score in relation to other departments. Using our program we can allowed us to get to the conclusion that the Electrical and Computer Engineering is more positive than Computer Science and Software Engineering(See Graph 1). Below are the different sentimental score prior to dividing by total document count and after dividing the sentimental score by the total document count. In both case Electrical and Computer Engineering is still more positive than Computer Science and Software Engineering(See Graph 1).



Graph 1: Computer Science/Software Eng VS Electrical Comp Eng.

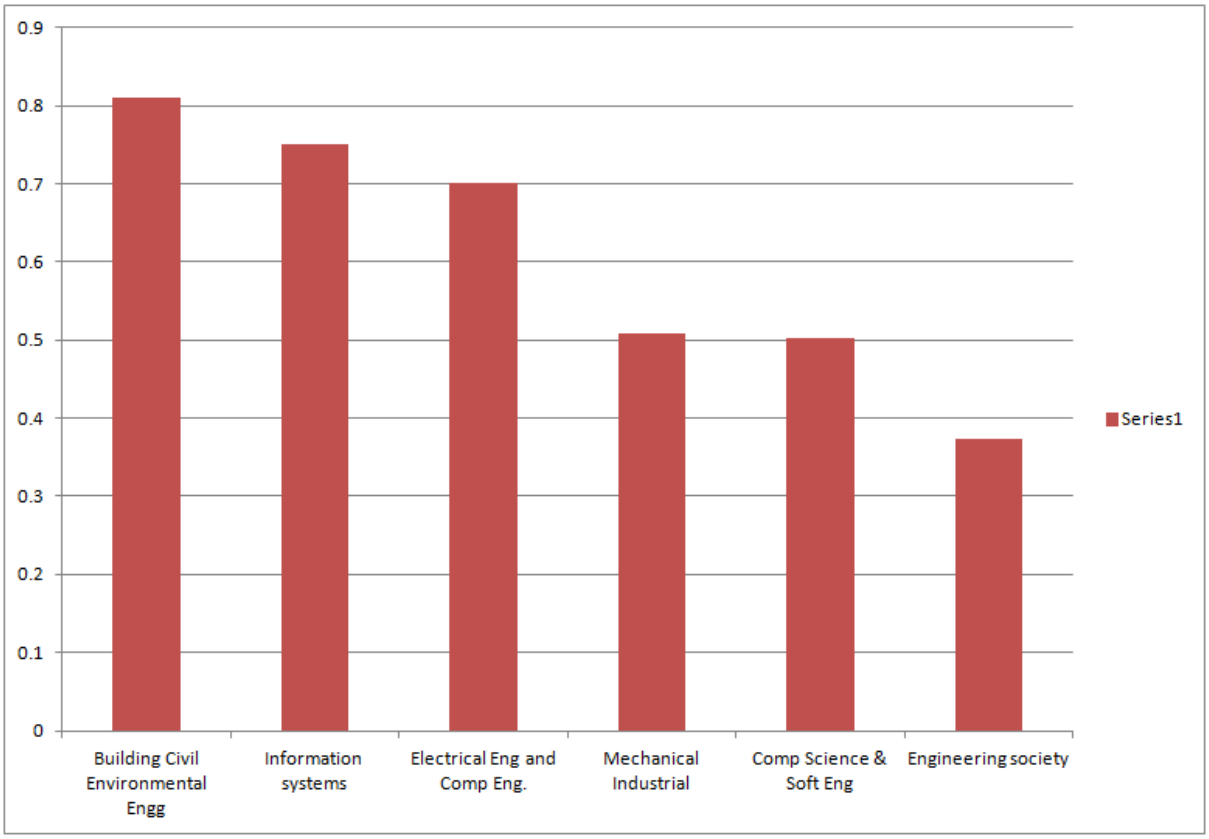
Ranking the departments in ENCS by sentiment score of their web documents

Department	Department sentiment score
Building Civil Environmental Engg	0.8106
Information systems	0.7501
Electrical Eng and Comp Eng.	0.7003
Mechanical Industrial	0.5081
Comp Science & Soft Eng	0.5023
Engineering society	0.3734

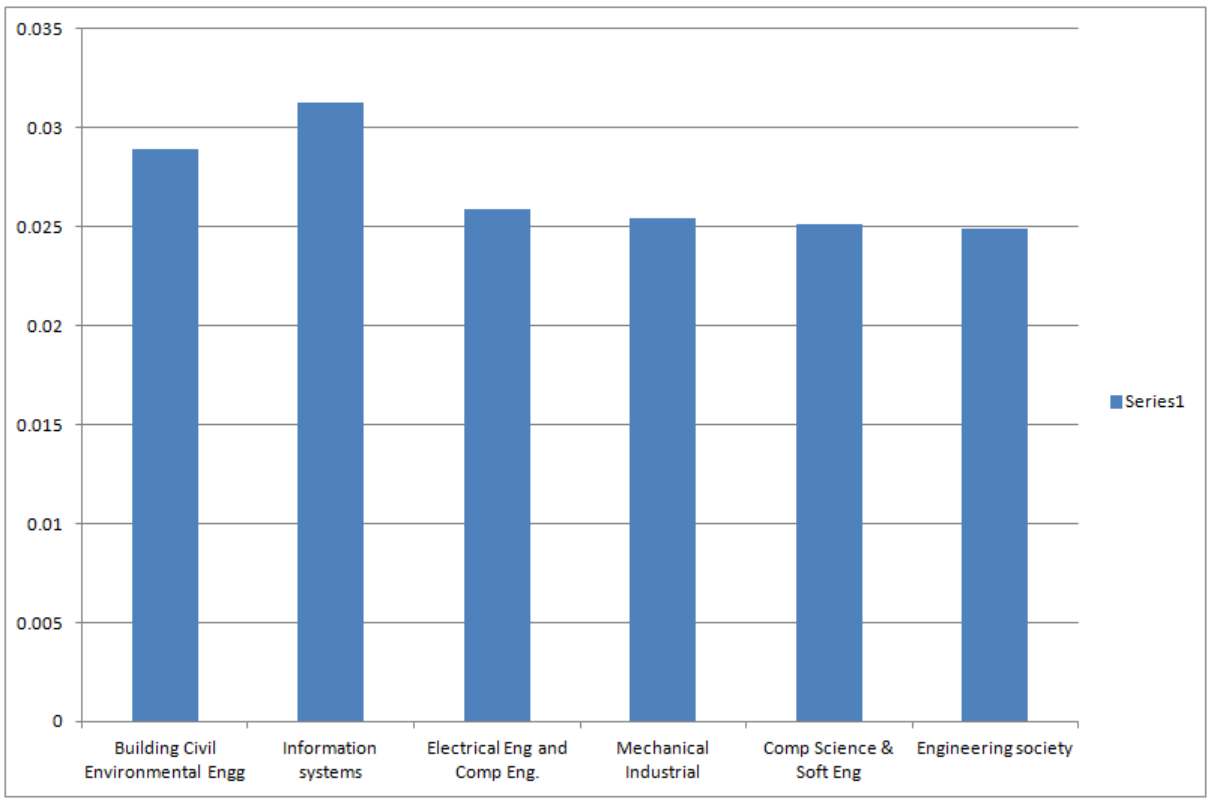
Table 1(Scenario 1): Department sentiment score = Total Score of each document

Department	Department sentiment score
Information systems	0.0313
Building Civil Environmental Engg	0.0289
Electrical Eng and Comp Eng.	0.0259
Mechanical Industrial	0.0254
Comp Science & Soft Eng	0.0251
Engineering society	0.0249

Table 2(Scenario 2): Department sentiment score = (Total score of each document / Total documents)



Scenario 1: No-Division by Total number of documents



Scenario 2: Division by Total number of documents (Averaged per document)

The Graphs above reveals some noticeable points that worth commenting on:

According to Scenario 1, based on the total documents sentiment scores of of each department; we notice from graph 1 that bcee was higher than info-sys department while According to Scenario 2, based on the average document sentiment score of each department;i we notice from graph 2, their ranking is reversed two departments vary significantly in their scores.

For the other 4 departments, Although the ranking doesn't really change, there is a significant change in the deviation, in Scenario 1; Electrical Eng and Comp Eng is remarkably higher while Engineering society is the lowest . In Scenario 2; the 4 departments are almost the same.

Based on this observations the reason behind this could be dependant on the number of more positive or more negative documents that is contained in the department. For example, between two departments, if they have equal ratio of positive to negative documents, then the department with higher value of these documents will have higher final score than the other department. However, by dividing their final scores with their respective number of documents removes the unfair advantage of the department with higher number of documents.

Moreover, Graph 1 shows clear separation between department scores compared to second Graph 2. However, the least scored department is same in both cases.

In conclusion Scenario 1 gives the sentiment score of a department while Scenario 2 give the average sentiment score of a document, hence In our opinion Scenario 1 is more less biased to judge the department rank

Classifying the departments in ENCS with a three way classifier into Positive, Negative, and Neutral

clustering

Scenario 1			Scenario 2		
Section	***Department***	***Range value***	***Section***	***Department***	***Range Value***
Positive	BCEE, INFO-SYS, ELEC	> 0.7	**Positive**	INFO-SYS BCEE	>0.0270
Neutral	MECH, COMP-SC	>0.5 && <0.7	**Neutral**	ELEC	>0.0255 && <0.0270
Negative	ENG-SOC	<0.5	**Negative**	MECH, COMP-SC ENG-SOC	<0.0255

Table 3: Classification

What was the hardest step?

When given the project it was we decided to get together and start breaking down the different tasked needed to be done in order to have a good deliverable. We started working with what we initially had from assignment 1 and assignment 2. As a group decision we used one of our group members version of the Assignment 1 and Assignment 2. The reason behind this because of the language they used, which was python and the different libraries that will facilitate the work for us. We used BeautifulSoup as our parser, NLTK for tokenizing/normalizing, and Porter-Stemmer for stemming. Combining the different tools went great; however, modifying the code from last the assignment to calculate sentiment analysis based on aFinn, navigating subdirectories, classifying documents according to each department and modifying the code accordingly was somehow of a challenge which we successfully overcame. In the end our program was able to function based on the expectation required from the assignment description.

How big is the index?

After performing the necessary compression on the corpus obtained from crawling the websites. We have obtained the following information in regard to our index size.

Remove Punctuation :	True
Case Fold :	False
Filter Numbers :	False
Filter StopWords :	False
Stemming :	False

Total Number of Tokens :	192608 tokens
Total Number of dictionary terms:	7713 terms
Total number of postings:	53331 postings

What observations did you make during your experiments?

As per the graphs shown below, we noticed significant changes; like in the first graph there were two departments which are equally higher while in the second graph, there was quite clear separation between these two departments.

Experiments

While calculating the final score for each department, we considered two scenarios and making our observations on the results of these two scenarios.

Scenario 1 : Final department score equals to the sum of each of its documents

Scenario 2 **:** Final department score equals to the sum of each of its documents and dividing this sum with total number of documents(Average sentiment score per document for a department) .

Outputs: every document sentiment score
001 bcee_About the department ====> 0.024066390041493777
002 bcee_Contact ====> 0.017628205128205128
003 bcee_Facilities & services ====> 0.028783658310120707
004 bcee_News & events ====> 0.03178694158075601
005 bcee_Programs ====> 0.011922503725782414
006 bcee_Research ====> 0.03904170363797693
007 bcee_Student life ====> 0.031568228105906315
008 bcee_Current students ====> 0.023305084745762712
009 bcee_Faculty members ====> 0.01490780698313064
010 bcee_Job opportunities ====> 0.03821656050955414
011 bcee_Staff ====> 0.027079303675048357
012 bcee_Lab Safety ====> 0.027156549520766772
013 bcee_Research Labs ====> 0.026859504132231406
014 bcee_Teaching Labs ====> 0.025962399283795883
015 bcee_Notices ====> 0.047058823529411764
016 bcee_Building Engineering ====> 0.03059490084985836
017 bcee_Civil Engineering ====> 0.02098849018280298
018 bcee_Co-op ====> 0.03383458646616541
019 bcee_Environmental Engineering ====> 0.011627906976744186
020 bcee_Graduate programs ====> 0.02364864864864865
021 bcee_Undergraduate programs ====> 0.025472473294987676
022 bcee_Building Engineering ====> 0.03333333333333333
023 bcee_Research centres ====> 0.042
024 bcee_Civil Engineering ====> 0.026476578411405296

025 bcee_Environmental Engineering ====> 0.03185840707964602
026 bcee_Researchers ====> 0.011431184270690443
027 bcee_Student & Professional Associations ====> 0.029596412556053813
028 bcee_Scholarships & awards ====> 0.07443531827515401

bcee ======> Total number of documents = 28
Sentiment Score = 0.8106419032554333
Average Sentiment Score = 0.0289514965448369

---

029 computer-science-software-engineering_About the department ====> 0.028146989835809225
030 computer-science-software-engineering_Contact ====> 0.01527331189710611
031 computer-science-software-engineering_News & events ====> 0.024744027303754267
032 computer-science-software-engineering_Programs ====> 0.015232292460015232
033 computer-science-software-engineering_Research ====> 0.02432179607109448
034 computer-science-software-engineering_Facilities & services ====> 0.020879940343027592
035 computer-science-software-engineering_Student life ====> 0.025693730729701953
036 computer-science-software-engineering_Current student resources ====> 0.028763183125599234
037 computer-science-software-engineering_Faculty members ====> 0.017264276228419653
038 computer-science-software-engineering_Job opportunities ====> 0.029789719626168224
039 computer-science-software-engineering_Administration & staff ====> 0.030381944444444444
040 computer-science-software-engineering_Notices ====> 0.019409282700421943
041 computer-science-software-engineering_Graduate programs ====> 0.027796161482461945
042 computer-science-software-engineering_Undergraduate programs ====> 0.016566265060240965
043 computer-science-software-engineering_Research centres ====> 0.046806649168853895
044 computer-science-software-engineering_Grants & funding ====> 0.03025347506132461
045 computer-science-software-engineering_Research groups ====> 0.01619644723092999
046 computer-science-software-engineering_Industry sponsors ====> 0.03155522163786627
047 computer-science-software-engineering_Visiting researchers ====> 0.02388535031847134
048 computer-science-software-engineering_Student associations ====> 0.029343629343629343

computer-science-software-engineering ======> Total number of documents = 20
Sentiment Score = 0.5023036940693407
Average Sentiment Score = 0.025115184703467037

---

049 electrical-computer_About the department ====> 0.034220532319391636
050 electrical-computer_Contact ====> 0.01638176638176638

051 electrical-computer_Facilities & services ====> 0.033707865168539325
052 electrical-computer_News & events ====> 0.03037037037037037
053 electrical-computer_Programs ====> 0.026525198938992044
054 electrical-computer_Programs ====> 0.026525198938992044
055 electrical-computer_Research ====> 0.023397761953204477
056 electrical-computer_Student resources ====> 0.030983733539891558
057 electrical-computer_Faculty members ====> 0.018370607028753993
058 electrical-computer_Job opportunities ====> 0.03913630229419703
059 electrical-computer_Staff ====> 0.019981834695731154
060 electrical-computer_Computer facilities ====> 0.02211126961483595
061 electrical-computer_Department calendar ====> 0.023662551440329218
062 electrical-computer_Teaching labs ====> 0.014186851211072665
063 electrical-computer_Utilities & tools ====> 0.04103671706263499
064 electrical-computer_Computer Engineering (BEng) ====> 0.01572052401746725
065 electrical-computer_Electrical Engineering (BEng) ====> 0.01678445229681979
066 electrical-computer_Co-op programs ====> 0.03580901856763926
067 electrical-computer_Electrical & Computer Engineering (MASc) ====>
0.03349056603773585
068 electrical-computer_Electrical & Computer Engineering (MEng) ====>
0.029974307736226093
069 electrical-computer_Electrical & Computer Engineering (PhD) ====> 0.030078125
070 electrical-computer_Research areas ====> 0.02559150169000483
071 electrical-computer_Professional activities ====> 0.021359223300970873
072 electrical-computer_Recent publications ====> 0.023835319609967497
073 electrical-computer_Recent theses ====> 0.023758099352051837
074 electrical-computer_Researchers ====> 0.021205357142857144
075 electrical-computer_Capstone ====> 0.022119815668202765

electrical-computer ======> Total number of documents = 27
Sentiment Score = 0.700324871378646
Average Sentiment Score = 0.025937958199209108

---

076 eng-society_About the Faculty ====> 0.01649928263988522
077 eng-society_Academics ====> 0.027055150884495317
078 eng-society_Contact ====> 0.01740506329113924
079 eng-society_Events ====> 0.023529411764705882
080 eng-society_Facilities & services ====> 0.032454361054766734
081 eng-society_News & events ====> 0.032148900169204735
082 eng-society_Research ====> 0.03873598369011213
083 eng-society_Student life ====> 0.027385159010600707
084 eng-society_About the Centre ====> 0.02356902356902357
085 eng-society_Courses ====> 0.02090032154340836
086 eng-society_Individualized program ====> 0.0244926522043387
087 eng-society_Research ====> 0.028288543140028287
088 eng-society_Faculty members ====> 0.015498154981549815
089 eng-society_Honours Research Project (ENGR 412) ====> 0.023547880690737835
090 eng-society_Technical Report (ENGR 411) ====> 0.02185430463576159

eng-society ======> Total number of documents = 15
Sentiment Score = 0.3733641932697581
Average Sentiment Score = 0.024890946217983875

091 info-systems-eng_About the Institute ====> 0.03169572107765452
092 info-systems-eng_Contact ====> 0.01146288209606987
093 info-systems-eng_Facilities & services ====> 0.03333333333333333
094 info-systems-eng_News & events ====> 0.026785714285714284
095 info-systems-eng_Programs ====> 0.025261860751694395
096 info-systems-eng_Research ====> 0.027436140018921477
097 info-systems-eng_Student life ====> 0.029608404966571154
098 info-systems-eng_Current students ====> 0.02521613832853026
099 info-systems-eng_Faculty members ====> 0.030723488602576808
100 info-systems-eng_Job opportunities ====> 0.03951612903225806
101 info-systems-eng_Staff ====> 0.02787769784172662
102 info-systems-eng_Notices ====> 0.023827252419955324
103 info-systems-eng_Seminars ====> 0.032915360501567396
104 info-systems-eng_3D Graphics & Game Development (Grad. Cert.) ====>
0.028160200250312892
105 info-systems-eng_Co-operative education (Co-op) ====> 0.039401103230890466
106 info-systems-eng_Information & Systems Engineering (PhD) ====>
0.030458715596330274
107 info-systems-eng_Information Systems Security (MASc) ====>
0.03642671292281006
108 info-systems-eng_Information Systems Security (MEng) ====>
0.03128371089536138
109 info-systems-eng_Quality Systems Engineering (MASc) ====> 0.0441367373431415
110 info-systems-eng_Quality Systems Engineering (MEng) ====>
0.04483837330552659
111 info-systems-eng_Service Engineering & Network Management (Grad. Cert.) ====>
0.027743526510480888
112 info-systems-eng_Funding & grants ====> 0.028199566160520606
113 info-systems-eng_Researchers ====> 0.03512259774685222
114 info-systems-eng_Student & Professional Associations ====>
0.03870967741935484

info-systems-eng ======> Total number of documents = 24
Sentiment Score = 0.7501410446381551
Average Sentiment Score = 0.03125587685992313

115 mechanical-industrial_About the department ====> 0.037527593818984545
116 mechanical-industrial_Contact ====> 0.019815994338287332
117 mechanical-industrial_Facilities & services ====> 0.031746031746031744
118 mechanical-industrial_News & events ====> 0.028374892519346516
119 mechanical-industrial_News & events ====> 0.02570694087403599
120 mechanical-industrial_Programs ====> 0.025992779783393503
121 mechanical-industrial_Research ====> 0.011111111111111112
122 mechanical-industrial_Student life ====> 0.03046218487394958
123 mechanical-industrial_Current students ====> 0.02282453637660485
124 mechanical-industrial_Job opportunities ====> 0.04209690230341541
125 mechanical-industrial_Lab Safety ====> 0.02947845804988662
126 mechanical-industrial_Research Labs ====> 0.015894955079474776
127 mechanical-industrial_Teaching Labs ====> 0.023391812865497075
128 mechanical-industrial_Notices ====> 0.025490196078431372
129 mechanical-industrial_Degree accreditation ====> 0.027168234064785787

130 mechanical-industrial_Graduate programs ====> 0.023349436392914653
131 mechanical-industrial_Undergraduate programs ====> 0.024026512013256007
132 mechanical-industrial_Student & Professional Associations ====>
0.028094820017559263
133 mechanical-industrial_Graduate students ====> 0.016937669376693765
134 mechanical-industrial_Undergraduate students ====> 0.018569087930092845

mechanical-industrial ======> Total number of documents = 20
Sentiment Score = 0.5080601496137528
Average Sentiment Score = 0.02540300748068764

Total documents = 134
Inverted Index is created in : 24.691413164138794 seconds

Raw scores for each department

[["bcee", 0.8106419032554333], ["info-systems-eng", 0.7501410446381551], ["electrical-computer", 0.700324871378646], ["mechanical-industrial", 0.5080601496137528], ["computer-science-software-engineering", 0.5023036940693407], ["eng-society", 0.3733641932697581]]

Average scores per document for each department

[["info-systems-eng", 0.03125587685992313], ["bcee", 0.0289514965448369], ["electrical-computer", 0.025937958199209108], ["mechanical-industrial", 0.02540300748068764], ["computer-science-software-engineering", 0.025115184703467037], ["eng-society", 0.024890946217983875]]

What did you learn from your experience?

Overall, working on this project has contributed a lot to our knowledge about information retrieval. Prior to undertaking this project, we didn't have a lot of knowledge about web crawling and sentimental analysis. Moreover, we had no understanding of how this process was done over a large corpus of data. We got the overview of how sentiment analysis could be done using information retrieval techniques which could be used for analysing the data and making important conclusions (like which department would attract more users based on their sentiment score). We can now comfortably say that we are familiar with this topic of information retrieval. Furthermore, this project also helped enhance our python development skills as well as our problem solving and team-work skills.