MSDS 5163 Data Mining

Draft Final Group Project Assignment Fall 2016

Due date as shown on lipscomb.blackboard.com

Unstructured Internet Data Mining with Spectral Clustering and NLP Document Similarity

THE PROBLEM

You are a data scientist with Big Data Science Incorporated (BDSI), and you are instructed to perform spectral clustering to investigate the hidden relational objects on the Internet (or later your own intranet) of interest to your company's research keywords and URL's.

The overall goal is to data mine the Internet to find content which is "invisible" to the research team, but that must discovered by the data scientist in order for the contracting client company to remain competitive.

Your task is to data mine unstructured content on the Internet to find the relationship of the company research pages and or items of interest.

1. Several URL's and keywords in a CSV or space delimited text file list will be used as input
2. The actual file will be provided for the final project report, but is not available to you prior to the study.
3. For the k-th URL/keyword element of the set {k=1,...,K}, find the top N=34 or so matches on gigablast.com, excluding circular URL domain matches from the original URL domain, and collect (GET) the HTML document file of each of the N reference hits (exclude advertisements and indirect findings).
4. You should assume that all keywords that are also domain names shall be prefixed with *http://* if they are an explicit link to a page; otherwise, the domain name is only a keyword.
5. Example code may be shared to advance the class, however, each student should endeavor to have a working version of the code on his or her computer.
6. Your final report should provide an interpretation of the results, and
7. you should demonstrate a means of testing the code model to validate your hypothesis.

8. Good visualizations of the similarity matrices, heatmaps, or network graphs should be employed if time permits.
9. The reporting format is prescribed in the syllabus for this class.
10. At least 50 percent of the content of your final report shall be your own work. Content from other sources must be given attribution immediately next to the content as is feasible.

Implementation Notes

1. It is recommended that you use the gigablast.com search engine that will allow you to make a scripted API query in Python3 and return a captured result, which can be parsed, such as in JSON or XML. The top N URL matches will then be retrieved in HTML and converted to text using HTML to text rendering tool (e.g. python beautiful soup), excluding the circular references.
2. You will place these top N matches to each member of the set {k=1,…,K} URL's and keywords into the appropriate directory for computation of the cosine similarity matrix which may be performed as demonstrated in class using code from http://computergodzilla.blogspot.com.tr/2015/01/calculate-cosine-similarity-using.html or other code of your own creation as long as it is clearly validated. Recent results using the R library "tm" for unstructured documents show good promise.
3. Perform spectral clustering of the similarity matrix above using R or Python. If you use specclust(), then you will have to modify the function to accept the similarity matrix. A reference copy of spectral clustering in R is available as written by myself.
4. Graph or plot the results to illustrate the clusters as best possible within the available time. If possible, identify the key cluster memberships and make interpretations in your final report.
5. Write final report and submit to lipscomb.blackboard.com in original ODT or DOC and PDF format.

   (a) submit a working copy of all code used to implement your solution and

   (b) a console log of a typical execution session.

   Note: you can submit multiple documents for a single upload assignment.

   The actual set of {k=1,…,K} URL's will be provided in a text file as a list with one URL or keywork per line (ended by a line return).

Group Instructions

You are encouraged to work in small groups to complete the work on the project. Each team member shall contribute parts of the group work towards the final report and presentation. Final reports must be each submitted by each student and shall include **attribution** for portions of the report content which are not the original work of the submitting student, such as figures, programs, etc., which are the works of the team, or team members (actual names must be given), or other external material references in the literature. The main body of text in the report should mostly be your own composition, but the minimum original content by each student shall be at least 40 percent. The SafeAssign resource on blackboard.com will be used to compute the percent of shared content and shown to you upon submission.

Formal Report Format and Submission Requirements

The format of the report should be in the example scientific format given in the syllabus for this course in section F subsection IV. The final report will be due on the last day of class.

Formal Presentation of Project

Each group and member of the group will be required to present their results and provide discussion of hypothesis, analysis, and conclusions on the last day of class. The projector/HDTV will be available with HDMI in the Spark classroom. Teams are encouraged to demonstrate running programs of various components or analytics developed.

Note 9/2/16: Recently, gigablast has started requiring a user key to manipulate the data stream/query. The fee is $5.00 for a key, and I will be happy to reimburse the group for the key.

Questions?

Please email or post questions to the class blog, and I will post responses to the class as appropriate.