CS595 Intro to Web Science, Assignment #9

Valentina Neblitt-Jones

December 5, 2013

List of Tables

Listings

List of Figures

Question 1

Create a blog-term matrix. Start by grabbing 100 blogs; include:

- `http://f-measure.blogspot.com/`
- `http://ws-dl.blogspot.com/`

and grab 98 more as per the method shown in class.

Use the blog title as the identifier for each blog (and row of the matrix). Use the terms from every item/title (RSS) or entry/title (Atom) for the columns of the matrix. The values are the frequency of occurrence. Essentially you are replicating the format of the "blogdata.txt" file included with the PCI book code. Limit the number of terms to the most "popular" (i.e., frequent) 500 terms, this is **after** the criteria on p. 32 (slide 7) has been satisfied.

Answer to Question 1

Use the book here [1]

Question 2

Create an ASCII and JPEG dendrogram that clusters (i.e., HAC) the most similar blogs (see slides 12 & 13). Include the JPEG in your report and upload the ASCII file to GitHub (it will be too unwieldy for inclusion in the report.)

Answer to Question 2

Question 3

Cluster the blogs using K-Means, using k=5, 10, 20 (see slide 18). How many iterations were required for each value of k?

Answer to Question 3

Question 4

Use MDS to create a JPG of the blogs similar to slide 29. How many iterations were required?

Answer to Question 4

Question 5 - Extra Credit (5 points)

Re-run Q2, but this time with proper TFIDF calculations instead of the hack discussed on slide 7 (p. 32). Use the same 500 words, but this time replace their frequency count with TFIDF scores as computed in assignment #3. Document the code, techniques, methods, etc. used to generate these TFIDF values. Upload the new file to GitHub.

Compare and contrast the resulting dendrogram with the dendrogram from Q2.

Note: Ideally you would not reuse the same 500 terms and instead come up with TFIDF scores for all the terms and choose the top 500 from that list, but I am trying to limit the amount of work necessary.

Answer to Question 5

Use the book here [1]

A Appendix Title

References

[1] Segaran, T. *Programming Collective Intelligence: Building Smart Web 2.0 Applications.* O'Reilly, 2007.