Big Data Analytics Final

Pubmed Central Topic Visualization

**Task:**

In this final, you will utilize the Pubmed Central (PMC) open access dataset, implement a system to process the PMC dataset with a user-provided term, and generate a JSON file for use in D3-driven visualization. The goal of this final is to produce a visualization of the topics that most commonly co-occur in the pubmed documents with a user provided term or phrase. During this final project, there are also some learning objectives:

1. Become more familiar with Big Data text processing
2. Become familiar with methods of representing results in an intermediate data interchange format (e.g. flare.json)
3. Explore various methods of visualizing the results of your analysis

This is an end-to-end data analysis challenge and will require you to make use of everything you have learned about data preparation, feature selection, and classification. This challenge also leaves you a great deal of room to improve upon what you have learned, as well as make use of tools or techniques we have not yet covered in class.

The assessment objectives of this final project are to measure how well you:

* Form problem statements / Hypotheses
* Design experiments to test these hypotheses
* Derive creative features or methods to gain access to latent information in the data
* Record your experiments and the exact parameters/methods used
* Record the results of your experiments
* Interpret and connect the pattern of experiment results
* Develop a narrative elucidating what you have discovered
* Identify actionable information obtained
* Identify insights that may lead to actionable information, as well as a description of what that actionable information may ultimately be

In addition to evaluation, this provides you with an opportunity to become more familiar with running end-to-end analysis projects. It is essential to understand how each of the phases of the project affects each other.

You are free to utilize any computational libraries you feel comfortable with to execute this task, as long as you can provide a detailed explanation of why you are using the package and exactly what it is doing. The grading criterion not only considers overall performance, but also the tools used to obtain those results (fancier tools will carry the expectation of better performance), logical thinking, reasoning, creative problem solving, diligence, eloquent explanation, and most importantly error analysis for each iteration.

**You are required to use at least one of these D3 visualization libraries:**

http://dpoetry.com/theplains/Hierarchie-gh-pages/

http://bl.ocks.org/mbostock/4339607

http://bl.ocks.org/mbostock/4063269

http://bl.ocks.org/mbostock/7607535

http://bl.ocks.org/mbostock/972398

http://bl.ocks.org/mbostock/1046712

http://bl.ocks.org/mbostock/4341134

You are free to utilize additional visualizations beyond what is listed for a more impressive result.

Note that you are not required to create a non-technical audience presentation in this assignment. You are however required to document ALL of your work in Jupyter or iPython, such that your results can easily be reproduced. Your record keeping, organization, and explanation in the lab notebook is a large and critical part of your grade.

**Minimum deliverables:**

Source code of your pubmed topic detection and visualization solution

README.txt if more than one source file

Library files for any external libraries used

Jupyter/iPython Analytic lab notebook

**Data:**

You will work with PMC Open Access Subset. Raw text files for the data can be downloaded at:

ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/articles.A-B.tar.gz

ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/articles.C-H.tar.gz

ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/articles.I-N.tar.gz

ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/articles.O-Z.tar.gz

XML annotated versions of these files that contain some metadata are available at:

ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/articles.txt.0-9A-B.tar.gz

ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/articles.txt.C-H.tar.gz

ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/articles.txt.I-N.tar.gz

ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/articles.txt.O-Z.tar.gz

**Due Date:**

December 18, 2015

**Hints**:

* Think through all the combinations of experiments you want to perform and consider the most logical way to organize the results for inclusion in your lab notebook, and error analysis.
* Remember that you also need to try to explain why a model is performing well, not just show good performance.
* Be sure to highlight the edge cases where classification is difficult and show examples to illustrate the specific causes of the difficulty.
* In this task you will be working with images. Remember that you need to convert the image into a set of labeled features for learning. Think through different strategies for converting the pixel matrix of the image into a vector.

**Readings and Resources:**

*Jupyter/iPython:*

http://jupyter.org

http://ipython.org

*Topic Detection:*

http://stackoverflow.com/questions/16831532/text-clustering-and-topic-extraction

https://www.quora.com/What-are-good-tools-to-extract-key-words-and-or-topics-tags-from-a-random-paragraph-of-text

http://www.markhneedham.com/blog/2015/03/05/python-scikit-learnlda-extracting-topics-from-qcon-talk-abstracts/

https://radimrehurek.com/gensim/

https://gist.github.com/alexbowe/879414

http://opendata.stackexchange.com/questions/1518/tool-to-extract-the-main-concepts-topics-from-web-pages

https://www.cs.princeton.edu/~blei/papers/Blei2012.pdf

*Flare.json:*

http://flare.prefuse.org

https://groups.google.com/forum/?fromgroups=#!topic/d3-js/L3UeeUnNHO8/discussion

http://www.delimited.io/blog/2013/11/2/creating-nested-json-for-d3

http://schoolofdata.org/2013/09/02/an-introduction-to-cargo-culting-d3-visualizations/

http://stackoverflow.com/questions/20015857/d3-tree-layout-from-flare-json-how-to-include-only-some-of-nodes

*D3 Visualization:*

http://d3js.org

https://github.com/mbostock/d3/wiki/Tutorials

https://github.com/mbostock/d3/wiki/Arrays#-nest

*PMC Open Access Subset:*

http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/