Identifying Highly Relevant Skills for Data Science Positions Captured 2023 through Topic Modeling with LDA, Top2Vec, and BERTopic

Jared Sarabia
https://jmsarabia.github.io/
jared.sarabia@gmail.com

Natural Language Processing

Dr. Alianna J. Maren

August 27, 2023

Abstract

As data science positions solidified in recent years to have clearer distinctions between roles and responsibilities, the skillsets and toolkits have changed just as quickly. To identify the current state of these skills and tools to their relative positions in 2023, this paper applies the unsupervised topic-modelling techniques of Latent Dirichlet Allocation (LDA), BERTopic, and Top2Vec to extract this information into categories of skills from 3,197 Artificial Intelligence (AI), Machine Learning (ML), and Data Science (DS) job postings collected into a Kaggle data set from a single job board. The clusters of skills and tools found in the Top2Vec model presented the highest overall coherence score across all topics, though the topics in LDA presented the most human interpretable topics.

Keywords
Natural language processing
Topic modeling
Skill extraction
Job advertisements
Modeling Comparison

1. Introduction

1.1 Background

Job listings display the real-time state of the job market for existing and newly created roles. With how quickly roles and responsibilities change, workers in the data science field become confused as to which roles they are qualified to apply to. As industries start to identify how these roles perform, responsibilities for these roles change and sometimes separate into new roles. For example, the data scientist role, once a catch-all for all data related roles with skills expected to range a vast array of competencies, eventually branched off more broadly into data scientists, data engineers, and data analysts. Furthermore, technical roles were molded to be more in tune with data rather than being domain-specific, such as quality assurance engineers specializing into data quality engineers.

There are several challenges to extracting skillsets from data science job listings. Due to the rapidly changing data science field, roles and responsibilities change quickly, so those interested in breaking into the field and those already in the field are constantly playing catch-up to what the latest tools, methodologies, and problems companies are trying to solve. Another issue stems from differences in industries and companies. The same role in different industries

will more than likely have different requirements, yet that same role in the same industry may have different requirements due to the different companies' organizational structure, size, and technological infrastructure. Conversely, there are job titles that are different yet overlap in skill requirements and responsibilities, such as clear communication skills being a requirement in all positions.

To address these problems, this paper propose using the unsupervised methods of Latent Dirichlet Allocation (LDA), Top2Vec, and BERTopic on a snapshot-in-time subset of the global data science job market. This is considered a subset as it is only the data science related jobs posted to one site ai-jobs.net, and would require scraping all English-speaking job boards available such as LinkedIn Jobs and Indeed to be considered a more holistic picture of the global data science job market. These topic modeling approaches represent a few of the highest performing and most recent approaches, though current literature has approached the job skill analysis through various approaches. These approaches will be discussed in Section 2 in more detail. Unsupervised methods were selected over supervised to ensure preconceptions of roles do not affect the skill clustering.

These specific methods were selected as they are high performing points along the evolution of topic modeling algorithms. LDA is a classic bag-of-words method that uses a Term Document Matrix (TDM) to form document-topic and term-topic matrices that can be easily interpreted by humans. Top2Vec takes this a step further by including a documents' semantic information, such as appearance of words next to each other, through the Doc2Vec word embedding. BERTopic similarly gathers semantic information through various word embeddings, but uses Transformers to gather this information instead.

2. Related Work

While there has been research into skill extraction from the total job market (Ao et al. 2023), there has yet to be a published paper focused solely on skills for data science jobs. Economics-based approaches to skill extraction focus primarily on manually-built word-counting methods, which establish skill categories that encapsulate many general skills. For example, the created topic "Cognitive Skills" could encapsulate any requirements such as "problem solving" and "analytical" (Deming & Kahn 2018). While these may be good starting points for deciding topics, these lists can grow out of date very quickly as new tools are developed or as roles develop different traits over time.

The computer and data science literature offers different solutions. Among the most popular methods are LDA (Blei et al. 2003), BERTopic (Grootendorst 2022), and Top2Vec

(Angelov 2023). The LDA approach uses a Bayesian approach of applying Dirichlet priors to estimate the probabilistic distributions of documents to latent topics. Each topic is then described by a probabilistic distribution of terms to that topic. The BERTopic approach extracts topics by generating document embeddings from pre-trained language models built from Sentence-Bidirectional Encoder Representations from Transformers (SBERT). The document embeddings are ultimately fed into a Term Frequency-Inverse Document Frequency (TF-IDF) procedure to define the final classes. Similar to BERTopic, Top2Vec uses the pre-trained Doc2Vec to create word, document, and topic embeddings. The

While these approaches have pros and cons, they have yet to be applied to small-to-medium sized datasets with an individual overarching topic. Whereas previous literature (Ao et al. 2023) examines a variety of fields to represent the global workforce, this paper examines these applications on a subset of the workforce, specifically the data science community.

3. Data description and preparation

The data set was scraped from https://ai-jobs.net/ on 06/17/2023 using the Python package Selenium (Shil 2023). The relevant data was identified then extracted from their HTML elements, to be saved to a comma-separated-values (CSV) file. The final CSV file, consisting of 8 columns and 3,197 data science related jobs, was then uploaded to Kaggle to be distributed to interested parties. The extracted format with definitions of the columns can be found in Appendix A.

While there are exploratory data analysis notebooks on Kaggle working with this dataset, this paper applies topic modeling techniques to compare skills required to each role. This has the additional effect of ordering the importance of skills to jobs through their semantic relevance to each other through the word embeddings as seen in Top2Vec. Furthermore, skills are naturally clustered together through these methods, so complimentary skills are easily identifiable.

4. Methods

Prior to modeling, the “Requirement of the company” column was cleaned by tokenizing the words, fixing capitalization, removing stop words, and removing non-alphabetic tokens, specifically for LDA. The tokens were then one-hot-encoded for ingestion by each topic modeling approach (Brownlee 2017). The “Job Title” column was further simplified into 11 jobs: 'Data Analyst', 'Data Engineer', 'Data Scientist', 'Machine Learning Engineer', 'BI Analyst', 'AI Engineer',' Software Engineer', 'DevOps Engineer', 'Architect', 'Big Data Engineer', and 'Research Scientist.' Only jobs that had one of these combinations of words were treated as part

of the posting. Otherwise, the "Job Title" was not included in the final string of words that would be considered an individual document. Each document would be a job posting consisting of the columns "Job Title," "Job Type," and "Requirements." This was done as inclusion of other columns muddied the interpretability of each topic.

For the topic modeling, this paper compares the process complexity and the final topic results from the LDA, Top2Vec, and BERTopic approaches. For the results, I will compare the number of optimal topics, the number of outliers after the topics are found, their coherence scores, and how easy to visualize these topics are given existing packages (Albanese 2022). The specific coherence score, CV, was selected due to its performance when compared to human benchmarks (Röder 2015).

5. Results

The raw results for each topic created by each algorithm can be found in Appendix B. However, high level results for the models will be compared here. The best number of topics for LDA, with the highest coherence score of 0.33 and best interpretability among the number of topics tested, was 6. The Top2Vec model was found to have the highest coherence score among all the models, ending with 5 topics generated. The BERTopic model generated 10 topics, with a coherence score 0.41 across all the topics. Image 1 displays the comparison of the models through their coherence scores.

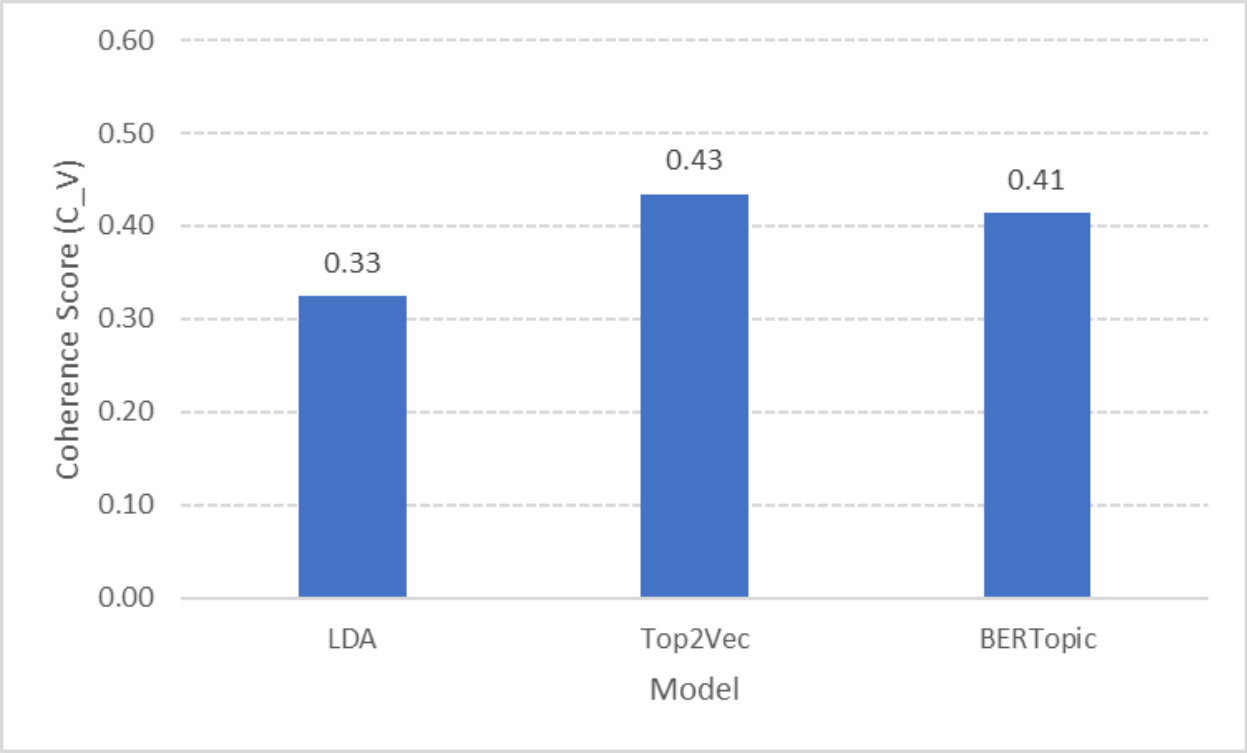

Image 1: Coherence scores by model, with Top2Vec having the highest overall coherence across all the topics generated.

6. Discussion

The number of topics for LDA was selected based on coherence score and interpretation of clusters. With larger topics, more skills were found correlated to each other, yet only three jobs (Data Analyst, Data Scientist, and Data Engineer) were found to be highly correlated to topics, which makes sense given how many postings included some form of these three job positions. Also, the topics made sense to a certain degree, as Excel, PowerBI, and SQL were found to be highly correlated skills for entry level data analysts within this topic (Appendix B).

Top2Vec had the highest coherence score, though the human interpretability seems more difficult. Topics 3 and 4 can be easily described as highly technical engineering topics. However, the other topics are harder to summarize, as their technical aspects range across various fields within the same topic. Furthermore, visualizing the intertopic distance between topics is not natively built into packages for Top2Vec, as it is for LDA and BERTopic. The word clouds generated

BERTopic resulted in 10 topics, though the -1 topic could be ignored as it is a catch-all topic for tokens, according to the documentation. The coherence score was low, though hardware limitations did require model constraints that likely limited its performance. The topics generated provide some insight to sets of tools used at the high levels. For example, in Topic 8, the senior level architects seem to be related to more front-end tools than back-end.

Table 1 describes the differences found between the three algorithms, with the related key of colors outlined in the caption.

Metric	LDA	Top2Vec	BERTopic
Ease of Set Up/Available Packages	Gensim is all that is needed for the base model	Multiple package dependencies can get tricky	Multiple package dependencies can get tricky
Data Preparation	Pre-processing is absolutely needed	Pre-processing not typically needed	Pre-processing not typically needed
Document-topic relationship	Each document can be assigned to multiple topics	Each document is assigned only one topic	Each document is assigned only one topic

Topics	Semantic information lost in bag-of-words	Semantic embeddings create better topics	Semantic embeddings create better topics
Optimizing Number of Topics	Start with guesses of topic numbers, then refine	Hierarchal topic reduction aids in finding optimal topic number	Hierarchal topic reduction aids in finding optimal topic number
Outliers	Can be ran without outliers	Creates outliers	Creates outliers (can be handled with other clustering methods)
Document Sizes	Performs well on all sizes	Performs better on shorter documents	Performs better on shorter documents
Dataset Size	Performs well on all sizes	Performs better on larger datasets	Performs better on larger datasets
Computational Resources	Fast	Longer training times and may require GPUs	Longer training times and may require GPUs

Table 1: This table compares the three algorithms from various metrics based on literature review and findings. Green means there were few complications, yellow means moderate issues, and red means significant number of difficulties.

7. Future Work

This dataset could be supplemented with data scraped from other job sites such as LinkedIn using the same process that Shil implemented to create this dataset, using packages such as Selenium or using the job board's API. Another option is to supplement with other Kaggle datasets, such as Olonade's *Data Science Job Postings (Indeed USA)*. This paper chose to exclude Olonade's data for the moment as it requires more processing work, is focused on the USA, and is from an earlier time (2022). A future project could be to have a more focused look into the US data science job market, but on a more updated timeline that better reflects recent economic trends such as how the early 2023 layoffs have affected the job market.

8. Conclusions

Topic modeling can prove to be a powerful tool in keeping data science workers informed of the latest tools and skills required by different roles. By modeling skills and job titles into topics, workers can identify complementary skills and trends that span across different roles. The topic models used to complete these tasks represent part of the evolution of topic modeling Starting with bag-of-words techniques such as LDA, algorithms were improved with semantic embeddings such as in Top2Vec and are currently being worked on with Transformers such as in BERTopic.

Appendix A

The **data_science_job.csv** file is a structured dataset containing information related to data science job opportunities. It consists of the following columns:

1. Company: This column lists the names of the companies offering data science positions. It provides insights into the diverse range of organizations hiring in the field.

2. Job Title: This column specifies the job titles associated with each position. It helps job seekers and researchers understand the various roles available within the data science domain.

3. Location: The location column indicates the geographical location of each job opportunity. It helps individuals identify job prospects in specific regions or countries.

4. Job Type: This column classifies the job types available, such as full-time, part-time, or remote positions. It assists job seekers in understanding the nature of the roles and the flexibility they offer.

5. Experience Level: This column indicates the desired experience level for each job. It helps applicants assess their suitability for the positions based on their professional background.

6. Salary: The salary column provides information about the expected salary range or compensation package for each job opportunity. It enables candidates to gauge the financial aspects of the roles.

7. Requirement of the Company: This column highlights the specific skills, qualifications, and prerequisites that companies seek in potential candidates. It provides insights into the qualifications and expertise required for data science roles.

8. Facilities: The facilities column outlines the additional benefits or perks offered by the companies, such as healthcare benefits, flexible working hours, professional development opportunities, or other amenities. It gives job seekers a glimpse into the additional advantages provided by employers.

Appendix B

Output Files

References

Albanese, Nicolo Cosomo. "Topic Modeling with LSA, PLSA, LDA, NMF, Bertopic, Top2vec: A Comparison." Medium. September 22, 2022. https://towardsdatascience.com/topic-modeling-with-lsa-plsa-lda-nmf-bertopic-top2vec-a-comparison-5e6ce4b1e4a5.

Angelov, Dimo. 2020. "Top2Vec: Distributed Representations of Topics." arXiv:2008.09470v1 [cs.CL]. (Accessed July 02, 2023; available online at arXiv:2008.09470v1 [cs.CL].)

Ao, Ziqiao, Gergely Horváth, Chunyuan Sheng, Yifan Song, and Yutong Sun. 2023. "Skill Requirements in Job Advertisements: A Comparison of Skill-Categorization Methods Based on Wage Regressions." *Information Processing & Management* 60, no. 2 (Accessed July 02, 2023; available online at https://doi.org/10.1016/j.ipm.2022.103185.)

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993-1022. (Accessed July 01, 2023; available online at https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.)

Brownlee, Jason. 2017. *Deep Learning for Natural Language Processing: Develop Deep Learning Models for your Natural Language Problems* (version 1.1). *Machine Learning Mastery*. 1.1ed. Machine Learning Mastery. https://machinelearningmastery.com/deep-learning-for-nlp/.

Deming, David, and Lisa B. Kahn. 2018. "Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals." *Journal of Labor Economics* 36, no. S1. (Accessed July 01, 2023; available online at https://doi.org/10.1086/694106.)

Grootendorst, Maarten. 2022. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." *arXiv*:2203.05794v1 [cs.CL]. (Accessed July 02, 2023; available online at arXiv:2203.05794v1 [cs.CL].)

Olonade, Yusuf. *Data Science Job Postings (Indeed USA)*. 2022. Distributed by Kaggle. (Accessed July 04, 2023; available online at https://www.kaggle.com/datasets/yusufolonade/data-science-job-postings-indeed-usa?select=data_science_jobs_indeed_usa.csv.)

Röder, Michael, Andreas Both, and Alexander Hinneburg. 2015. "Exploring the Space of Topic Coherence Measures." Association for Computing Machinery. (Accessed August 18, 2023; available online at https://doi.org/10.1145/2684822.2685324.)

Shil, Joy. *Scraped Data on AI, ML, DS & Big Data Jobs*. 2023. Distributed by Kaggle. (Accessed July 01, 2023; available online at https://doi.org/10.34740/KAGGLE/DSV/5963366.)