Binary Classification of Job Posts

Distinguishing IT from Non-IT Roles

Ahmed Abdulrahim
Bimsara Siman Meru Pathiranage
Simranjeet Kaur
Efemena Theophilus Edoja
Himanshu

Problem Definition and Research Inquiry

Problem Statement:

- **Objective**: Develop a machine learning model to classify job postings as IT or non-IT based on textual content.
- **Challenge**: Effectively process and analyze the nuanced language and terminologies used across various industries to ensure accurate classification.

Research Question:

- Can a machine learning model reliably distinguish between IT and non-IT job posts using textual data, and what are the key features that influence this classification?

Project Significance and Relevance

HR Efficiency

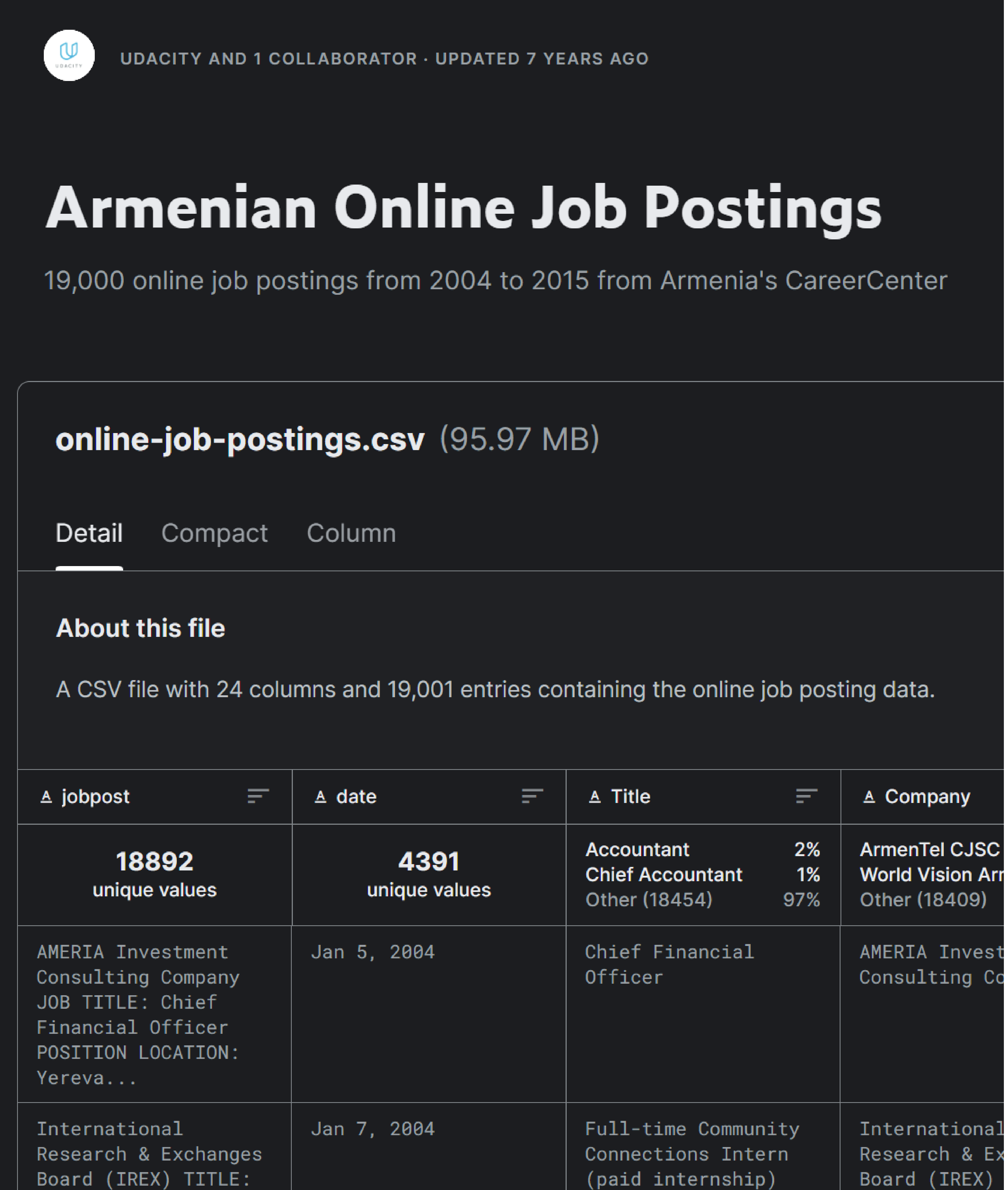
Streamlines Recruitment: Automates the sorting of job postings, significantly enhancing HR efficiency.

Data Analytics Education

Theory to Practice: Provides practical experience in NLP and machine learning within real-world contexts.

Job Search Optimization

Enhanced Targeting: Filters out non-relevant job postings, improving efficiency and relevance in job searches.



UDACITY AND 1 COLLABORATOR · UPDATED 7 YEARS AGO

Armenian Online Job Postings

19,000 online job postings from 2004 to 2015 from Armenia's CareerCenter

online-job-postings.csv (95.97 MB)

About this file

A CSV file with 24 columns and 19,001 entries containing the online job posting data.

jobpost	date	Title	Company
18892 unique values	4391 unique values	Accountant 2% Chief Accountant 1% Other (18454) 97%	ArmenTel CJSC World Vision Arr Other (18409)
AMERIA Investment Consulting Company JOB TITLE: Chief Financial Officer POSITION LOCATION: Yereva...	Jan 5, 2004	Chief Financial Officer	AMERIA Invest Consulting Co
International Research & Exchanges Board (IREX) TITLE:	Jan 7, 2004	Full-time Community Connections Intern (paid internship)	International Research & Ex Board (IREX)

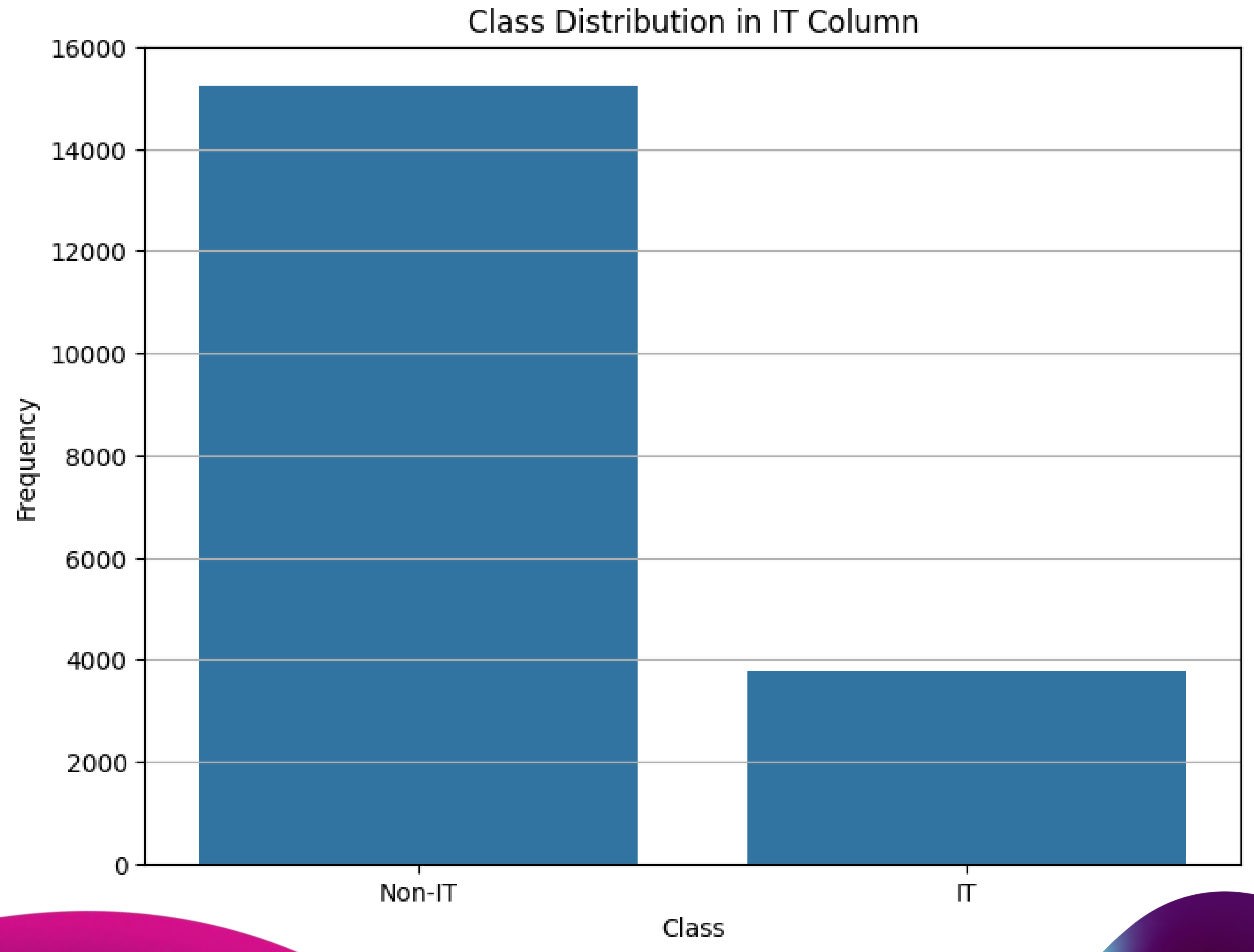
Data Source

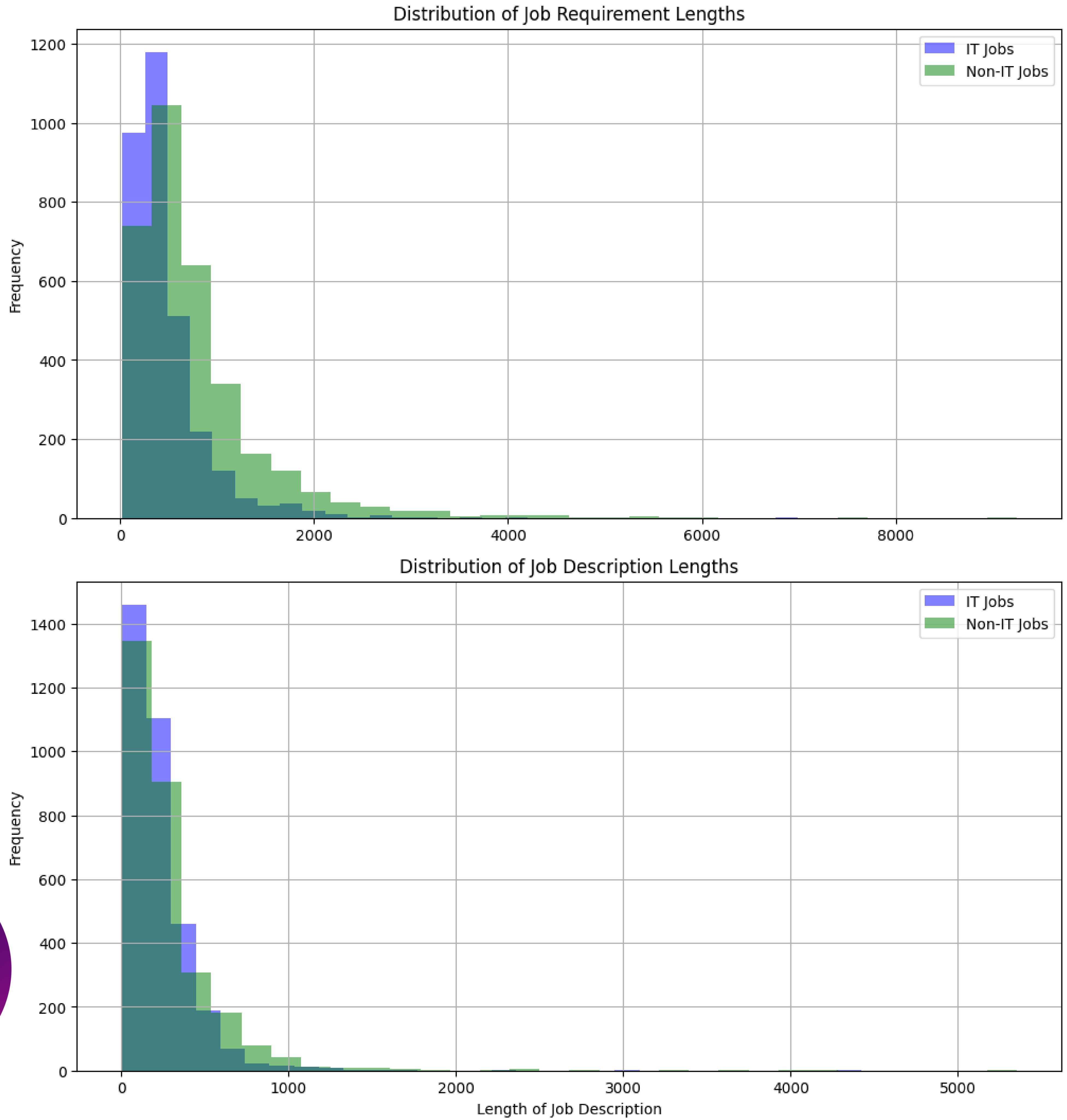
Brief Overview: This dataset contains **19,000** structured job postings from 2004 to 2015, sourced from CareerCenter, an Armenian job portal, and includes key details like job titles, company names, and descriptions.

Data Collection

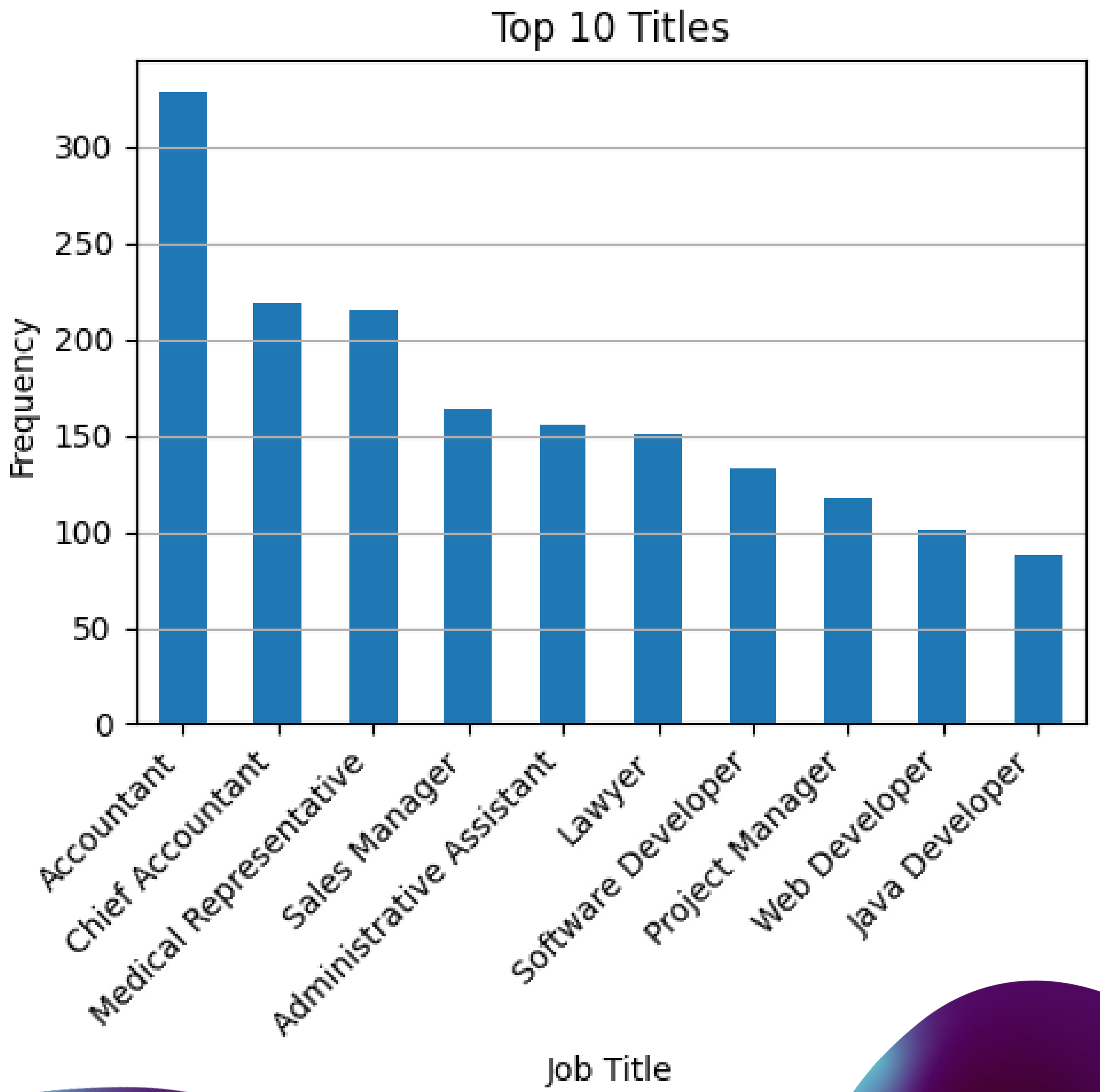
Our dataset is sourced from Kaggle and consists of **24 columns**, including a '**Job Post**' column that encompasses the entire job listing. Another key column, '**IT**', is a Boolean field indicating whether the job is IT-related, encompassing roles such as computer science, software engineering, data analysis, network administration, and cybersecurity.

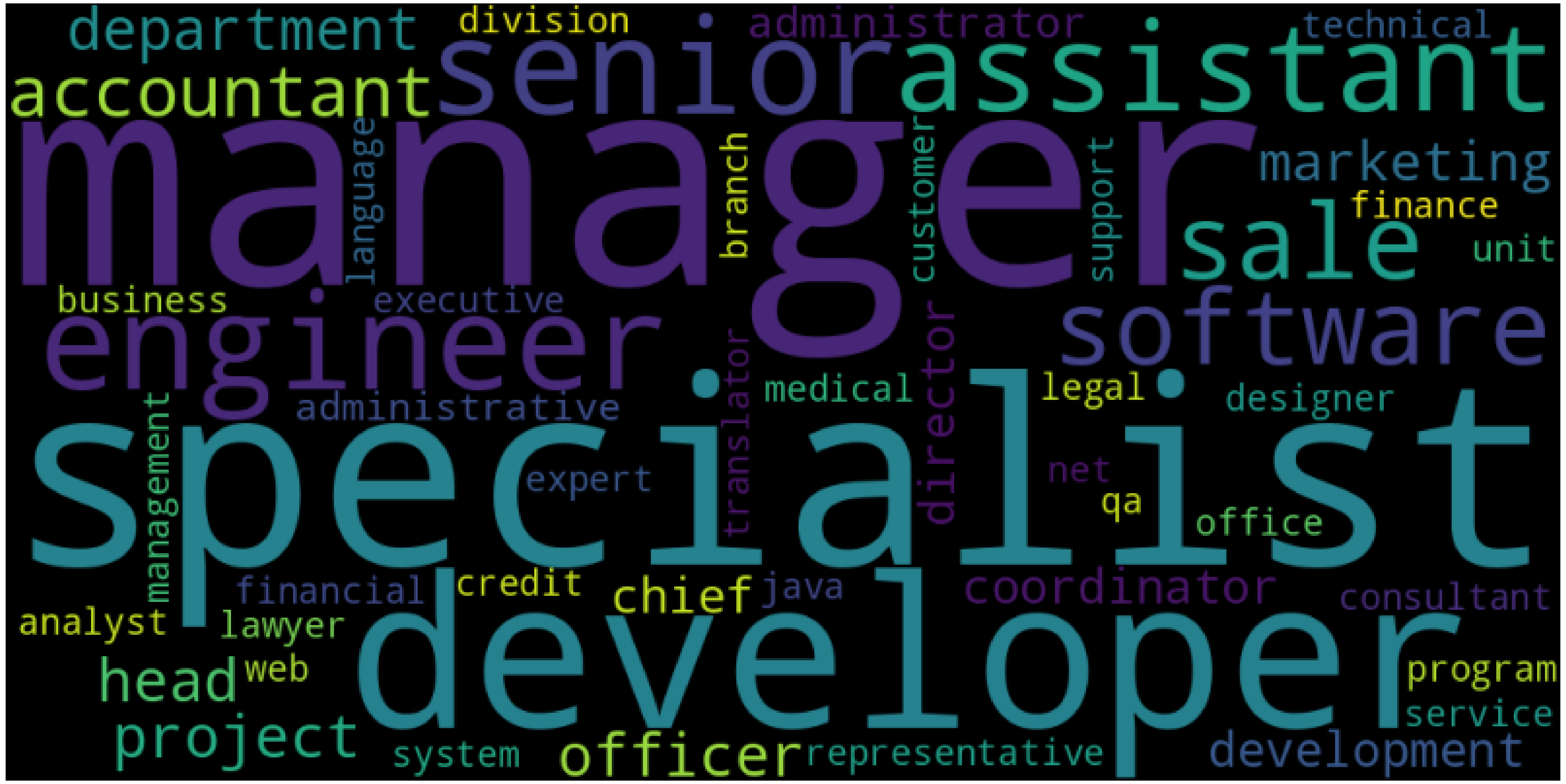
Data Exploration





Data Exploration



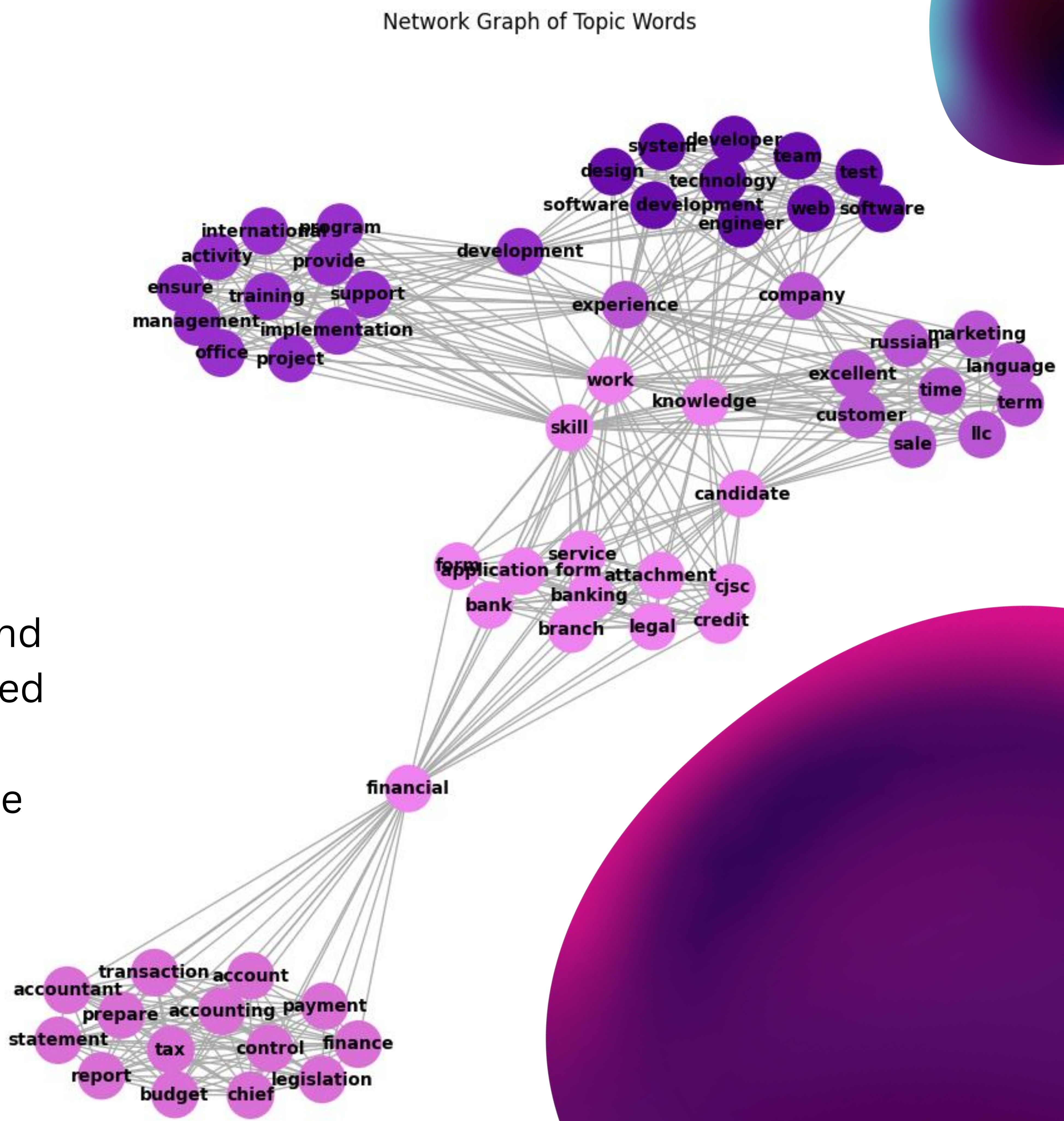


Job Title Tokens
WordCloud

Data Exploration

Topic Modeling with LDA: Leveraging Latent Dirichlet Allocation (LDA) to identify prevalent topics in job postings.

Latent Dirichlet Allocation (LDA) technique, is a type of topic modeling, to analyze and understand the main themes or natures of jobs from tokenized job postings. The goal is to identify clusters of words that represent significant topics within the job posts, which can help in understanding the nature and profiles of jobs.



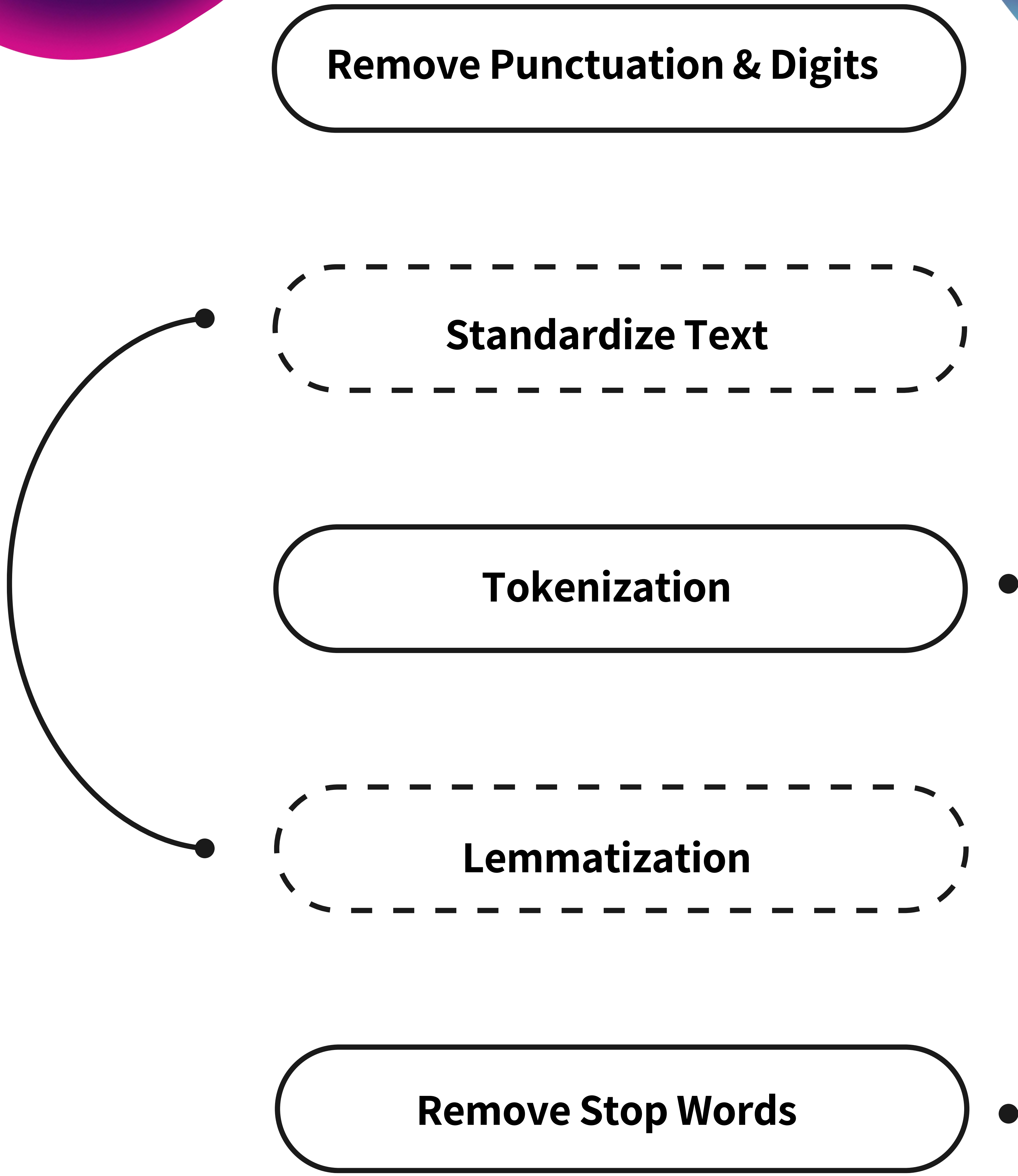
Data Preparation

Feature Extraction:

- TF-IDF (Term Frequency-Inverse Document Frequency): Quantifying the importance of words relative to their frequency across all documents.

Data Balancing:

- Apply SMOTE (Synthetic Minority Over-sampling Technique) to balance the distribution of IT and non-IT job labels.



Remove Punctuation & Digits

Standardize Text

Tokenization

Lemmatization

Remove Stop Words

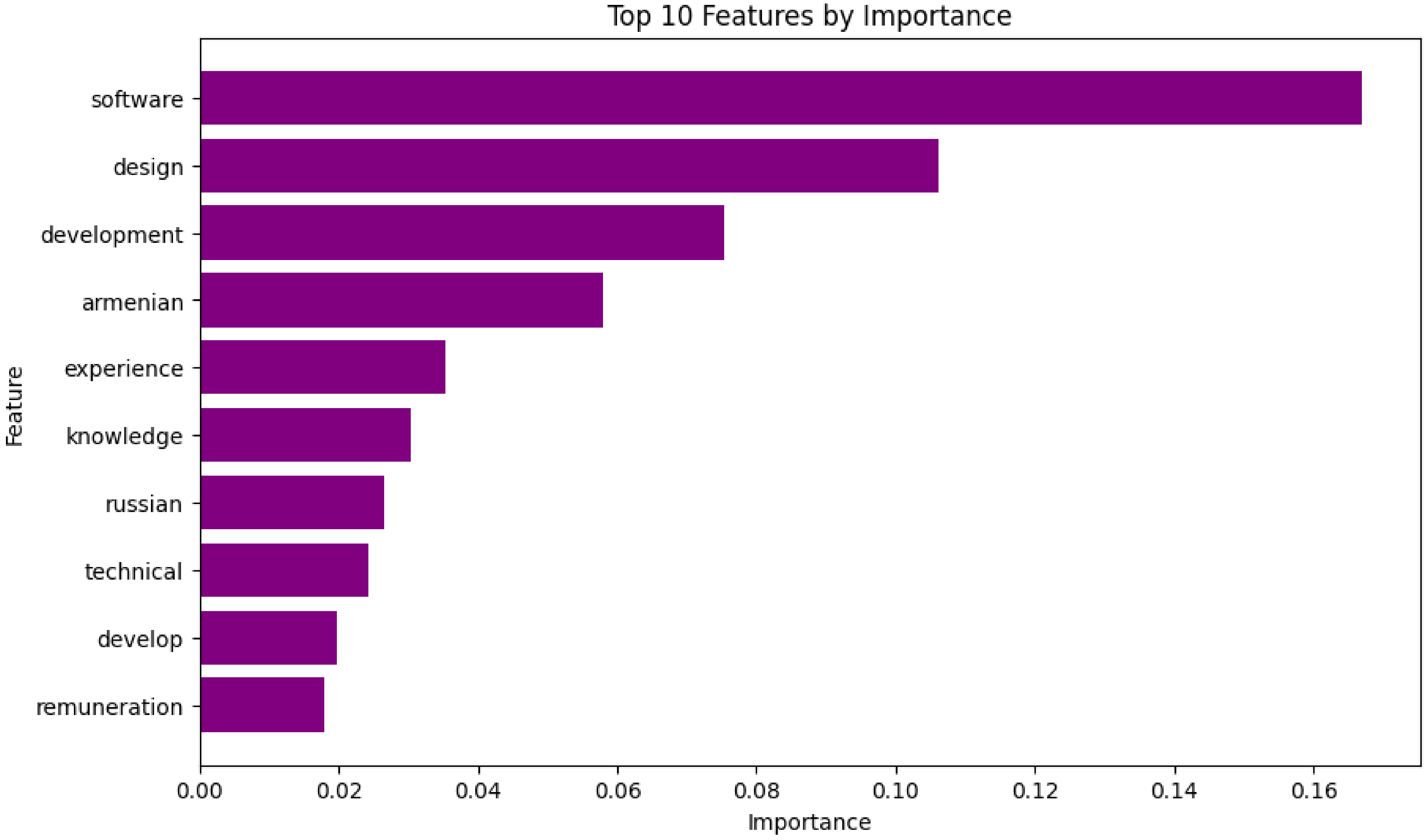
Modeling

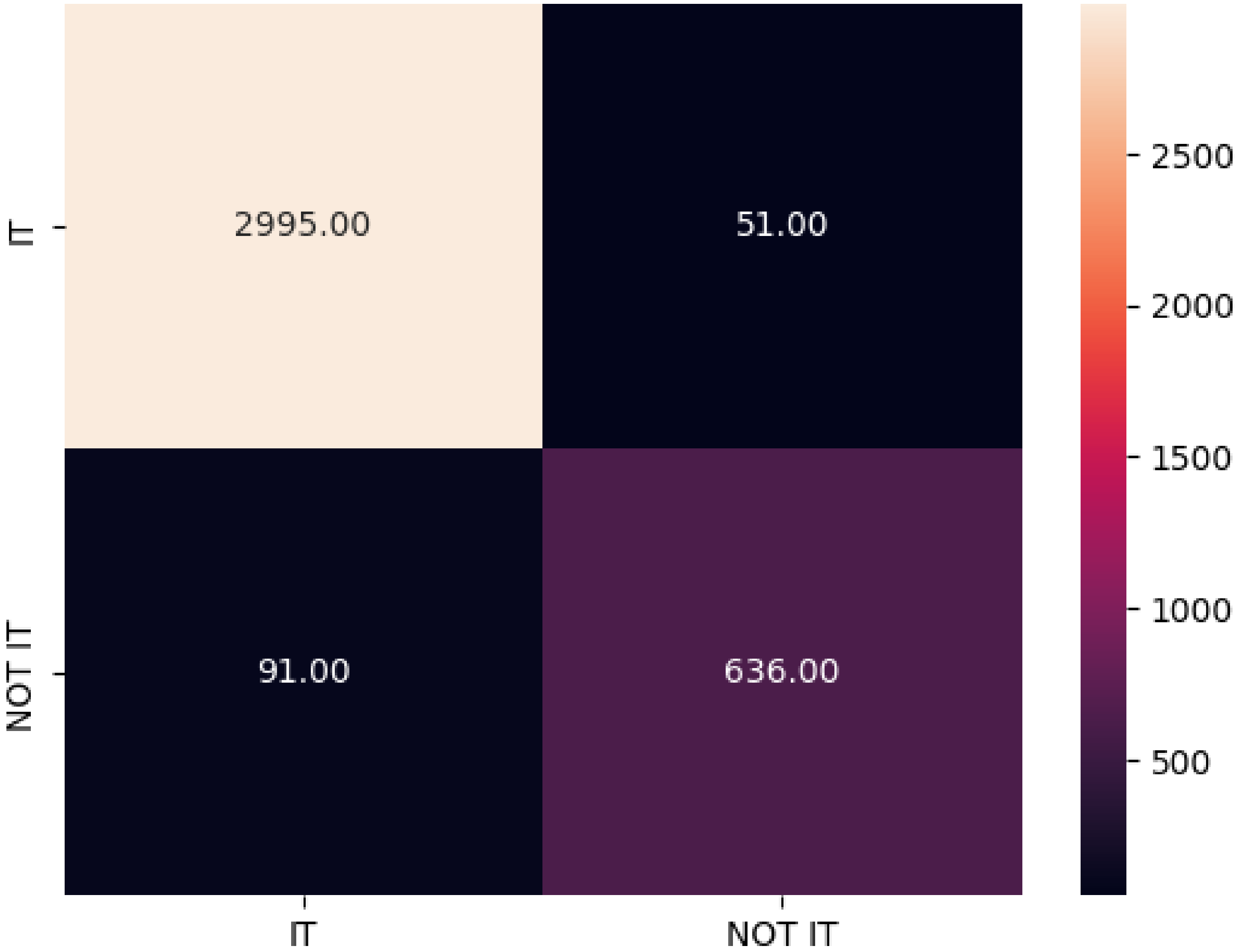
- **Random Forest:** max_features=25, n_estimators=180
- **RNN:** Embedding layer (128 units), LSTM layer (64 units), Dense output layer with softmax activation.

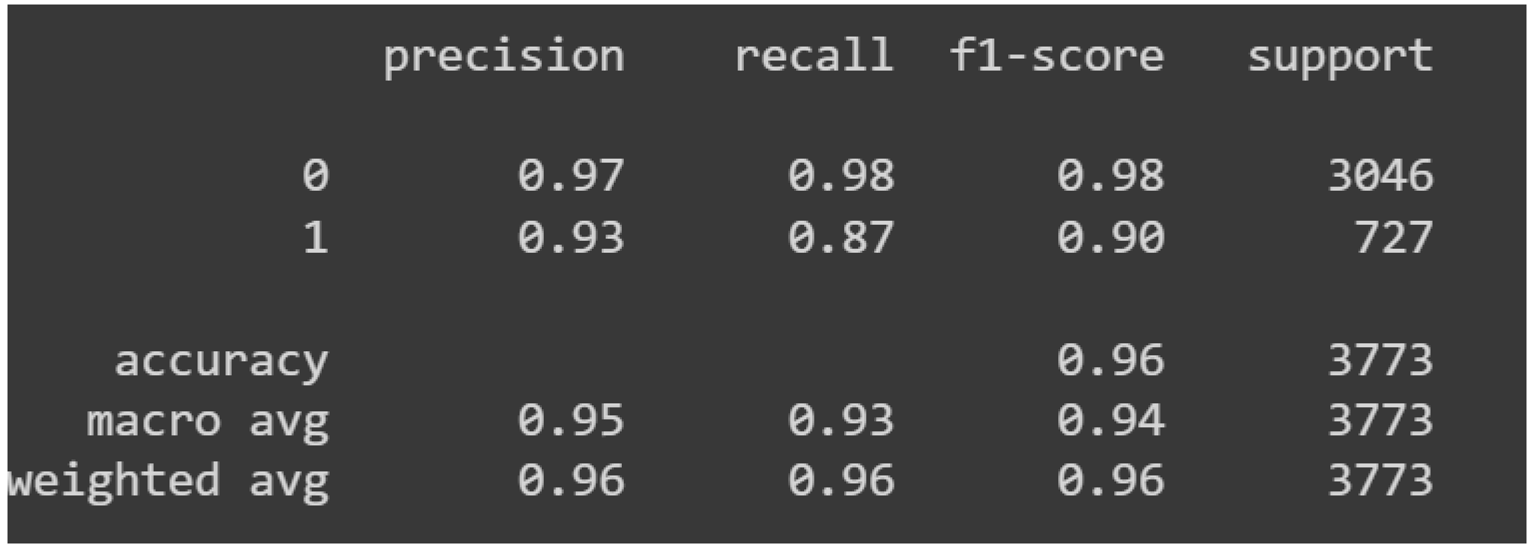
Model	Dataset	Accuracy	Precision	Recall	F1 Score	ROC-AUC Score
Random Forest	Train	0.9986	-	-	-	0.9986
	Test	0.9626	-	-	-	0.9290
RNN	Train	0.9796	0.9709	0.9651	0.9680	-
	Test	0.9244	0.8776	0.8804	0.8790	-

Model Evaluation

Random Forest



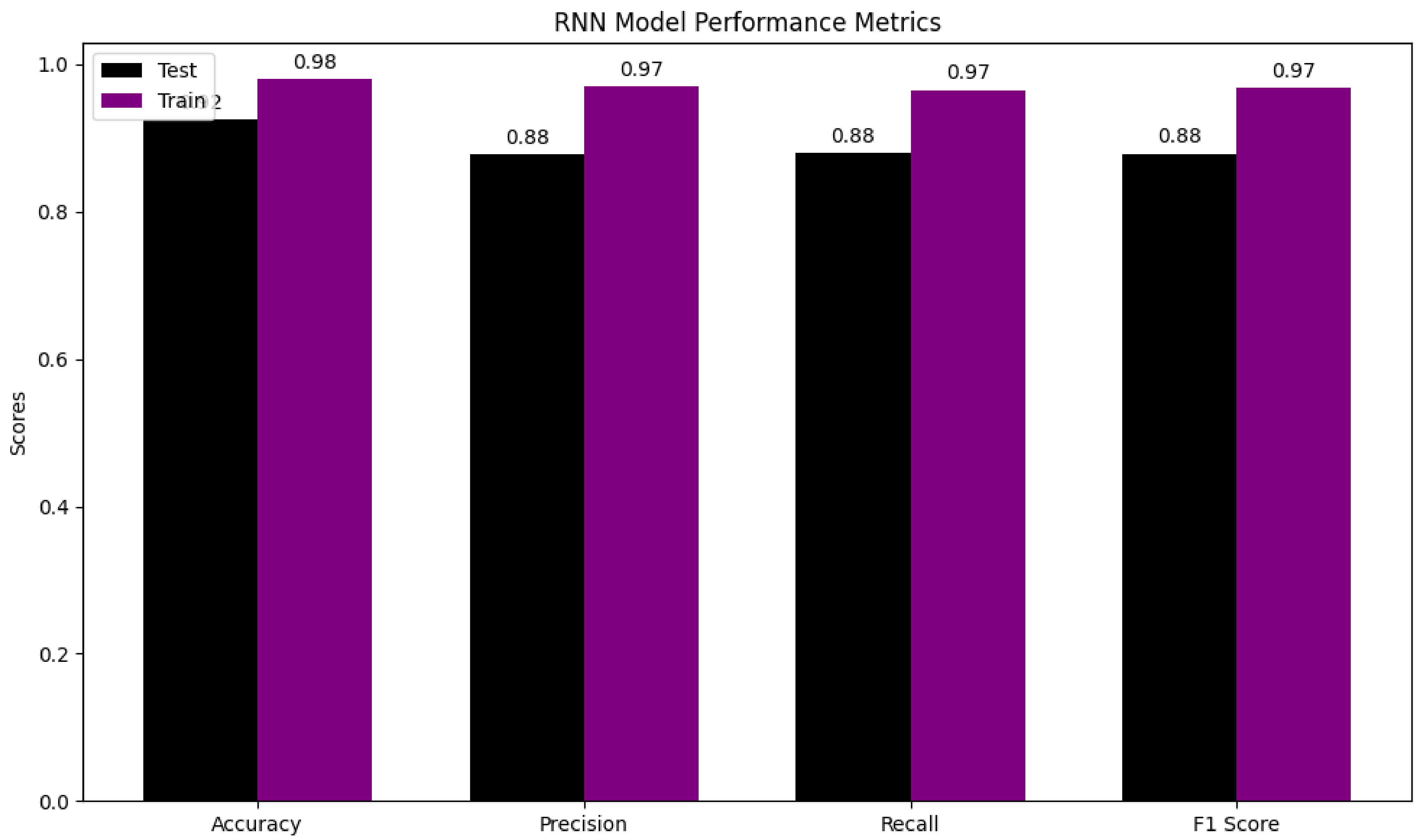


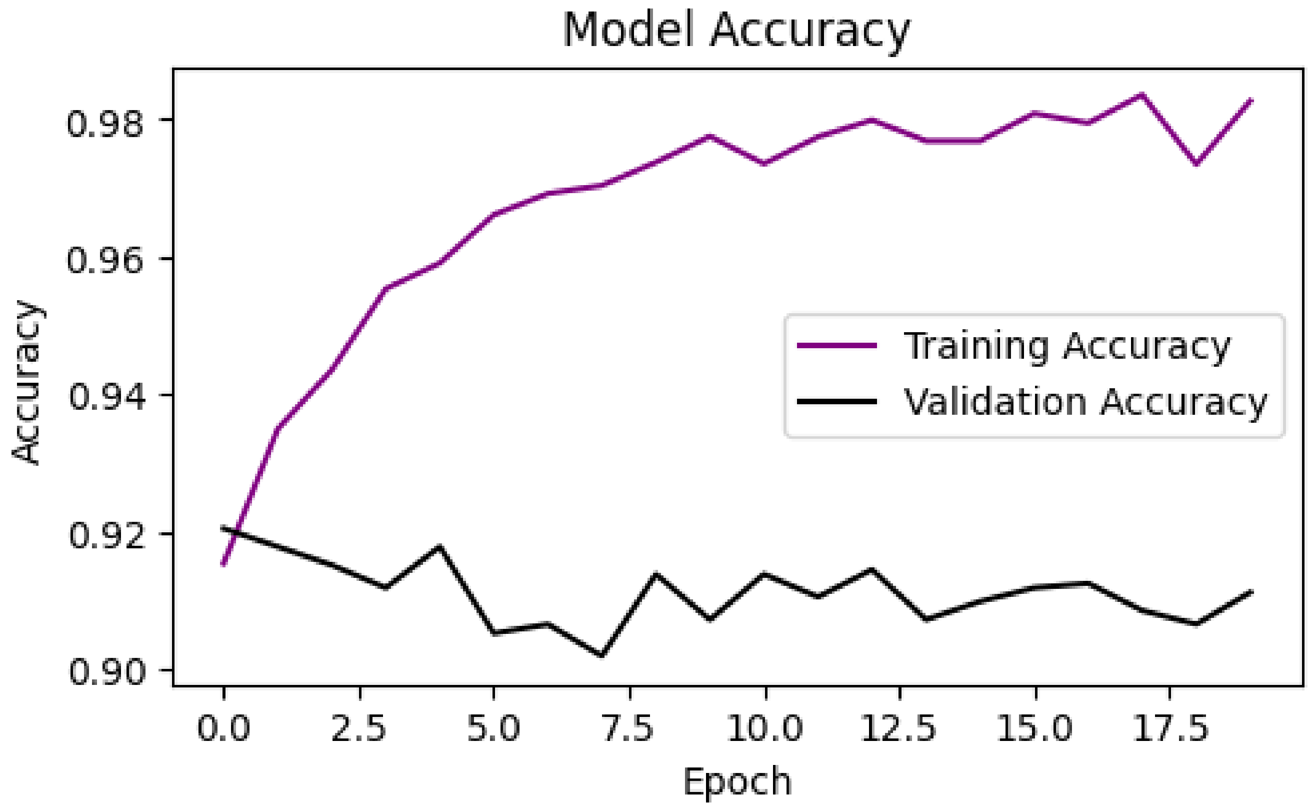


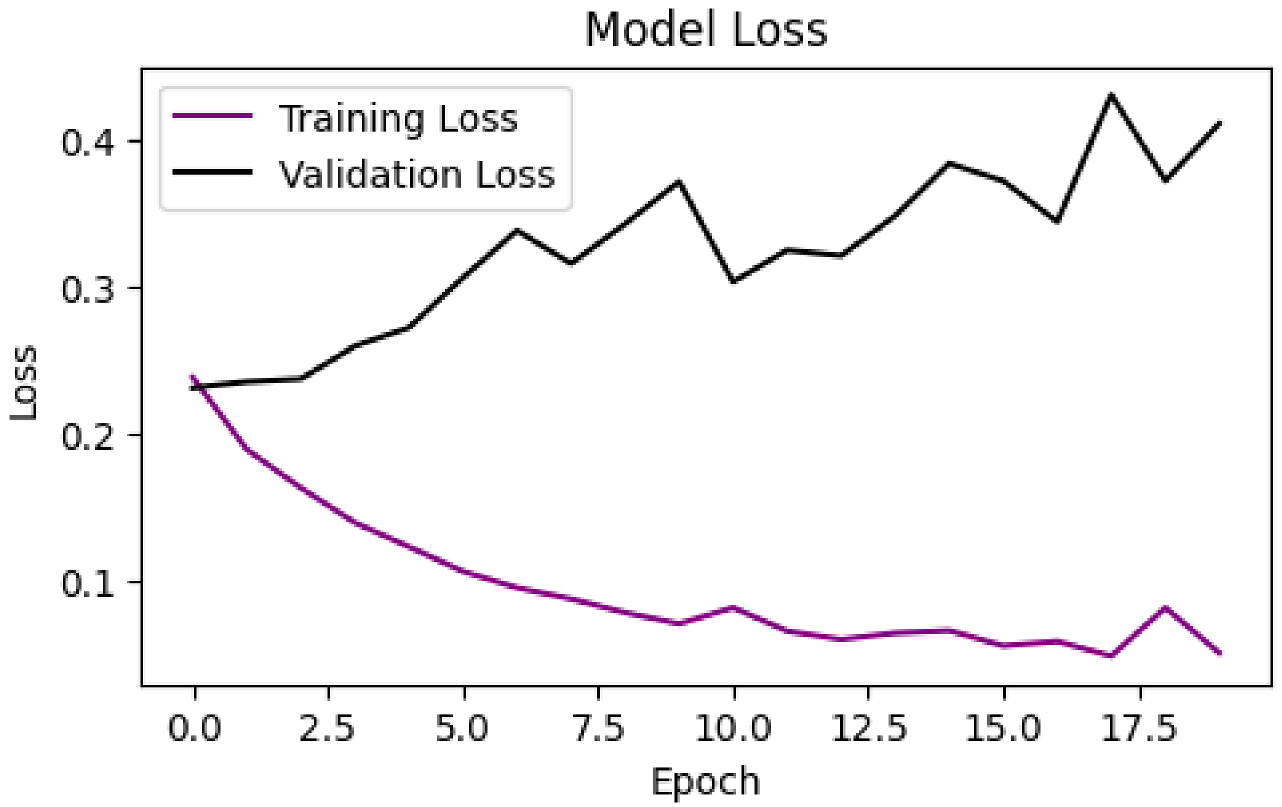
```
              precision    recall  f1-score   support

           0       0.97      0.98      0.98      3046
           1       0.93      0.87      0.90       727

    accuracy                           0.96      3773
   macro avg       0.95      0.93      0.94      3773
weighted avg       0.96      0.96      0.96      3773
```

Model Evaluation

Recurrent Neural Network

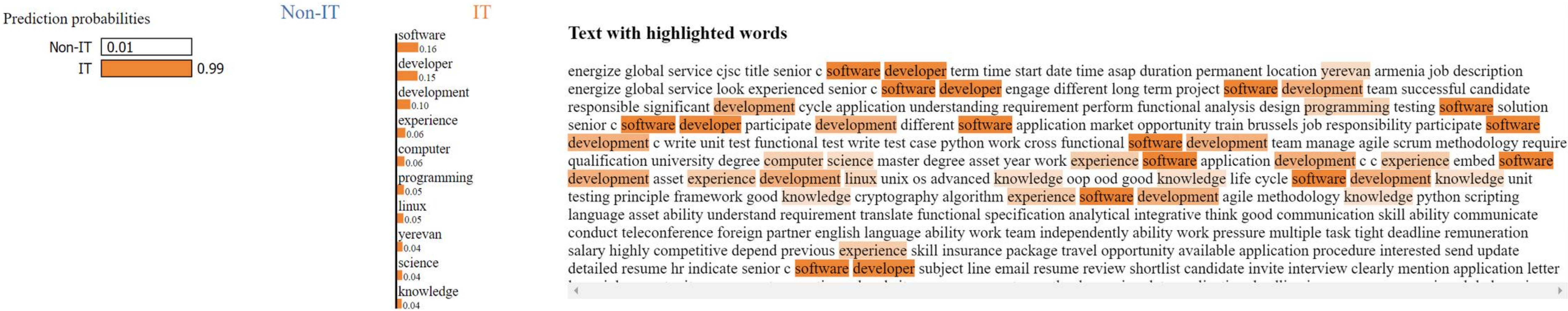






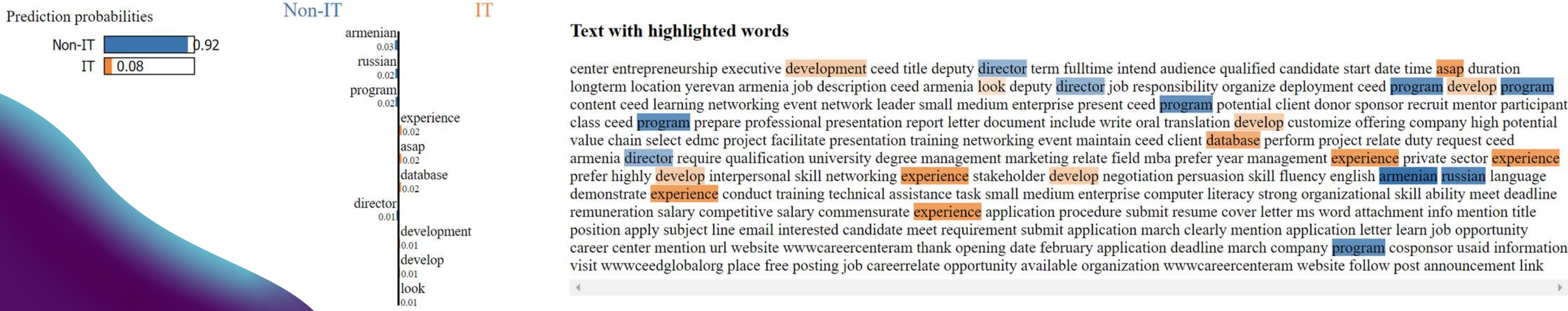
LIME

Local Interpretable Model-Agnostic Explanations



Text with highlighted words

energize global service cjsc title senior c software developer term time start date time asap duration permanent location yerevan armenia job description energize global service look experienced senior c software developer engage different long term project software development team successful candidate responsible significant development cycle application understanding requirement perform functional analysis design programming testing software solution senior c software developer participate development different software application market opportunity train brussels job responsibility participate software development c write unit test functional test write test case python work cross functional software development team manage agile scrum methodology require qualification university degree computer science master degree asset year work experience software application development c c experience embed software development asset experience development linux unix os advanced knowledge oop ood good knowledge life cycle software development knowledge unit testing principle framework good knowledge cryptography algorithm experience software development agile methodology knowledge python scripting language asset ability understand requirement translate functional specification analytical integrative think good communication skill ability communicate conduct teleconference foreign partner english language ability work team independently ability work pressure multiple task tight deadline remuneration salary highly competitive depend previous experience skill insurance package travel opportunity available application procedure interested send update detailed resume hr indicate senior c software developer subject line email resume review shortlist candidate invite interview clearly mention application letter



Text with highlighted words

center entrepreneurship executive development ceed title deputy director term fulltime intend audience qualified candidate start date time asap duration longterm location yerevan armenia job description ceed armenia look deputy director job responsibility organize deployment ceed program develop program content ceed learning networking event network leader small medium enterprise present ceed program potential client donor sponsor recruit mentor participant class ceed program prepare professional presentation report letter document include write oral translation develop customize offering company high potential value chain select edmc project facilitate presentation training networking event maintain ceed client database perform project relate duty request ceed armenia director require qualification university degree management marketing relate field mba prefer year management experience private sector experience prefer highly develop interpersonal skill networking experience stakeholder develop negotiation persuasion skill fluency english armenian russian language demonstrate experience conduct training technical assistance task small medium enterprise computer literacy strong organizational skill ability meet deadline remuneration salary competitive salary commensurate experience application procedure submit resume cover letter ms word attachment info mention title position apply subject line email interested candidate meet requirement submit application march clearly mention application letter learn job opportunity career center mention url website wwwcareercenteram thank opening date february application deadline march company program cosponsor usaid information visit wwwceedglobalorg place free posting job careerrelate opportunity available organization wwwcareercenteram website follow post announcement link

Discussion: Results & Limitations

Key Findings

The application of LDA revealed distinct topics within the job postings dataset, with a clear demarcation between IT-related and non-IT-related roles.

Limitations

The dataset included labels solely for IT roles, lacking categorization for other job sectors that could have provided deeper insights into industry-specific keywords and features.

Model Performance

The RNN model demonstrated dependable performance in identifying IT roles with 92% accuracy, while the Random Forest model outperformed with a 96% accuracy and 94% macro-average F1 score, indicating its robust classification capabilities.

References

- Kaggle. (n.d.). Armenian online job postings. Retrieved from https://www.kaggle.com/datasets/udacity/armenian-online-job-postings?select=online-job-postings.csv

- GeeksforGeeks. (2024). ML linear discriminant analysis. Retrieved from https://www.geeksforgeeks.org/ml-linear-discriminant-analysis/

- Ribeiro, M. T., Singh, S., & Guestrin, C. (n.d.). “Why should I trust you?” Explaining the predictions of any classifier. Retrieved from https://homes.cs.washington.edu/~marcotcr/blog/lime/

- Soshace. (n.d.). NLP preprocessing using SpaCy. Retrieved from https://soshace.com/nlp-preprocessing-using-spacy/

THANK YOU!