Project Sample Resources for CS512 – Spring 2016

By James Abello

This is a sample list of resources (by no means exhaustive) that you could use for some of your projects.

Reference Papers

Ask Graph View, J. Abello, F. Van Ham;

CGV: C. Tominski, J. Abello,

Computational Folkloristics, J. Abello, T.Tangherlini

Exploratory Search: from finding to understanding, CACM, 49(4):41-46, 2006, G. Marchionini.

Sample Videos

https://vimeo.com/113233823

https://www.youtube.com/watch?v=BYzy0j5Z9Bo&feature=youtu.be

Sample Software: D3, Tulip, Gephi, Graph Stream, Semavis.net/dblp/#

Sample Data Sets

Some data sets that you may consider include: data feeds from Tweeter, YouTube, news streams, stocks, joke collections, movies, songs, online encyclopedias (Ex: OEIS, Algorithm and Software repositories), transportation schedules, WordNet, Motion Capture Data, data analytics blogs, funding agencies, startups, computer science educational materials, internet of things, …..

The following list consists of "curated" data sets that are very close to ready for use in algorithmic data exploration projects.

SNAP data Sets: Stanford Large Network Data Set Collection J. Leskovec and A. Krevl http://snap.standford.edu/data, June 2014

Patent citation network: https://snap.stanford.edu/data/cit-Patents.html

Global Media Monitoring Marko Grobelnik (email: Marko.Grobelnik@ijs.si)

http://eventregistry.org

Stream Access: http://newsfeed.ijs.si/stream/

Python Scripts: http://newsfeed.ijs.si/http2fs.py

Document Enriching http://enrycher.ijs.si/ **Cross-Linguality (Cross lingual similarity) http://xling.ijs.si**

Dmoz: The biggest taxonomy on the web;

Enron Email data: **http://www.cs.cmu.edu/~enron/** Used in **http://eliassi.org/papers/abello-icdm10.pdf**

LBL IP communication data: http://eliassi.org/data/lbl-20041215-1142.tar.gz
Used in http://eliassi.org/papers/abello-icdm10.pdf

*This needs special pwd to get access. Ask me if you want to use this data set.