Homework 5 – Bag of Words

Arya Rahmanian

COMP 614

November 15, 2024

CodeSkulptor Link: https://py3.codeskulptor.org/#user309_bGzaB1xQVo_0.py

1. **Find the top 30 words in "rice_university.xml". Provide the words and their (raw/absolute, not normalized) counts in the form of a list of tuples of (word, count), sorted in descending order. Do the results align with your expectations? Why or why not?**

Top 30 words in rice_university.xml:

Word	Count	Word	Count
the	616	students	54
and	341	at	45
of	299	its	43
in	257	with	40
rice	190	research	38
a	173	first	37
to	159	student	36
for	126	rice's	35
was	78	are	35
as	74	has	33
on	67	that	30
is	65	colleges	30
by	64	an	27

university	63	which	26
campus	58	center	26

The top 30 words align with what is expected. Unless we filter out common words and conjunctions like “the”, “and, “for”, these will be our top words. But once those words are out of the way, words like “student” and “campus” are very expected to be in the top 30 as it is a large university.

2. **Execute the provided run() function (be careful not to modify this function), and select the number corresponding to "Rice University" (141). Which are the five nearest neighbors of the article that you chose? Do the results align with your expectations? Why or why not?**

The 5 nearest neighbors of Rice University are:

University of Southern California

University of Michigan

Trinity University (Texas)

University of Florida

Columbia University

These results align with what is expected. Most of the nearest neighbors are large and successful private universities like USC, and all the nearest neighbors are large US universities. Nothing out the ordinary.

3. **Choose a different article that is not about a university, and perform the same two analyses for that article. What is the title of the article that you selected? What are the top 30 words in the article? Provide the words and their counts in the form of a**

list of tuples of (word, count), sorted in descending order. What are the five nearest neighbors of the article that you chose? Do the results of these two analyses align with your expectations? Why or why not? Identify one nearest neighbor that is unexpected or one article from the provided training corpus that you expected to be a nearest neighbor but wasn't, and hypothesize why that article may have been present or absent in the collection of nearest neighbors.

Topic Chosen: **Television**

a. **Top 30 words**

Word	Count	Word	Count
the	802	that	94
of	391	first	78
in	389	an	66
a	355	from	66
and	351	it	64
television	310	s	61
to	251	color	60
was	139	at	56
for	137	system	54
by	135	be	53
as	122	tv	48
on	106	used	46
or	100	which	45
is	99	are	45
with	94	also	44

b. Five Nearest Neighbors

The 5 nearest neighbors of Television are:

The Beatles

Manhattan Project

Herman Melville

Global warming controversy

Pottery

The nearest neighbors here are quite surprising; I would have anticipated seeing other significant technological inventions, but none are listed. What stands out most is that the Manhattan Project is the second nearest neighbor to Television, which I would never have expected. Upon researching, I found that the television was invented in 1927. My only theory is that both the Manhattan Project (Nuclear Bomb) and television were developed around the same period, during the WWII era. Furthermore, while both are considered monumental inventions for humanity, they have vastly different impacts and reasons for their significance.

Reflection

1. **What were the two most important concepts and/or skills that were reinforced by this assignment? Why do you believe that these concepts are important?**

   The two most important concepts introduced in this assignment is data collection/cleaning and k nearest neighbors. Data collection is an important part of data science, as we need data to build models on. But most of the time the data is messy, so we must clean it for our specific use case. Next, KNN is a widely used and well-known statistical tool to help visualize and group concepts in data.

2. **How do the skills and concepts that you applied in this assignment transcend the specific application/problem that you were tackling? How could you envision applying these skills and concepts again in the future?**

   As I mentioned earlier, all data scientists work with data. Whenever data is involved, someone must gather and clean it so we can draw conclusions. This concept applies to all scientists who deal with data. I can envision myself applying these principles to datasets—such as cancer rates across the country—and using KNN to cluster areas with high cancer rates.

3. **What do you believe you did well on this assignment? If you could do this assignment over, would you do anything differently? Why or why not?**

   I believe I did well on writing efficient code to scan and gather all the text elements and clean it up without too much overhead. I am confident that my solution was very efficient. If I had to redo this assignment again, I would not do anything different, as I am very happy with my result.

4. **Do you think you're comfortable enough with the concepts covered in this assignment that you would be capable of teaching them to a peer? Why or why not?**

   Yes, I am comfortable with the concepts in this assignment. I have used KNN before on other assignments in previous courses. In my opinion, it’s an incredibly useful but not so difficult concept to learn. Additionally, I have worked with large datasets before and had to clean up the information so I can effectively develop a model off it.