

CS560 Knowledge Discovery and Management

Midterm: Hackathon

Submission Deadline: April 6 (S), Midnight

I. Submission

- 1) **Source code submission** to CS560 blackboard/Exam/MidtermSourceCode. Submit a word document file including all source codes (IMPORTANT: just in case you use someone's code and include the reference – URL)
- 2) **GitHub submission:** Submit your report and source package (source code, library, date, etc) to your GitHub site and include the GitHub link to your report
- 3) **Report submission to CS560 blackboard/Exam/MidtermReport.** Submit your report (word document). Check out the report guideline below.

II. Overview

The KDM hackathon aims to explore students to hack on anything that involves big data, processes big data, or produces big data and to do some interesting and innovative things. This hackathon focuses on the real time big data integration and aggregations by connecting different sources and providing high quality recommendation or customized mobile apps. This will foster opportunities to learn and contribute big data research in service computing community grow.

Each student is expected to output your “hacked” deliverables as follows:

- Publish cloud-delivered big data analytic services.
- Design new algorithms or models to integrate data from different sources
- Develop mobile/web apps that leverage big data.

Some optional deliverables are include:

- Optimize performance and scalability
- Visualize big data for better user experiences.

At the end of the hackathon, a couple of students will be selected/recognized for a demo and presentation about their project. We expect that you could benefit from this hackathon:

- Embrace and learn more about big data.
- Share your ideas with the people attending.
- Find out about interesting apps being created, and get to play around with them as well.

We are problem solvers, let's make it work and have fun!

III. Design & Implementation

You have the freedom to choose, to make your own choices for your application domain, datasets, and big data analytics technologies (e.g., Cloudera/Solr/Mahout/R).

1. Selection of datasets from at least two different domains (e.g., twitter and Youtube)
2. Big data integration and aggregations models and algorithms
3. Predictive/recommendation models and algorithms
4. Mobile App/Web design

You can use any open source projects or available APIs/Web services. But you need to cite them properly both in report and code. You are not allowed to discuss and share your solutions with other students. Your source code should be deployed to your GitHub site.

Solr upload Service:

134.193.136.127:8080/SolrWS/jaxrs/generic/uploadJson2Solr//home/cloudera/Desktop/test6.json

You need to specify your absolute file directory as parameter in the URL. The file may be located in your local Cloudera VM. The file must be in JSON format.

Solr Query Example:

If you want to retrieve all information related to id "Television", here is the url for invoking the service:

http://134.193.136.127:8983/solr/collection1_shard1_replica1/select?q=id%3ATelevision&wt=json&indent=true

IV. Examples

We list some examples of big data mashup applications that aim big data integration and aggregations by connecting different sources and providing high quality recommendation or customized mobile apps.

1) MusicMind

This machine learning music recommendation engine operates on images of album covers. When a user uploads a collage of their favorite albums, it will recognize and suggest similar music on the basis of the images alone. It can also tell you how expected or unexpected your particular music tastes are so the next time someone calls you a hipster you can refute them with scientific proof!

2) See Click Predict Fix

To rank the people's weighted perception about the importance of various issues based on a given dataset. The dataset should be rich and consist of a variety of common environmental factors such as:

1. Description of task/issue
2. Tag of Task/Issue
3. Source of Information
4. Time of issue/task creation
5. Location of issue/task creation

For this Hackathon, we attempt to utilize common machine learning libraries in order build a predictive model which is based on multiple datasets from diverse domains.

3) Attire Decider

This app is intended to simplify the way you get dressed in the morning. We retrieve information about your current location, the current temperature, humidity, chance of precipitation, wind, etc., and determine what articles of clothing and accessories you should wear to be most comfortable throughout the entire day. Even better, we customize such decisions of clothing based off of your gender.

4) Safe Neighborhood

- How can we make cities more equitable, more participatory, more open using IoT + Urban Data + Data Science?
- How can we use public urban data to improve citizen's urban experience?
- How can we remove barriers to accessing & distributing urban sensor, device data?
- How can we combine IoT + Data Science to provide equitable insight to citizens and councils?

Develop a predictive model, data visualization, or proto app on any of the following areas:

- Smart meters, energy reduction and carbon footprint reduction
- Crime & neighborhoods
- Urban flows, pedestrian flows, traffic flows, urban sensors
- Clean Streets, quality of our streets, lighting, and impacts on streets
- Antisocial behavior, noise complaints
- Impacts of licensing and planning on streets, and impacts on social environment

V. Report Guideline

1. Design
 - Data model (features)
 - Integration model and algorithm
 - Predictive/recommendation model and algorithm
 - Selection of datasets
 - Mobile App/Web design
2. Features Implemented
 - Integration algorithms
 - Predictive algorithms
 - Solr indexing: Your own data services
 - Software Services
 1. Your own services
 2. Existing services/APIs
 - Mobile User Interface
 1. Your own design
 2. Existing design
3. Outputs: description with screenshots of the Features
4. All the Web Service and Web Site URLs
5. Your midterm Github URL
6. Limitations
7. References

VI. Datasets

Here is a list of datasets you can use for this hackathon. You can use other external datasets as long as you use the ones provided.

- Public Data Sets on AWS: <http://aws.amazon.com/publicdatasets/>
- Yahoo Big Data Sets: <http://m.cnbc.com/bigdata/>
- U.S. Government's open data <http://www.data.gov/>
- DataGov Interactive Platform <https://explore.data.gov/>
- Infochimps datasets <http://www.infochimps.com/datasets>
- Data repositories http://oad.simmons.edu/oadwiki/Data_repositories
- IBM academic initiative University of Arkansas datasets <http://enterprise.waltoncollege.uark.edu/IBM.asp>
- City Forward Datasets http://cityforward.org/wps/wcm/connect/cityforward_en_us/city+forward/about/data+catalog/data+catalog+table+content
- City of Chicago Data Portal Datasets <https://data.cityofchicago.org/>
- University of California Machine Learning Datasets <http://archive.ics.uci.edu/ml/>
- Open Datasets <http://www.data.gov/opendatasites>

Image Datasets

- ImageNet <http://www.image-net.org/>
- 80 Million Tiny image dataset (from Antonio Torralba's group) <http://groups.csail.mit.edu/vision/TinyImages/>
- CIFAR-10 <http://www.cs.toronto.edu/~kriz/cifar.html>

Text Datasets:

- Open web directory <http://www.dmoz.org/>
- Wikipedia http://en.wikipedia.org/wiki/Wikipedia:Database_download

Network Datasets and collections:

- A collection of moderately-sized to larger datasets <http://graphlab.org/downloads/datasets>

- Twitter 2010 graph with 1.4B edges <http://bickson.blogspot.com/2012/03/interesting-twitter-dataset.html>
- A collection of moderately sized datasets <http://socialcomputing.asu.edu/pages/datasets?att=numAttributes&order=ASC>
- Trec datasets <http://trec.nist.gov/data.html>
- KDD cup datasets <http://www.kddcup2012.org/>
- Click-through rate dataset from HW1 <http://www.kddcup2012.org/c/kddcup2012-track2>
- Lemur project <http://lemurproject.org/clueweb09/>
- The Graph 500 Benchmark <http://graph500.org/>
- **Data repositories** <http://www.kdnuggets.com/datasets/>
- AWS (Amazon Web Services) Public Data Sets, provides a centralized repository of public data sets that can be seamlessly integrated into AWS cloud-based applications. <http://aws.amazon.com/publicdatasets/>
- BigML big list of public data sources. <http://blog.bigml.com/2013/02/28/data-data-data-thousands-of-public-data-sources/#comment-7538>
- Bioassay data, described in *Virtual screening of bioassay data*, by Amanda Schierz, J. of Cheminformatics, with 21 Bioassay datasets (Active / Inactive compounds) available for download. <http://www.jcheminf.com/content/pdf/1758-2946-1-21.pdf>
- Bitly 1.usa.gov data, anonymized clicks on gov links. <http://www.usa.gov/About/developer-resources/1usagov.shtml>
- Canada Open Data, pilot project with many government and geospatial datasets <http://www.data.gc.ca/>
- Causality Workbench data repository. <http://www.causality.inf.ethz.ch/repository.php>
- Corral Big Data repository at Texas Advanced Computing Center, supporting data-centric science. <http://www.tacc.utexas.edu/resources/data-storage/#corral>
- Data Source Handbook, A Guide to Public Data, by Pete Warden, O'Reilly (Jan 2011) <http://shop.oreilly.com/product/0636920018254.do>
- Datacatalogs.org, open government data from US, EU, Canada, CKAN, and more. <http://datacatalogs.org/>
- Data.gov.uk, publicly available data from UK (also London datastore.) <http://data.gov.uk/>
- Data.gov/Education, central guide for education data resources including high-value data sets, data visualization tools, resources for the classroom, applications created from open data and more <http://www.data.gov/education>
- DataMarket, visualize the world's economy, societies, nature, and industries, with 100 million time series from UN, World Bank, Eurostat and other important data providers. <http://datamarket.com/>
- Datamob, public data put to good use. <http://datamob.org/>
- DataSF.org, a clearinghouse of datasets available from the City & County of San Francisco, CA. <http://datasf.org/>
- DataFerrett, a data mining tool that accesses and manipulates TheDataWeb, a collection of many on-line US Government datasets. <http://dataferrett.census.gov/>
- Delve, Data for Evaluating Learning in Valid Experiments <http://www.cs.toronto.edu/~delve>
- EconData, thousands of economic time series, produced by a number of US Government agencies. <http://inforumweb.umd.edu/econdata/econdata.html>
- Enron Email Dataset, data from about 150 users, mostly senior management of Enron. <http://www.cs.cmu.edu/~enron/>
- Europeana Data, contains open metadata on 20 million texts, images, videos and sounds gathered by Europeana - the trusted and comprehensive resource for European cultural heritage content. <http://data.europeana.eu/>
- FEDSTATS, a comprehensive source of US statistics and more <http://www.fedstats.gov/>
- FIMI repository for frequent itemset mining, implementations and datasets <http://fimi.cs.helsinki.fi/>

- Financial Data Finder at OSU, a large catalog of financial data sets. <http://fisher.osu.edu/fin/fdf/osudata.htm>
- GDELT: The Global Data on Events, Location and Tone, described by Guardian as "a big data history of life, the universe and everything. <http://www.guardian.co.uk/news/datablog/2013/apr/12/gdelt-global-database-events-location>
- GEO (GEO Gene Expression Omnibus), a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval. <http://www.ncbi.nlm.nih.gov/geo/>
- GeoDa Center, geographical and spatial data. <http://geodacenter.asu.edu/datalist/>
- Google ngrams datasets, text from millions of books scanned by Google. <http://ngrams.googlelabs.com/datasets>
- Grain Market Research, financial data including stocks, futures, etc. <http://www.grainmarketresearch.com/>
- Hilary Mason research-quality Big Data sets collection - many text and image datasets. <https://bitly.com/bundles/hmason/1>
- ICWSM-2009 dataset contains 44 million blog posts made between August 1st and October 1st, 2008 <http://www.icwsn.org/2009/data/>
- Infochimps, an open catalog and marketplace for data. You can share, sell, curate, and download data about anything and everything. <http://infochimps.org/>
- Investor Links, includes financial data <http://www.investorlinks.com/>
- KDD Cup center, with all data, tasks, and results <http://www.sigkdd.org/kddcup/index.php>
- Kevin Chai list of datasets, for text, SNA, and other fields. <http://kevinchai.net/datasets/>
- KONECT, the Koblenz Network Collection, with large network datasets of all types in order to perform research in the area of network mining. <http://konect.uni-koblenz.de/>
- Linking Open Data project, at making data freely available to everyone. <http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>
- Million Song Dataset <http://labrosa.ee.columbia.edu/millionsong/>
- MIT Cancer Genomics gene expression datasets and publications, from MIT Whitehead Center for Genome Research. <http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi>
- ML Data, the data repository of the EU Pascal2 networks. <http://mldata.org/>
- NASDAQ Data Store, provides access to market data <https://data.nasdaq.com/>
- National Government Statistical Web Sites, data, reports, statistical yearbooks, press releases, and more from about 70 web sites, including countries from Africa, Europe, Asia, and Latin America <http://www.archive-it.org/>
- National Space Science Data Center (NSSDC), NASA data sets from planetary exploration, space and solar physics, life sciences, astrophysics, and more <http://nssdc.gsfc.nasa.gov/>
- Open Data Census, assesses the state of open data around the world. <http://census.okfn.org/>
- OpenData from Socrata, access to over 10,000 datasets including business, education, government, and fun. <http://opendata.socrata.com/>
- Open Source Sports, many sports databases, including Baseball, Football, Basketball, and Hockey <http://www.opensourcesports.com/>
- Peter Skomoroch dataset Bookmarks <http://www.delicious.com/pskomoroch/dataset>
- PubGene(TM) Gene Database and Tools, genomic-related publications database <http://www.pubgene.org/>
- Quandl, a collaboratively curated portal to millions of financial and economic time-series datasets. <http://www.quandl.com/>
- qunb, a platform to find and visualize quantitative data. <http://www.qunb.com/>
- Robert Schiller data on housing, stock market, and more from his book *Irrational Exuberance*. <http://www.econ.yale.edu/~shiller/data.htm>

- SMD: Stanford Microarray Database, stores raw and normalized data from microarray experiments. <http://genome-www5.stanford.edu/MicroArray/SMD/>
- Jerry Smith dataset collection, with Finance, Government, Machine Learning, Science, and other data. <http://datascientistsinsights.com/2013/02/02/data-monetization-road-paved-on-top-of-data-sets/>
- SourceForge.net Research Data, includes historic and status statistics on approximately 100,000 projects and over 1 million registered users' activities at the project management web site. <http://www.nd.edu/~oss/Data/data.html>
- StatLib, CMU Datasets Archives <http://lib.stat.cmu.edu/datasets/>
- STATOO Datasets part 1 and STATOO Datasets part 2
- <http://www.statoo.com/en/resources/anthill/Datamining/Data/>
- http://www.statoo.com/en/resources/anthill/Data_Sets/
- Time Series Data Library <http://robjhyndman.com/TSDL/>
- Visual Analytics Benchmark Repository. <http://kdd.ics.uci.edu/>
- UCI KDD Database Repository for large datasets used in machine learning and knowledge discovery research. <http://archive.ics.uci.edu/ml/>
- UCI Machine Learning Repository. <http://archive.ics.uci.edu/ml/>
- UCR Time Series Data Archive, offering datasets, papers, links, and code. http://www.cs.ucr.edu/~eamonn/time_series_data/
- United States Census Bureau <http://www.census.gov/>
- Wikiposit, a (virtual) amalgamation of (mostly financial) data from many different sites, allowing users to merge data from different sources <http://wikiposit.org/>
- Wolfram Alpha disease and patient level data <http://blog.wolframalpha.com/2010/06/29/disease-and-patient-level-statistics-with-wolframalpha/>
- Yahoo Sandbox datasets, Language, Graph, Ratings, Advertising and Marketing, Competition <http://webscope.sandbox.yahoo.com/catalog.php>
- Yelp Academic Dataset, all the data and reviews of the 250 closest businesses for 30 universities for students and academics to explore and research. http://www.yelp.com/academic_dataset

VII. Useful Sites for web services/APIs

- Foursquare website <https://foursquare.com/>
- Foursquare API <https://developer.foursquare.com/start>
- Google Earth Vehicle Simulators <http://www.gelib.com/simulators.htm>
- Text-to-speech http://www.oddcast.com/home/demos/tts/tts_example.php
- Google Developer Site <https://developers.google.com/>
- Amazon Web Service API <http://docs.aws.amazon.com/AWSECommerceService/latest/DG/Welcome.html>
- Yahoo Developer Site <http://developer.yahoo.com/>
- Twitter Developer Site <https://dev.twitter.com/>
- Facebook Developer Site <https://developers.facebook.com/>