IS620 Week 4: High level plan

Alex Satz

Sept 25 2016

## APIs

<http://www.alexhadik.com/blog/2014/6/12/create-pubmed-citations-automatically-using-pubmed-api>

Or <http://dev.elsevier.com/api_key_settings.html>

## Goal

Generate a set of nodes (Author names, affiliation) with edges (coauthorship) and weights (# of coauthorships). Assign to each node a gender by matching firstname to a list of ‘male’ names. Determine centrality measures for the nodes, and find average centrality measures based upon gender and affiliation.

## Data acquisition

1. The data would be generated by starting with a single research topic, for instance ‘crisper’, using the api.
2. The first author is then taken from the first paper from the list of #1 above.
3. The author of #2 is given a node consisting of their full name and affiliation.
4. Another search is done, this time of the author name of the node in #3 and the search term ‘crisper’.
5. Each article returned in #4 is iterated over and for each co-author the number of co-authorships counted. For each co-author another node is created, with an edge equal to the number of coauthored publications.
6. We then return to the list of #1, and move onto the second author of the first paper, and repeat steps 3-5. When all the authors on the first paper of list #1 have been iterated over, we move onto the second article of list #1. In this way we create nodes for all listed authors, and weighted edges between all authors.
7. Lastly, we match author names to a list of typical ‘male’ names, to assign gender.

With the above data, we can determine which gender and institutions have played a dominant role in crisper research. We can also determine if these institutions have worked together or not. The research topic investigated has not been determined. I would prefer a new topic, with manageable number of papers such that the limits of the api searches are not reached.