Zirui Jiao, Shiyue (Shirley) Zong

Project Overview

We choose to use Wikipedia as our major source and we use “computing summary statistics” and “natural language processing” as two techniques to analyze the word on Wikipedia page. We choose Wikipedia page of four undergraduate school in Boston area, namely “Babson College, Boston University, Boston College and Bentley College” as the text since we are curious about what are similarities and differences between schools to introduce the institution and how Wikipedia describe the schools (negatively or positively). We also want to learn what keywords each institution emphasizes when describing itself.

Implementation

For the two techniques, we used accordingly fitting algorithms. As to get the most frequent words for each school’s introduction, we utilized series of functions that first get the words from the page to a dictionary, containing each unique word and its frequency of showing up on the page content. Secondly, we ran a text file called “stopwords.txt”, containing all the common words like prepositions, and we excluded all these words from our list to be analyzed. We made sure that the words that we are analyzing, are the ones that can actually describe the institute. Last be not least, we print out the 12 most common words appear in the institutes’ description.

Our second technique used to test the sentiment of the text in the page content we found. We imported the Sensitivity Intensity Analyzer and printed out score for each positive, neutral, or negative sentiment for each institute description.

Results

From our analysis, we found some expected results and unexpected ones.

It was pretty within our expectation that Wikipedia doesn’t have any biased attitudes added on the descriptions for such institutes. Therefore for each school, the highest sentiment score lies for the neutral sector. Even though the descriptions are mostly neutral, those content still tend to be more positive than negative. All the schools got 0.0 for negative sentiment. This strict no negativity was quite surprising.

What’s more interesting was about the most frequent words. Wikipedia is definitely one of the best way to present a school to the mess audience who first come across with the name. The more the keyword shows up in the description, the better off the SEO for the school’s marketing and branding. Among these school, we found Babson actually did best in emphasizing on its strengths— business and entrepreneurship, while for the other schools, the most frequent words were less relevant about their core strength in academic areas. Among the rest of the three, Boston College has the second best SEO friendly keywords. For its 10th most frequent word- Jesuit- it presents clearly the core identity of itself as a Catholic school.

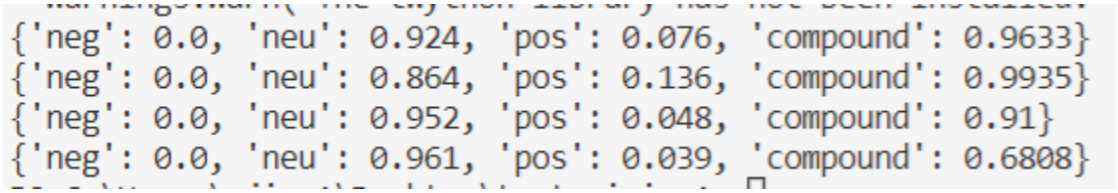
Although it makes sense that for bigger universities, it's harder to identify a specific strength to represent themselves. However, in order to promote the school for their online-presence and make sure the schools show up to their most relevant keyword searches by potential students or scholars who are interested, the institutes should work on developing their keywords and make sure the words get repeated to a good extent.

Reflection

The process of the project starts smoothly in the beginning when we decide the data and techniques to use. Even though the natural language processing is a new topic, its function is very easy to adopt and analyze. We met several difficulties when computing the top words in each texts, such as building the function, deleting the stopwords and processing the text from a webpage. As our model is based on Wikipedia, we can use it to analyze every Wikipedia pages in the future to analyze the top words and sensitivity tests. If we have more time, we could apply other techniques such as "text similarity" and data visualization to make the project more visual and straightforward.

As we choose two techniques, Shirley chooses to do "summary statistics" and I focuses on "natural language processing"; however, we help each along the assignment. For the write-up part, as both of us understand our topic, we are able to write every part individually and combine together. The main issue of the corporation is time arrangement during this week, as both of us are busy with exams, assignments and travel arrangement.

nltk sensitivity output:



Babson most frequent 12 keywords:

1. babson
2. college
3. business
4. former
5. students
6. one
7. babson's
8. ranked
9. founder
10. program
11. president
12. entrepreneurship

Bentley's

1. bentley
2. campus
3. university
4. men's
5. north
6. division
7. college
8. business
9. sigma
10. programs
11. team
12. school

BU's

1. university
2. boston
3. campus
4. bu
5. students
6. school
7. college
8. new
9. program
10. university's
11. center
12. research

BC's

1. boston
2. college
3. school
4. campus
5. university
6. hall
7. bc
8. student
9. program
10. jesuit
11. new
12. first