Data workforce needs: Disambiguation of roles using clustering and topic modeling

Cheryl A. Thompson & Craig Willis

Center for Informatics Research in Science & Scholarship, Graduate School of Library & Information Science, University of Illinois at Urbana-Champaign

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois

Introduction

As the data workforce evolves, data curation educators need information about emerging professional roles. To date, the literature has been populated with imprecise and inconsistent names and definitions for data professionals (Cox & Corrall, 2013; Swan & Brown, 2008). Disambiguation of data roles and responsibilities is key to preparing well-trained graduates that can meet the workforce demands.

The project goals were to:

1) disambiguate data professional roles.
2) understand how learning algorithms can be used in workforce analysis.

Method

Data Source

- 15,120 unique job advertisements were harvested from Indeed.com using 46 queries issued to the Indeed API to capture a range of data-related positions.
- 430 ads were identified as not data/ information positions.

Analysis

- Clustering: Job ads were clustered using the *k*-means algorithm (k=50, cosine distance). Ads in each cluster were ranked by their distance from the centroid. The top 10 cases in each cluster were manually reviewed for cluster labelling.
- Topic models: Latent Dirichlet Allocation (LDA) was used to generate 200 unigram topics. The top 10 terms for each topic were used for topic labelling.
- Qualitative Analysis: For a sample of 30 ads, analytical codes were developed using both an inductive and deductive approach. A codebook was created with the final set of codes. Job ads were manually coded.

Acknowledgments

This work is supported by the Data Curation Education in Research Centers (DCERC) IMLS Award #RE-02-10-0004-10. The job ad corpus was provided by Indeed.com. We also would like to thank Dr. Catherine Blake for her assistance and guidance.

Example job ad with sample topics, assigned cluster, and qualitative codes

The terms in the top 10 LDA-assigned topics for this ad are highlighted below. This is intended only to illustrate the LDA-assigned topics in a document. Below is also the cluster and a selection of qualitative codes assigned to this ad.

Data Scientist, Booz Allen - Rockville, MD-01155460

Key Role: Work with cross-functional consulting teams to design, develop, and execute analytical solutions to derive business insights and solve operational and strategic problems of the client. Support the development of analytical models and tools that improve existing processes and decision making with the effective use of data and analytic techniques within the healthcare space. Support consulting teams in the design, development, and implementation of new operating processes and models for clients. Build internal team capabilities in analytics techniques and methods to better serve clients and demonstrate thought leadership. Contribute to business and market development for the government and commercial market.

Basic Qualifications:

- Experience with programming languages, including Python, Perl, or Java
- Experience with statistical analysis software, including SAS or R
- Knowledge of statistical or mathematical analysis techniques, including classification, regression, optimization, and simulation
- Ability to manipulate, integrate, and analyze large and complex data sets
- Ability to communicate technical concepts and solutions to clients and internal teams
- Ability to obtain a security clearance

Additional Qualifications:

- Experience with machine learning, text mining, or natural language processing
- Experience with health data sets, including electronic health records a plus
- Knowledge of Big Data and Cloud computing technologies, including Hadoop, No-SQL, or map reduce
- MA or MS degree in CS, Statistics, Epidemiology, Mathematics, Physics, or a related field

Clearance: Applicants selected will be subject to a security investigation and may need to meet eligibility requirements for access to classified information.

Integrating the full range of consulting capabilities, Booz Allen is the one firm that helps clients solve their toughest problems, working by their side to help them achieve their missions. Booz Allen is committed to delivering results that endure. We are proud of our diverse environment, EOE, M/F/D/V.

Job: Program/Project Management; Primary Location: United States-Maryland-Rockville; Travel: Yes, 25 % of the Time

Top 10 LDA topics

Topic labels were manually assigned based on the top 10 terms per topic. Terms are assigned to topics at the collection level. Terms presented here comprise the topic and may not appear in each ad. The % value represents the probability that this topic is assigned to the sample ad.

- **Qualifications 30%**
- **Machine learning 13%**
- **Statistics 4%**
- **Software development 3%**
- **Development 3%**
- **Teamwork 3%**
- **Consulting 2%**
- **Client management 2%**
- **Medical domain 2%**
- **Analytics 2%**

Selection of qualitative codes

A. Candidate qualifications	B. Employer	D. Position
A2 Education	B1 Domain	C1 Position title
A3 Experience	B2 Employer name	C2 Role
A4 Knowledge	B3 Employer description	C3 Duties
A5 Skills	B4 Location	C9 Travel expectations
A6 Other qualifications		
	C. Tools, programming languages, and platforms	

Qualitative codes were created based on a detailed review of a random sample of 30 job ads. Only codes assigned to this example ad are displayed.

K-means cluster

Cluster labels were manually assigned based on review of top 10 ads in each cluster. The number (0.63) is the cosine distance of this job ad from the cluster centroid.

Data Scientist (0.63)

Key Findings

- Clusters effectively revealed different roles including data scientists, librarians, curators, data center personnel, data entry, and user support roles.
- Distinct clusters emerged for engineers including data center engineers, big-data software engineers, and software architects.
- Clusters also highlighted the importance of specific domains (e.g., medical, military).
- Clustering analysis successfully grouped the non-data positions (e.g., photo assistant, babysitter) into one cluster. It also identified job ads with near identical text and ads from the same employer (e.g., Bloomberg).
- Topic modelling was effective at identifying specific aspects of ads including roles, specific technologies, employment characteristics (e.g., full-time), geographic locations, and work environment.
- In comparison to qualitative analysis, topic modelling produced similar topics, further refined some codes (e.g., machine learning vs. statistics), and highlighted new themes in the ads such as enterprise architectures and boilerplate language (e.g., EEO).

Conclusions

- Qualitative coding and analysis offers the researcher greater control and produces themes relevant to the research questions. However, this approach is infeasible for large collections and can introduce bias or error since it is a subjective approach.
- Clustering and topic modelling techniques offer a repeatability via statistical and probabilistic models. While these methods are faster than qualitative analysis, both still require additional work to interpret and evaluate individual clusters or topics.
- While clustering does confirm the broad classes of data professions, it is too coarse and the aspects of the ads that contribute to clustering are often not useful for this type of workforce analysis (e.g., company boilerplate language).
- Topic modelling produces output closer to themes found in qualitative coding, but is more difficult to interpret.
- Topic modelling seems most promising for large-scale workforce-analysis using text-based job ads and descriptions.

References

Cox, A. M., & Corrall, S. (2013). Evolving academic library specialties. *Journal of the American Society for Information Science and Technology, 64*(8), 1526-1542.

Swan, A., & Brown, S. (2008). *The Skills, Role and Career Structure of Data Scientists and Curators: An Assessment of Current Practice and Future Needs. Report to the JISC.*