Text & Data Mining Project Summary

Chirs Reynolds | Spring 2024

Project Summary

This project looks at how machine learning can be used to classify academic library job postings based on department focus and shared skill sets. Using a dataset of job listings from the University of California (UC) and California State University (CSU) systems, the goal was to train a neural network (NN) to recognize patterns in job descriptions and group them into relevant categories.

The idea was partly inspired by San José State University's *MLIS Skills at Work* report, which takes a similar look at job titles and trends in the field. Over several assignments, the project moved through stages: web scraping, cleaning and preprocessing text, clustering documents, manually tagging categories, and building a neural network model. The final result is a working classifying NN that can distinguish between different roles in academic libraries.

Final Dataset Refinement

In the earlier stages of the project, job postings were collected from a number of academic institutions across the United States. These postings varied in structure, tone, and length. Many included extensive non-essential content (like institutional overviews or promotional language) that was unrelated to the actual job qualifications or responsibilities.

For the final model, the dataset was refined to improve consistency and focus. All job postings were sourced exclusively from the University of California (UC) and California State University (CSU) systems. This narrowed down my scope and provided uniformity in posting style and language use. Each listing was manually cleaned to include only the job description and qualification criteria.

Methods

This project involved developing a neural network model to classify academic library job postings into professional categories. The process followed an iterative, exploratory methodology over six assignments, each contributing a foundational component to the final system. The following sections outline each phase of development.

Data Collection via Web Crawling

The initial phase focused on identifying and retrieving job postings from academic libraries, primarily in the western United States. Using custom web crawling techniques, job listings were scraped from the library-specific employment pages of various institutions. Challenges included inconsistent webpage structures, non-persistent search queries, and inaccessible or broken links. Data collection was refined to focus on institutions with clearly structured job pages and yielded a modest corpus of academic job listings.

Text Preprocessing

Raw job descriptions were processed to prepare them for analysis. Using text mining tools in RapidMiner, preprocessing steps included:

- Tokenization
- Case normalization
- Stopword removal
- Generation of 3-grams
- Token length filtering
- Stemming (Porter Stemmer)

The resulting term-document matrix served as the basis for exploratory text analysis and later modeling tasks. Additional manual cleaning was done to remove HTML artifacts and irrelevant institutional information from the postings.

Document Clustering & Exploratory Analysis

Unsupervised clustering techniques were applied to understand possible groups within the job postings. Four clustering methods were tested:

- Random clustering (baseline)
- k-Means clustering
- x-Means clustering
- Hierarchical Top-Down clustering (with k-Means as a subprocess)

This phase was exploratory: providing insights into how the postings might cluster based on text similarity. Although interpretation was initially challenging, the results suggested latent groupings that informed future categorization.

Data Labeling for Supervised Learning

To prepare the data for supervised classification, job postings were manually labeled into thematic categories. Initial tagging included categories such as "Subject Librarian," "Archives," "Management," and "Instruction." This process was iterative and informed by both domain knowledge and clustering patterns found in the previous step.

Multiple tagging groups were tested, evolving from job types to levels of seniority (entry, mid, leadership) and finally to more nuanced professional categories based on role responsibilities. This phase showed the ambiguity and overlap within academic job roles.

Neural Network Model Development

A NN was trained to predict job categories from posting text. Initial architectures varied in complexity, exploring different combinations of:

- Training cycles (up to 1000)
- Hidden layers (up to 4)
- Node counts (ranging from 50 to over 600)
- Learning rates (0.05 to 0.2)

The most effective configuration included 4 hidden layers with 50 nodes each and a 0.2 learning rate, achieving reasonable performance on a validation test.

Model Testing & Refinement

In the final phase, the dataset was narrowed to postings from the UC and CSU systems for consistency. Job descriptions were further cleaned to isolate qualification-relevant content. The neural network was retrained using optimized settings (500 cycles, learning rate = 0.05, momentum = 0.9, hidden layers = [200, 200, 100]).

The final classification schema consisted of:

Archives & Special Collections

Metadata & Cataloging

Collections Management

Data Management & Systems

Outreach & Instruction

Reference & Public Facing Roles

Interns

Despite overlapping duties in many roles, the refined model demonstrated a promising ability to distinguish between categories with a small margin of error. Clustering analysis was revisited to validate grouping logic, which occasionally revealed unexpected job role associations (e.g., Special Collections with Archives).

Progression of Job Classifications though Model Development

	First Attempt (ex 7)	*Second Attempt*	*Final Groups*
1	Subject Librarian	Archives	Archives
2	Archives	Metadata & Cataloging	Metadata & Cataloging
3	Special Collection	Management & Admin	Collections, Acquisition & Circulation
4	Collections	Instruction	Management & Admin
5	Metadata	Public Interaction	Data Management & Systems
6	Coordinator	Collections	Outreach, Programming & Instruction
7	Assistant		Reference & Public Facing Roles
8	Management		Interns
9	Instruction/Reference		
10	Other		

Dept	Prediction
1	1
4	4
8	8
6	6
1	1
6	6
6	6
3	3
2	2
1	1
6	6
5	5
7	7
3	3
6	6
4	4
1	1
5	5
3	3
5	5
6	6
6	6
6	6
7	7
3	3
2	2
6	6
6	6
8	8
8	8
1	1
2	2
4	4
6	6
6	6
5	5
6	6
5	5
1	1
6	6

Validation Testing

On the right is my validation table which came out 100% correct!

Issues & Challenges

Personal notes

I’m not gonna lie, I almost gave up. I re-tagged my posts 6 times total. One of the main challenges was the significant overlap between job roles, which made it difficult to create distinct categories. To better understand how the model was interpreting the data, I re-ran the clustering analysis using the updated job postings and used those insights to inform my final classifications.

Final Thoughts

Overall, I really enjoyed this process even when I got a little turned around. It’s been exciting to see how each part of the workflow builds toward something functional. I recently attended an introductory machine learning seminar through ASIS&T and have started learning Python as well. I'm hoping these skills will help me keep exploring text and data mining techniques in more depth moving forward.