Data Industry Job Analysis

By Seth Chart

TOC

Overview

The job market for the data industry can be difficult to navigate because there are a plethora of job titles and the relationship between titles and roles is often not well defined. Two identical roles may have completely different titles. Two substantially different roles may have the same title. This limits the usability of job titles as a means to succinctly communicate about data industry jobs. This project will address the issue by directly analyzing full job descriptions from a corpus of job postings to provide three main deliverables.

Problems to solve

1. Job titles do not effectively summarize job descriptions in the data industry. We need a better way to give job seekers an at-a-glance summary of a job posting.

2. Because there is not direct correspondence between job titles and job descriptions in the data industry, it is not clear how many distinct roles exist. We need a better way to classify data industry jobs.

3. Having claimed that job titles in the data industry are flawed and ineffective, we need to quantify how they perform as classifiers of data industry roles.

Project objective

Find a data driven supplement to job titles for summarizing and classifying jobs in the data industry.

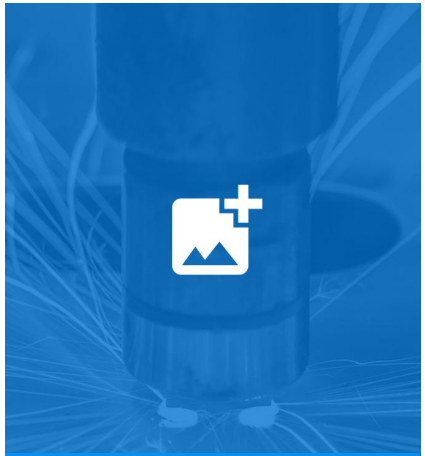
Understanding the market

Market trends

01

We used unsupervised learning to detect ten common topics within data industry job postings. Nine of these topics aligned well with requirements and responsibilities for data industry jobs. One topic captured boilerplate phrases.

Client Implications:
With this model, we could provide a job seeker with an at-a-glance ten number summary of a job posting that describes the prevalence of our ten topics within the job description.



Market trends

02

We used k-Means clustering to group job postings into ten job classes.

Client Implications:
These ten job classes represent ten distinct groups of jobs that are supported by data. These classes could be useful to job seekers as an even more succinct summary of a job. They could also help employers to more clearly organize their workforce.

Market trends

03

We showed that using job titles alone to predict job classes was only 55% accurate. This supports our hypothesis that job titles, in the data industry, are currently flawed and ineffective.

Client Implications:
This information demonstrates the need for better tools for summarizing job postings in the data industry. Low accuracy in job titles results in an inefficient job market where job seekers and employers are unable to connect.

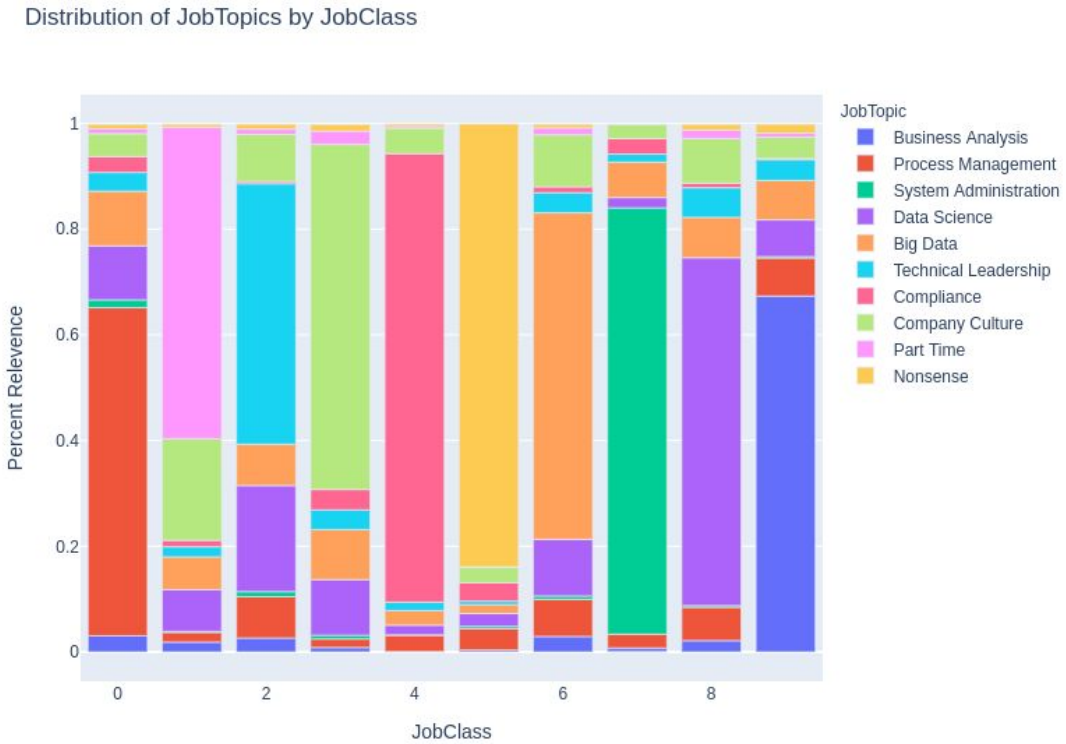
Trend analysis

To the right we see our ten JobClasses each displayed with their average JobTopic distribution.

JobClasses 0, 1, 2, 3, 6, 8 and 9 require a mixture skills.

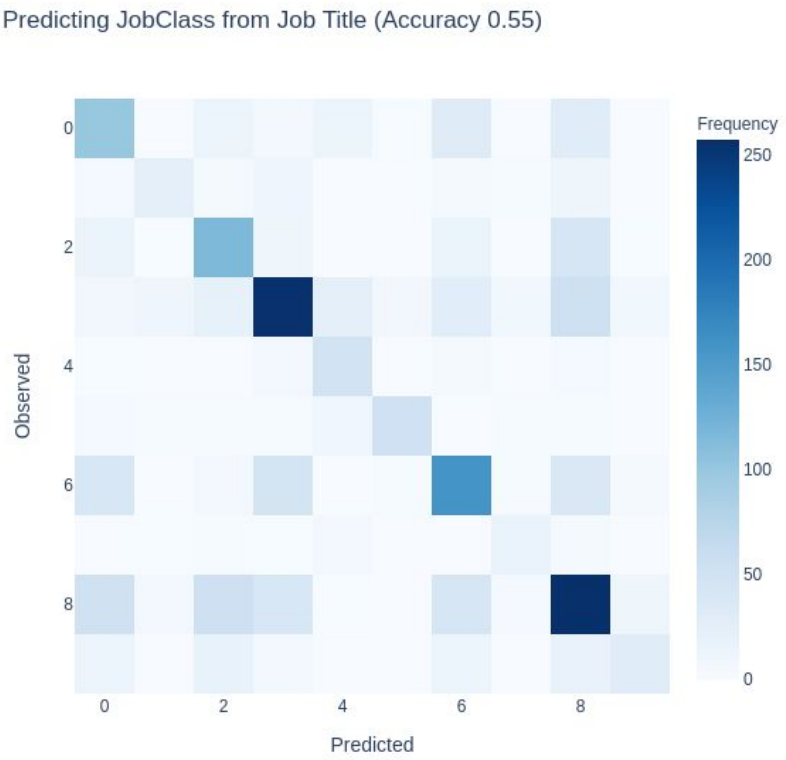
JobClasses 4 and 7 are fairly specialized.

JobClass 5 contains job descriptions without much meaningful content.



Trend analysis

To the right we see a confusion matrix for our predictions of JobClass using only job titles. We observed prediction accuracy of 55%.



Target audience

We wish to target data industry job seekers, employers, and job forums with a turnkey, data driven solution to supplement their existing job postings.

01 | Seekers connect to opportunities efficiently

02 | Employers receive more relevant applicants

03 | Job forums add value by improving alignment

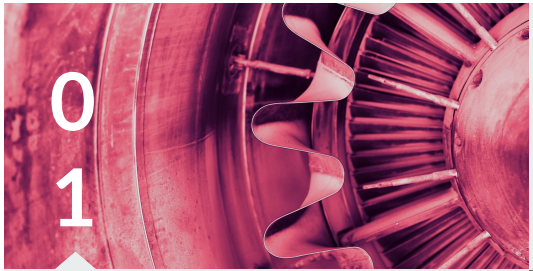
04 | Seekers get clearer picture of job market

05 | Employers can optimize postings and clarify roles

Proposed solution

A RESTful endpoint consumes raw job description text and returns both a ten number JobTopic summary and a job classification, providing a powerful quantitative supplement to job titles.

Process



Data collection and processing

Collect a large corpus of job postings from careerjet.com

Modeling

Use unsupervised learning techniques to extract topics and job classes from job descriptions. Use supervised learning to evaluate accuracy of job titles.

Deployment

Host trained models and provide an API to obtain topic distributions and job classification for a user provided job description.

Deliverables

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt labore dolore magna aliqua. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt labore dolore magna aliqua. Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Corpus of Job Postings

9.5K

Raw data provides the basis for our product.

High Quality ML Models

3

These models turn raw data into at-a-glance insight for job seekers and employers alike

Software as a Service Endpoints

2

Both TopicDistribution and JobClass provide turn key supplements to job titles

Vision

Initial Development

Data collection and modeling complete.

Refactor Codebase

With the initial data collection and modeling products in place, it is time to move code from the project notebook into modules and ensure that the code base is maintainable.

Launch API and Example Site

With a well designed and maintainable codebase, we are ready to deploy a public API and an example site to demonstrate functionality.

Design Product Lifecycle

Automate data collection, set model retraining schedule, and assess long-term model stability.

Product Maintenance

Ensure that product is available and reliable through regular maintenance and consideration of project scaling needs.

Project Lead

Seth Chart, PhD

Seth is a data scientist and mathematician living in the Baltimore area. Seth is currently seeking opportunities in Data Science.

- Website: sethchart.com
- Github: sethchart
- LinkedIn: sethchart
- Twitter: @sethchart
- Email: seth.chart@protonmail.com

Thank you.