Job Posting Crawler Text Analysis

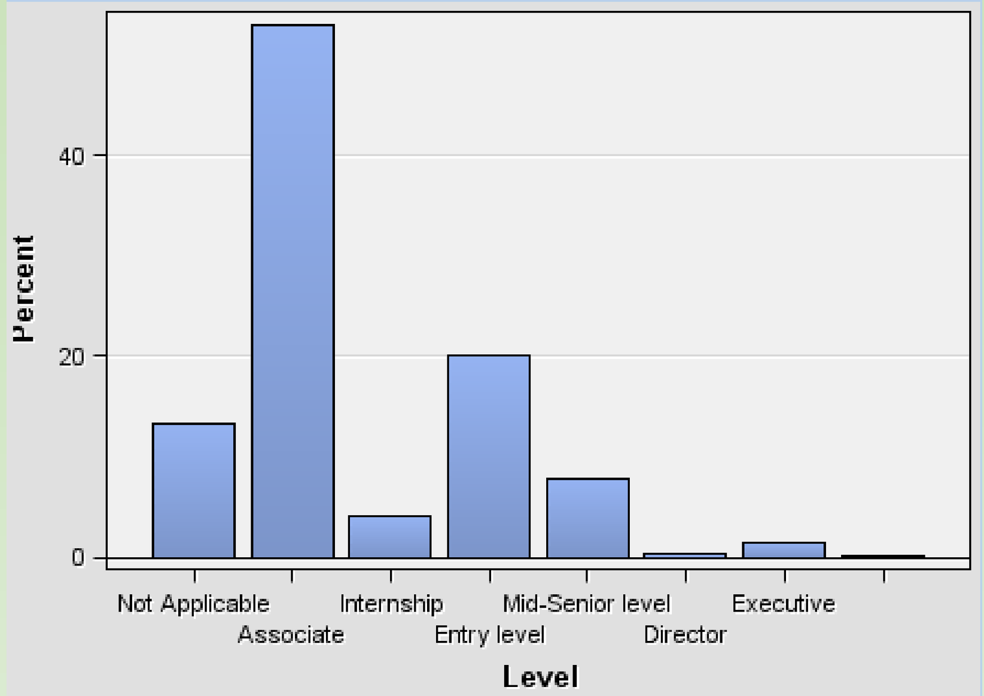
Tay Yiqin Timothy, 1982086I

Temasek Polytechnic, School of Informatics & IT – Specialist Diploma in Business Analytics

INTRODUCTION

People who aspire to enter an analyst profession may not know what employers are looking out for in their hires. In order to be informed about the right career choice, text analysis of job postings from popular sites like LinkedIn will help in providing insight into the main topics employers post as requirements, as such prospective hires can craft resumes more effectively, go for more targeted upskilling, or simply know what a job scope usually entails. We will aim to use SAS EM to conduct clustering & topic analysis on data scraped off LinkedIn postings around "Data Analyst in Singapore".

DATA EXPLORATION & PREPARATION

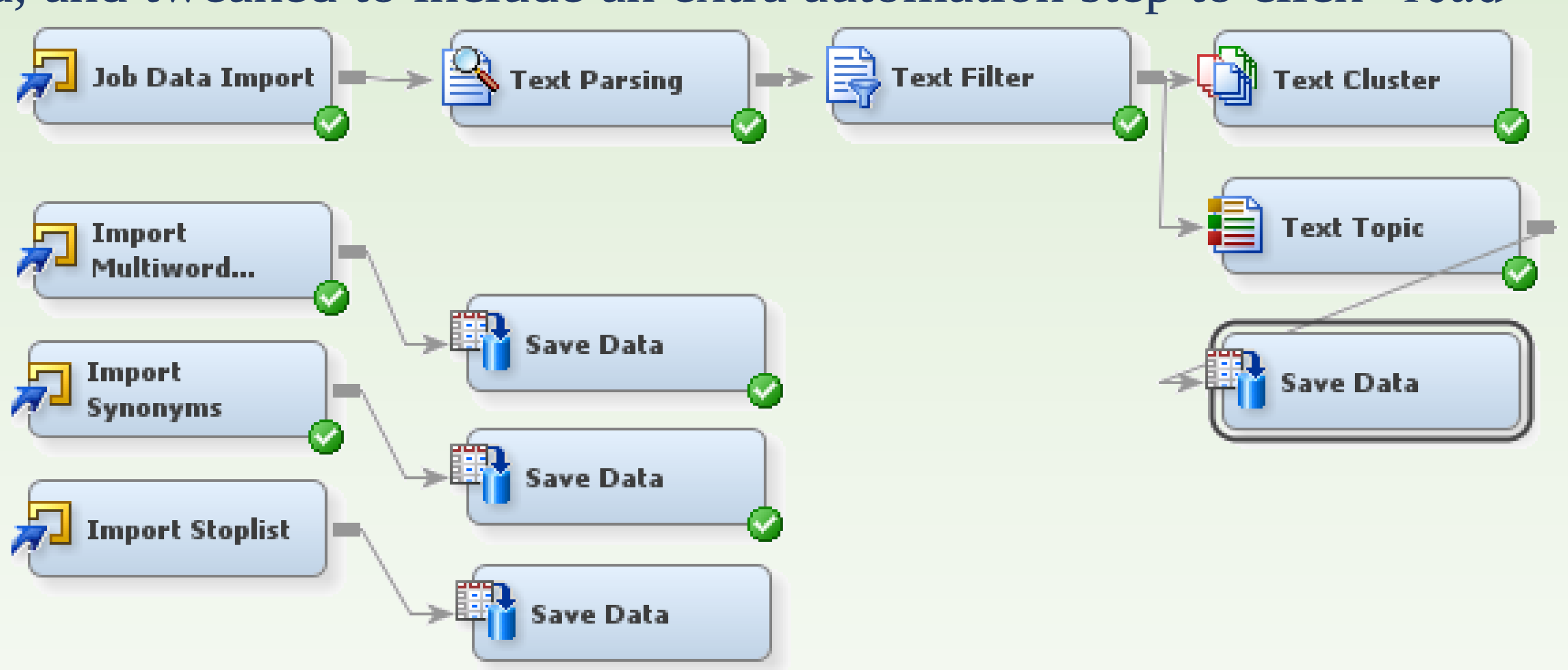


In order to scrape the data, we referenced a GitHub script to code a script in Python utilizing web browser automation (Selenium) and web text scraping libraries (BeautifulSoup) to scrape 1000 LinkedIn job postings around "Data Analyst in Singapore" retrieved on 15th August 2020, sorted by date posted. The results were placed into an excel file containing the description field, Job Posting ID & Posting Level which was ingested into SAS Enterprise Miner to form the basis of the text analysis.

The data of 1000 job postings consists of the following distribution: 53% Associate Level, 20% Entry level, 7% Mid senior level, 4% Internship with 13% unknown.

WORKFLOW

The LinkedIn scraper built by kirkhunter was used, and tweaked to include an extra automation step to click "read more" so the full description could be scraped.



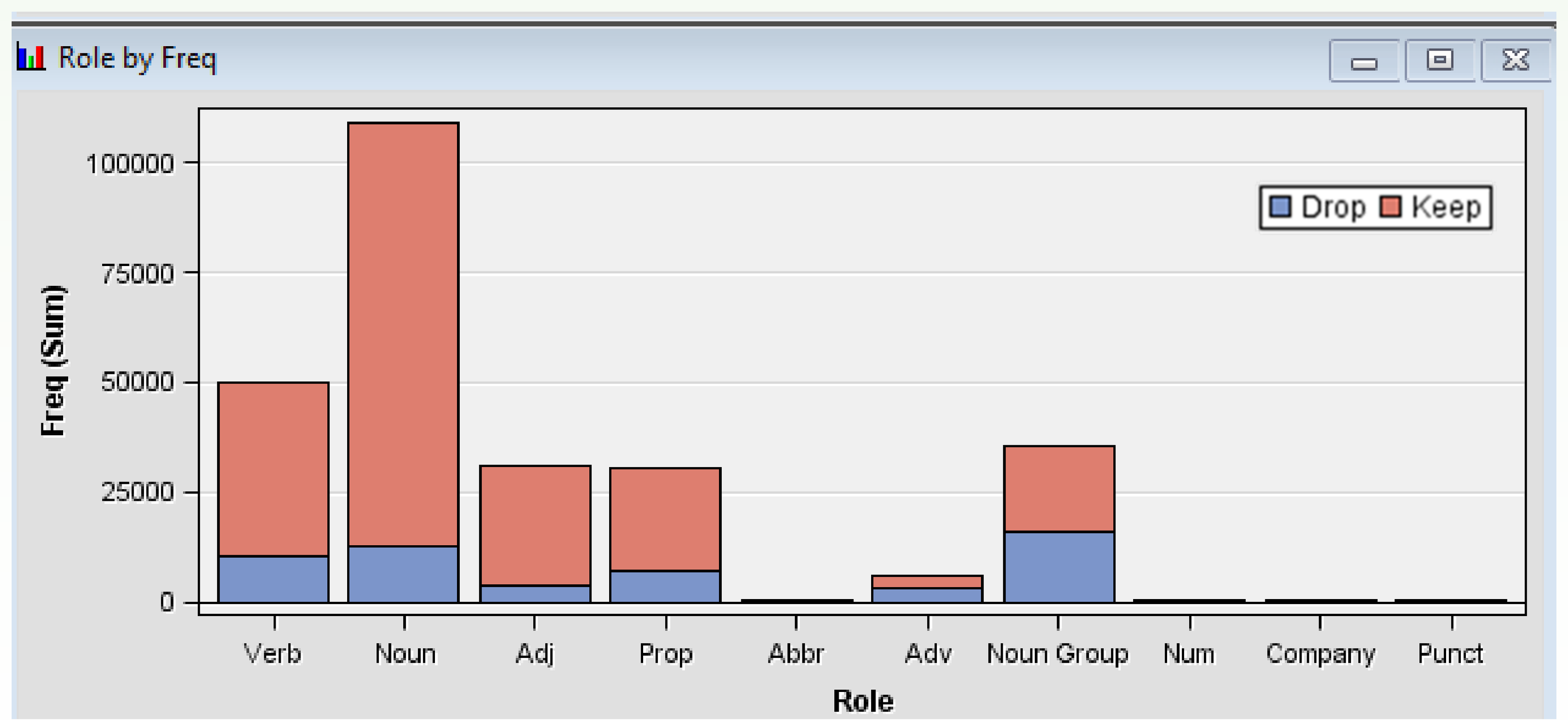
Import Node: The Description field was set to Text property & Job Posting ID set to ID.

Parsing Node: We used 189 multi-word terms, 10 synonyms & 10 stop terms with Stemming used. The stop terms include +be, +work, one, job responsibilities & skill-set, because they are common filler words in job postings. Multi-word terms include "business analysis" or "quantitive analysis" to capture the contextual difference between the terms.

Filter Node: Frequency Weight was set to Log, while Term Weight was set to Inverse Document Frequency because job postings are relatively long. After the parsing & filter node, roughly 79% of terms are kept.

TextCluster: Number of Clusters kept at 40, Probability used (Expectation Minimization).

TopicGenerator: Topics kept at 20, no correlated topics, no single-word topics.

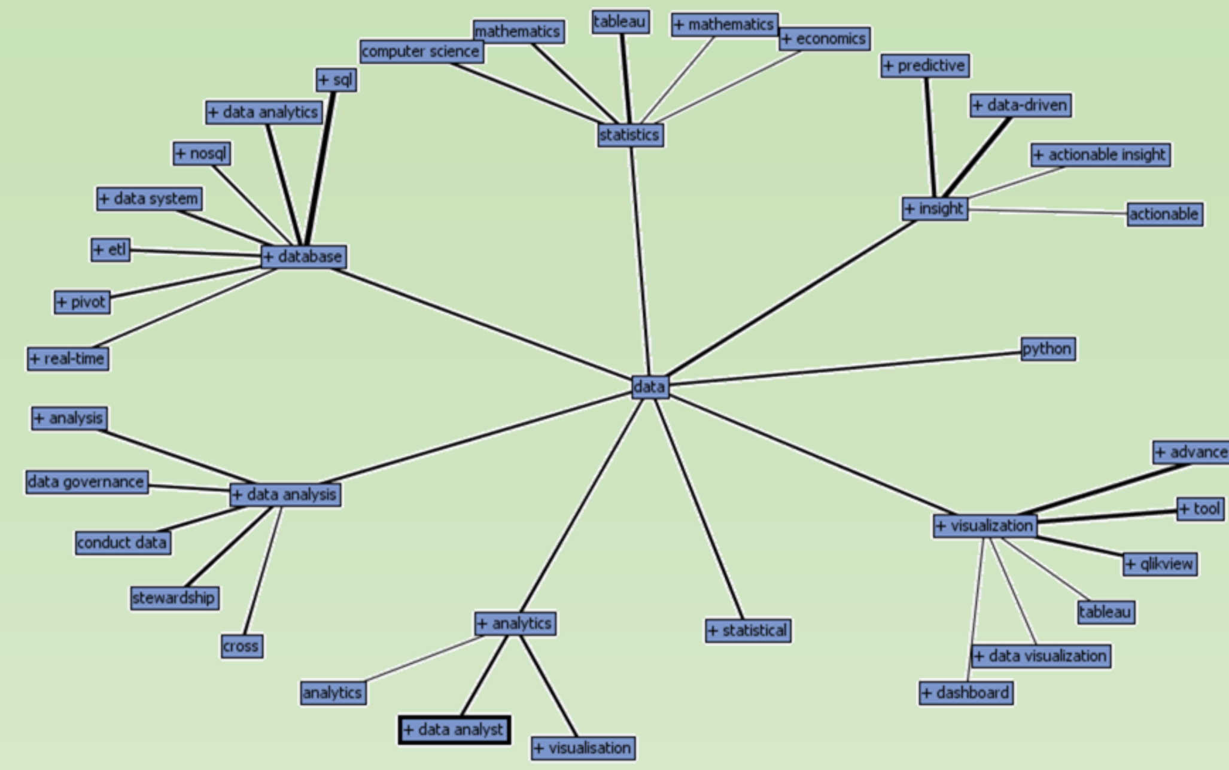


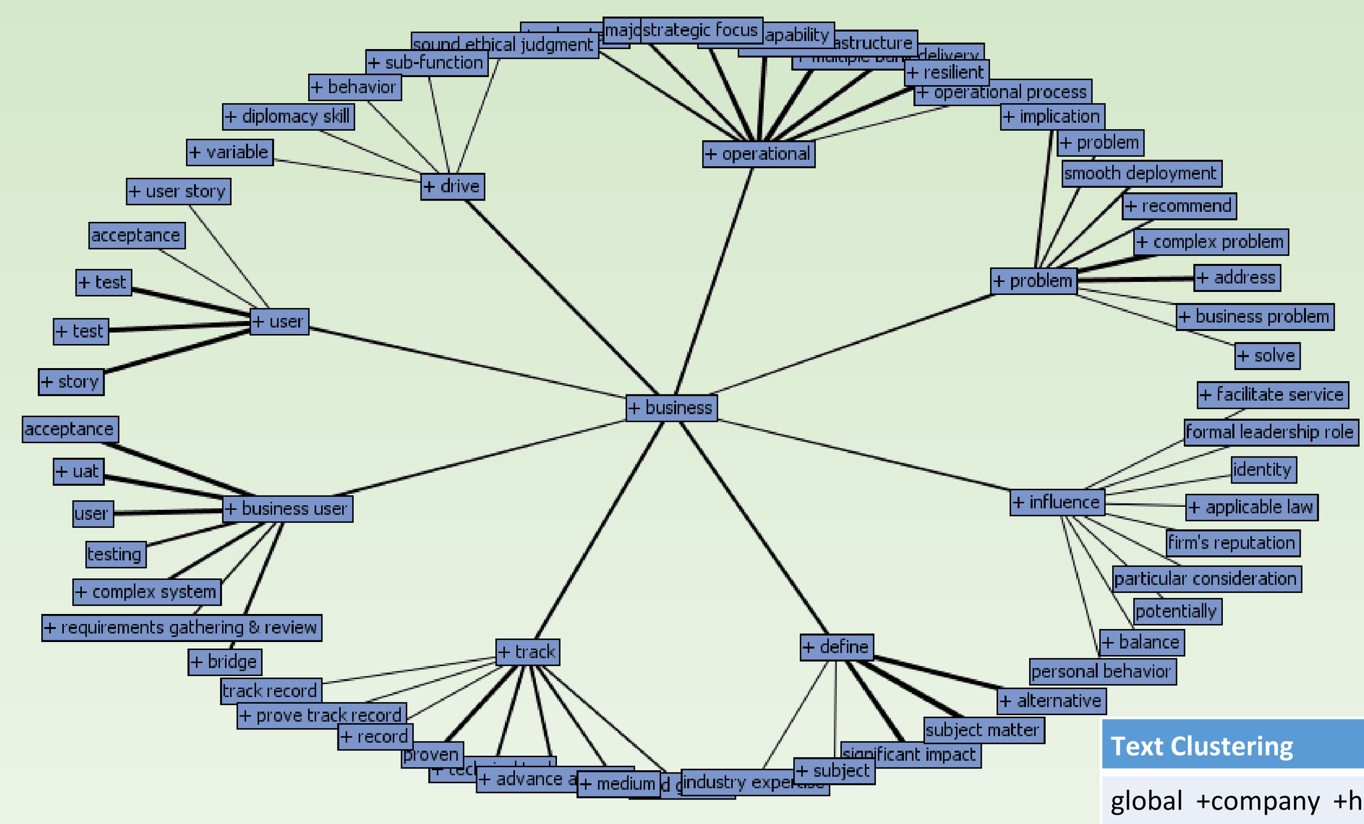
FINDINGS & RECOMMENDATIONS

The Interactive Filter Viewer was used to examine the terms. "data" was expected as "data analyst" was the search term used. However +business, +team, +requirement & +analysis were top terms, which were further expanded using the Concept Link tool.

Term	Freq	# Docs
data	3810	709
+business	2658	741
+team	1351	627
+requirement	1210	613
+analysis	1207	587
+project	1203	540
+system	1103	487
+solution	1081	511

Term	Freq	# Docs
+process	980	521
+knowledge	946	521
+ability	932	490
+management	911	484
+development	828	457
+technology	824	416
+product	807	398
+user	802	392





For postings with "data" (about 70%), major themes were around **database skills** like SQL, NoSQL & ETL; **data analysis topics** including data governance & stewardship; **visualization** involving technology like Tableau dashboards & Qlikview, & **ability to generate predictive, data-driven & actionable insights .**

For postings with "business" (about 70%), major themes were around 1) **Operations**, with a focus on process analysis & developing resilient infrastructure, 2) **Interfacing with the Business user**, involving UAT & Requirements Gathering of a complex system, 3) **a proven track record**, 4) able to **influence management**, and 5) **Problem Solving skills** to allow smooth deployment.

The Text Topic & Text Clustering node were run & results compared:

Text Clustering	Freq	%
global +company +help +market +opportunity highly +client +build +role +enjoy +individual +drive +singapore +commitment excellent	225	23%
+test +user acceptance functional +document +uat +documentation +'requirements gathering & review' 'business analyst' +implementation banking +technical +system +issue +solution	149	15%
python +insight +analytics +statistical +engineer tableau +model digital +sql +tool +dashboard data +trend +build +decision	114	11%
'computer science' +project +meet +test agile +implement +application +service +problem +system +development +good +manage +database +user	83	8%
ea +shortlisted +'shortlisted candidate' +resume +licence +candidate recruitment +singapore +bank +apply tableau +application analytical python +dashboard	78	8%

Topic	# Terms	% Docs
digital,+customer,+company,+sale,growth	651	15%
data,+report,+report,+dashboard,+analyze	549	12%
+user,+test,+business user,+uat,testing	385	12%
financial,+finance,+system,+murex,+office	527	10%
+project,+manage,+external,+deliver,+management	513	7%
+government,+application,ict,+influence stakeholder,conceptualisation	578	6%
pte,ea,ltd,personal data,+investment	377	6%
gender,+status,+origin,+national origin,+age	522	6%
+hadoop,+restriction,+visa,+ecosystem,python	373	5%
+click,+view,eeo,+career opportunity,citigroup	302	4%

The themes, which were auto-generated by SAS, complemented with the clusters, captures the following top topics in most job postings of "data analyst" in Singapore:

1) Involved with digital transformation, sales & growth of accounts;
2) Build BI dashboards & analyse data;
3) UAT Testing;
4) Finance Industry;
5) Project Management

To break into Data Analyst track, prospective job applicants can be aware that they can focus on upskilling through learning visualization skills, account management, project management, digital transformation or User Acceptance Testing frameworks. The Finance Industry has a sizeable number of analyst job postings which one can start from.

REFERENCES & CONTACT DETAILS

Kirkhunter's Python LinkedIn Scraper: https://github.com/kirkhunter/linkedin-jobs-scraper