MSAN 692: Data Acquisition Project Brief

Matthew Dixon

Due: 11:59pm September 27th, 2013

Overview

The data acquisition project is a team project with the goal of evaluating the data acquisition workflow for the purpose of analytics. Your goal is to simply pose a high level problem or hypothesis that you seek to address and then evaluate the availability of data and the challenges and limitations with data acquisition. Please note that you are not expected to actually perform the analytics - this is beyond the scope of the course. However, you may need to carefully think ahead and try to pre-empt some of the data challenges that you are likely to encounter.

Deliverables

Please turn in a 1 to 3 page group report intended for a non-technical person to understand (e.g. a business decision maker) to evaluate the data acquisition challenges in pursuing your analytics problem. You can assume that whoever will be reading the report determines whether your proposed analytics project will be funded.

Your report should address the following questions and items:

- What problem or hypothesis are you solving? Please state in non-technical terms and provide a short context to the problem.
- Summarize the main findings of your data acquisition project in a table or other suitable format.
- What is your recommendation for pursuing the problem further?
- What data acquisition experiments and evaluations did you perform?
- What challenges and issues did you encounter?
- How did you address this challenge, or how would you recommend addressing it?
- What assumptions did you make about the data and what are some of the main limitations of your evaluation?
- Include any useful URLs to data sources and data content sites that you found.

General Advice

- Remember that this is a data acquisition project which means that the focus is primarily on using the skills that you've learnt in this course.

- Try not to get lost in the details - it's important that you summarize the challenges in data acquisition rather than provide a set of instructions on how to access and clean the data.
- 'More is less' and 'content defines form' - conciseness is key (without being too vague) and the content of your report is more important than elegance, aesthetics and flashy designer logos etc.
- Try to make your insight actionable - how could someone who hasn't taken this course evaluate and utilize your findings?
- For this project, you can assume that the reader understands sparing use of basic terminology such as REST, but avoid excessive use of acronyms and/or technical jargon.

Grading

This project will be awarded up to 60 marks and is equivalent to two weekly homework assignments. Your report will be graded on the merit of:

- The clarity of the problem statement/ hypothesis and your solution approach(es)
- How challenging the problem that you are addressing is
- How you addressed the problem using skills learnt in this course and how you overcame or managed data acquisition problems.
- The strength and clarity of your summary and recommendations

Example report

The following example is for illustrative purposes only - the nature of the illustrated project was considerably more complex and extensive than you are expected to address in the very limited time frame.

Problem Statement

David Thomson's 'Blueprint To a Billion' defines a number of characteristics, or 'domains of concerns' necessary for a company to grow to a billion dollars in valuation. We seek to evaluate whether it is possible to identify such characteristics in cleantech growth stage companies based on a number of public and private data sources.

Executive Summary

This report evaluates various public and private data sources for their potential utility in assessing the likelihood of a growth stage company ever reaching a billion dollar valuation. We find, as of June 2012, that Dow Jones VentureSource has the strongest coverage of domestic and international venture backed 'cleantech' companies with coverage of 67k companies and 19k active investors. However, we identified that only 3775 companies are strictly clean tech companies - energy efficiency and resource innovation. Out of these only 11% have sufficient data to forecast future growth using time series techniques in financial econometrics.

Recommendations

We find that the data is too limited to use standard time series methods for revenue and valuation forecasting. Instead we propose evaluating a company along the domains of concerns set forth by David Thomson. For each domain of concern, we have identified data sources most suited to model building and quantification of company strengths in each domain. We recommend, as a next step, more extensive data acquisition processes and evaluation of each of the sources identified. This would culminate in the insight necessary to build fundamental and statistical models to translate qualitative fundamental variables into quantitative features which can be used by machine learning classifiers to discern likely winners and losers.

Data Acquisition Experiments

We implemented a python data acquisition application to extract JSON and XML files from major data sources and store them in a database. This application was automatically run daily over four weeks to avoid breaching throttle limits and to overcome limitations on the maximum number of companies returned by RESTful API queries. We subsequently merged data sources by transforming company names into canonical form.

Data Challenges

RESTful services provide limitations on the number of companies that are returned per query thus limiting the depth of data extraction. Furthermore, many of the companies return by a query with the keyword 'cleantech' returned companies which are not considered cleantech for the purpose of this project. This led us to believe that we were not recovering many actual cleantech companies which were mis-labelled.

Solution Approaches

By constructing a contingency table of company sector tags, we were able to infer a hierarchy of sector tags which are most commonly associated with actual cleantech companies. We then searched by combinations of tags to build up a more comprehensive set of data. To overcome the throttle limit, we also separated the search by continent and, within N. America, additionally by state.

Assumptions and Limitations

We assumed that the industry sector tags associated with each company are accurate - however we since learned that this too is open to interpretation. In hindsight, a better approach would be to work with an expert analyst to define a meaningful industry sector taxonomy and then map the tags to our own tags. The results of any forecasting are likely to be sensitive to this mapping. Also, we are assuming that the data provider does not re-map their tags, thus introducing complications in future data collection.

Summary of Results

We evaluated nine data sources, the name and classification of each are provided in Table 1. These sources are either classified as crowd-sourced RESTful services, gated public data (i.e. subscription based) or unstructured web content. Their assignment to each domain of concern is based on detailed analysis beyond the scope of this report.

Additional references

There are also various industry websites and newsletters where data could be collected from, but they are disparate sources that are too costly to be aggregating information from. Examples of such websites include:

`http://info.cleantech.com/Cleantech-Newsletter.html`
`http://www.greentechmedia.com/`
`http://gigaom.com/cleantech/`