Recently voted “Largest Upside Potential” by a panel of Silicon Valley venture capitalists, tealbook is just getting started building our supplier knowledge platform. We are building a searchable database of the world’s suppliers with details of their capabilities and contact information. We are looking for a research intern to crawl new data sources, to extract information and to apply machine learning techniques to leverage this information. This is an important role at tealbook. Our data, and the information that we extract from it, enable our supplier discovery and recommendation features which are key differentiators of our solution.  
  
The goal of this internship is to enrich our supplier database with additional suppliers and better information about existing suppliers such as name, description, contact information, classification, keywords, and reputation. This will be a large dataset with hundreds of millions of suppliers. Web crawling will be used to discover suppliers and find information about suppliers. As input you'll be provided with a list of website URLs.   
  
Challenges come in two forms:  
  
a) To manage the crawlers, and more importantly the associated data, in a manner that supports data enrichment pipelines and experiments.   
  
b) To extract structured data from unstructured data sources (html pages). Some fields such as contact details can be extracted successfully with simple algorithms. Other fields such as keywords and classification code mappings require more powerful machine learning algorithms to extract values and crowdsourcing to measure performance of that information.  
  
Specific objectives:  
  
a) Setup accounts on the Google cloud platform and choose data storage and compute options.  
  
b) Define the data layout for the crawler data.  
  
c) Configure a crawler to pull an initial page from each supplier website and run the crawler.  
  
d) Use regular expressions and heuristics to extract initial field values from crawled content.  
  
e) Evaluate coverage and quality of data extracted.  
  
f) Crawl additional pages from supplier site to gather additional text, for example use a simple heuristic to look for the contact page.  
  
As time permits:

g) Investigate better extraction techniques for extracting structured data.  
  
h) Investigate categorizing suppliers by assigning industry standard classification codes.

i) Extract keywords and phrases from websites using word frequency and analyze html structure to identify important terms.  
  
j) Apply NLP to improve search term understanding or data extraction.  
  
Research literature and support:  
  
There is background literature on setting up large scale crawlers and how to organize the data layout to support a crawl that will be useful in object b). The crawlers themselves provide some of this structure. This is a well understood mature field with good documentation available.  
  
Open ended tasks g) h) i) and j) involve deeper research problems. Simple rules will provide a benchmark performance level for information extraction but how can you achieve better coverage or quality? Using tools like Google’s natural language API and machine learning toolkit (TensorFlow), more powerful algorithms can be explored.