What Job Is This Anyway?

Using LLMs to Classify USAJobs Data Scientist Listings

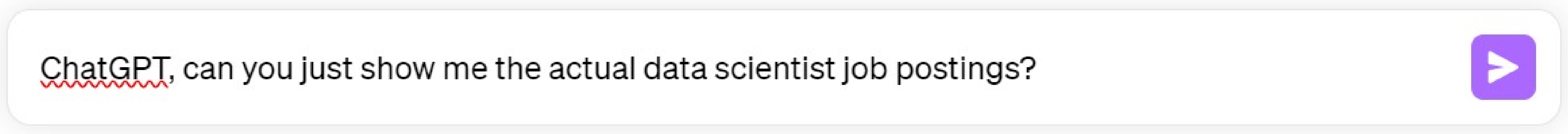
Author: Abigail Haddad Date: October 19, 2023

I'm a data scientist, not a {data analyst, business intelligence analyst, software engineer}

I'm a data scientist, not a {data analyst, business intelligence analyst, software engineer}

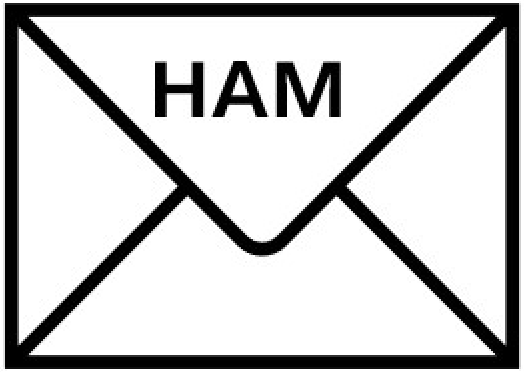
Solution: Text Classification

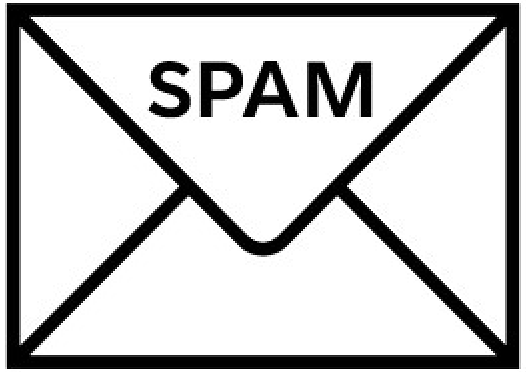
1. We can use LLMs.
2. This can help job seekers....
3. ..and enable internal government analysis and changes.



ChatGPT, can you just show me the actual data scientist job postings?

A detour to discuss classification problems





Assessing classification problems

Confusion Matrix

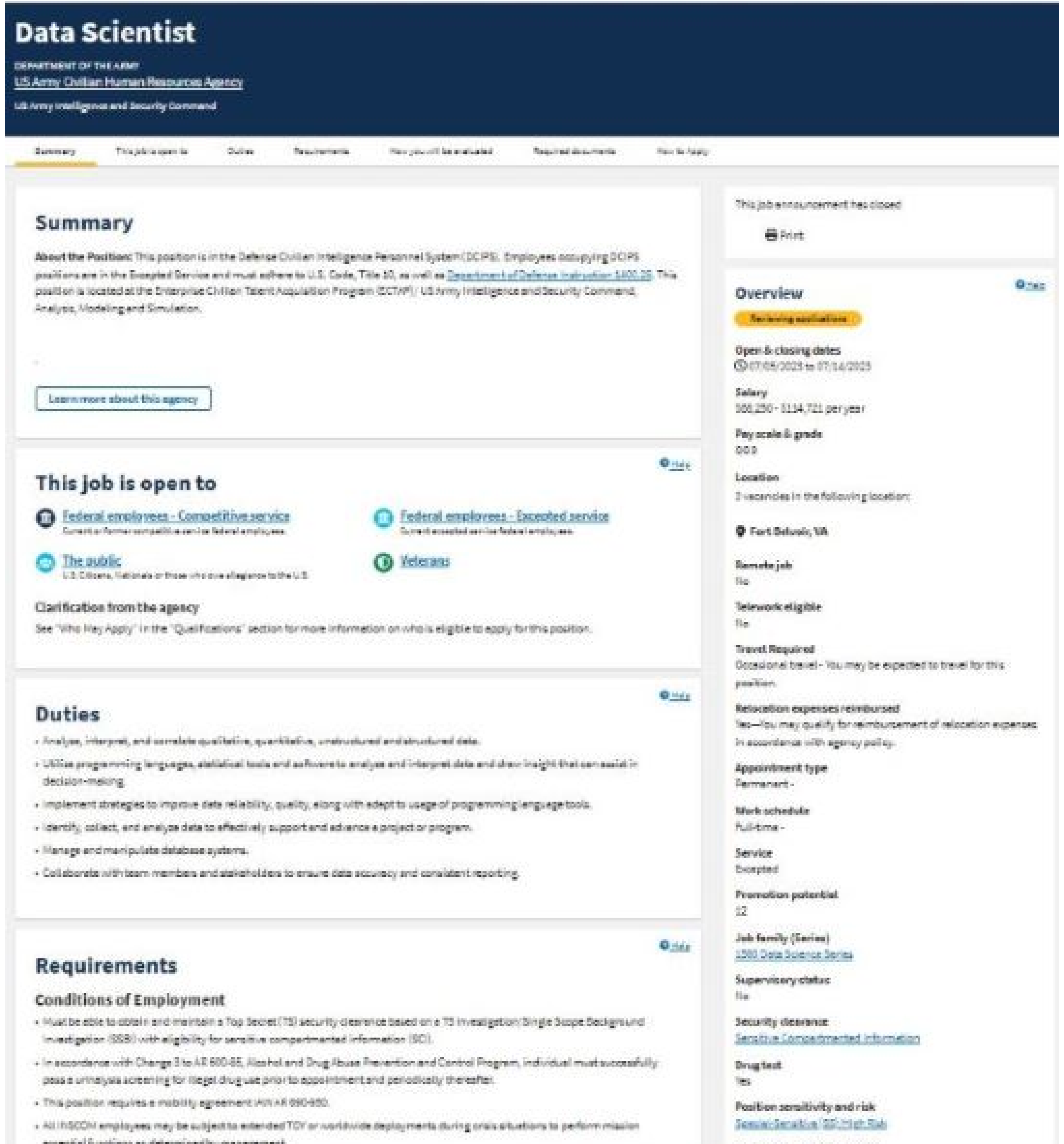
	Predicted Not Spam	Predicted Spam
Actual Not Spam	479	27
Actual Spam	24	470

Assessing classification problems

Derivative metrics

Metric	Value
Recall (True Positive Rate)	95%
Specificity (True Negative Rate)	94%
Precision (Positive Predictive Value)	94%
NPV (Negative Predictive Value)	95%

Back to USAJobs Listings



Data Scientist

DEPARTMENT OF THE ARMY
US Army Civilian Human Resources Agency
US Army Intelligence and Security Command

Summary

About the Position: This position is in the Defense Civilian Intelligence Personnel System (DCIPS). Employees occupying DCIPS positions are in the Excepted Service and must adhere to U.S. Code, Title 10, as well as Department of Defense Instruction 1400.25. This position is located at the Enterprise Civilian Talent Acquisition Program (ECTAP)/ US Army Intelligence and Security Command, Analysis, Modeling and Simulation.

This job is open to

Federal employees - Competitive service
Current or former competitive service federal employees.

Federal employees - Excepted service
Current excepted service federal employees.

The public
U.S. Citizens, Nationals or those who owe allegiance to the U.S.

Veterans

Clarification from the agency
See "Who May Apply" in the "Qualifications" section for more information on who is eligible to apply for this position.

Duties

- Analyze, interpret, and correlate qualitative, quantitative, unstructured and structured data.
- Utilize programming languages, statistical tools and software to analyze and interpret data and draw insight that can assist in decision-making.
- Implement strategies to improve data reliability, quality, along with adept to usage of programming language tools.
- Identify, collect, and analyze data to effectively support and advance a project or program.
- Manage and manipulate database systems.
- Collaborate with team members and stakeholders to ensure data accuracy and consistent reporting.

Requirements

Conditions of Employment

- Must be able to obtain and maintain a Top Secret (TS) security clearance based on a TS Investigation/Single Scope Background Investigation (SSBI) with eligibility for sensitive compartmented information (SCI).
- In accordance with Change 3 to AR 600-85, Alcohol and Drug Abuse Prevention and Control Program, individual must successfully pass a urinalysis screening for illegal drug use prior to appointment and periodically thereafter.
- This position requires a mobility agreement IAW AR 690-950.
- All INSCOM employees may be subject to extended TDY or worldwide deployments during crisis situations to perform mission essential functions as determined by management.

This job announcement has closed

Overview

Reviewing applications

Open & closing dates
07/05/2023 to 07/14/2023

Salary
$88,250 - $114,721 per year

Pay scale & grade
GG 9

Location
2 vacancies in the following location:

Fort Belvoir, VA

Remote job
No

Telework eligible
No

Travel Required
Occasional travel - You may be expected to travel for this position.

Relocation expenses reimbursed
Yes—You may qualify for reimbursement of relocation expenses in accordance with agency policy.

Appointment type
Permanent -

Work schedule
Full-time -

Service
Excepted

Promotion potential
12

Job family (Series)
1560 Data Science Series

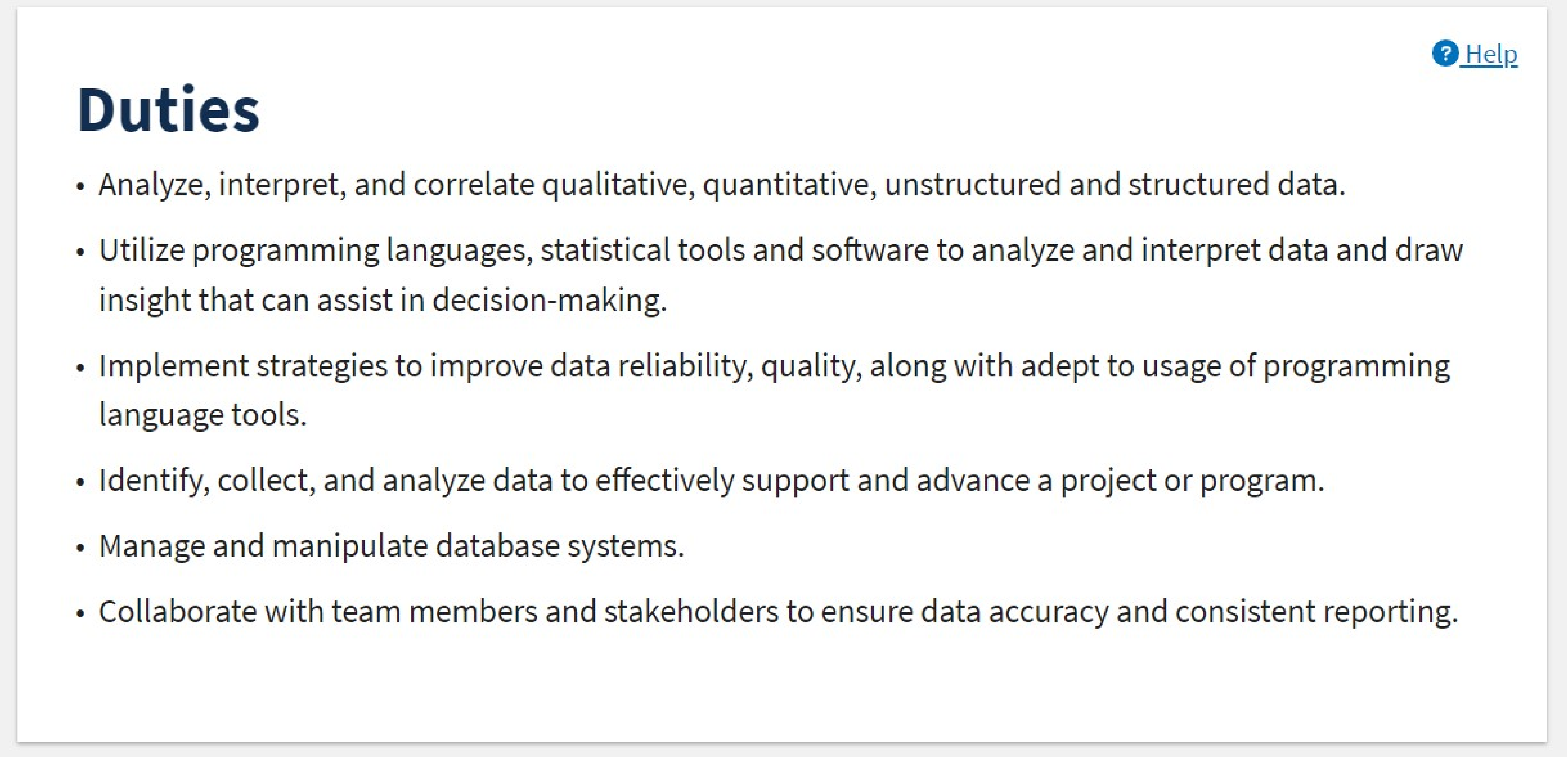
Supervisory status
No

Security clearance
Sensitive Compartmented Information

Drug test
Yes

Position sensitivity and risk
Special-Sensitive (SS)/High Risk

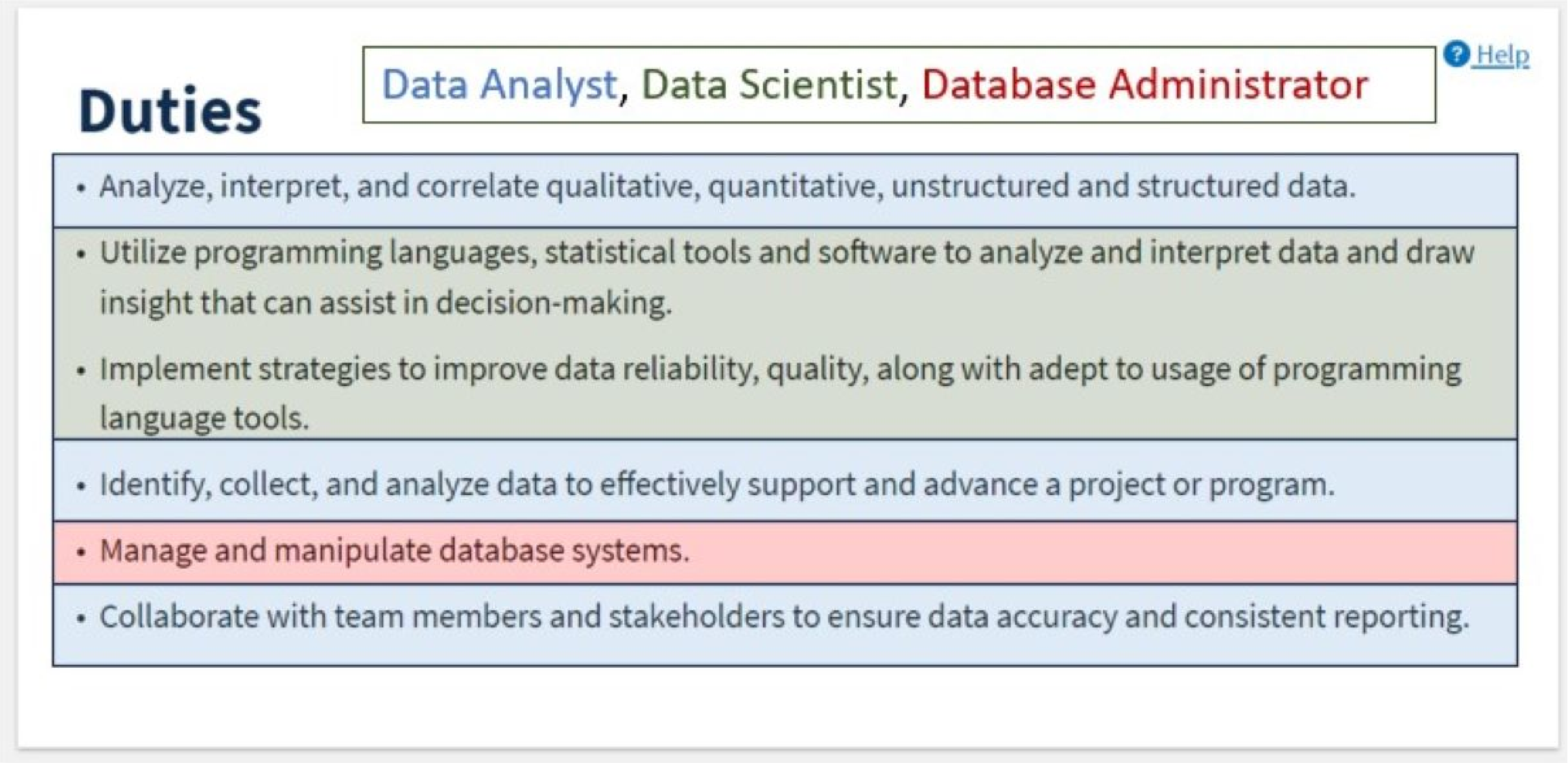
Example 'Duties' Text



Duties

- Analyze, interpret, and correlate qualitative, quantitative, unstructured and structured data.
- Utilize programming languages, statistical tools and software to analyze and interpret data and draw insight that can assist in decision-making.
- Implement strategies to improve data reliability, quality, along with adept to usage of programming language tools.
- Identify, collect, and analyze data to effectively support and advance a project or program.
- Manage and manipulate database systems.
- Collaborate with team members and stakeholders to ensure data accuracy and consistent reporting.

Example 'Duties' Text with Highlights



Duties

Data Analyst, Data Scientist, Database Administrator

- Analyze, interpret, and correlate qualitative, quantitative, unstructured and structured data.
- Utilize programming languages, statistical tools and software to analyze and interpret data and draw insight that can assist in decision-making.
- Implement strategies to improve data reliability, quality, along with adept to usage of programming language tools.
- Identify, collect, and analyze data to effectively support and advance a project or program.
- Manage and manipulate database systems.
- Collaborate with team members and stakeholders to ensure data accuracy and consistent reporting.

We can't overcome text ambiguity problems with LLMs. If the information isn't there, it's not there.

My MVP Project Workflow

1. Pulled 843 1560 Data Scientist job listings.
2. Dropped those with no duty list via length/keyword filter.
3. Multi-label classification via GPT-3.5 3a. Assess consistency scores/inter-rater reliability
4. Analysis of results 4a. BERT encoding of labels, clustering. 4b. GPT labels of clusters 4c. Word cloud 4d. Ad hoc validation of results

What can we actually get from this?

What can we actually get from this?

- A view of the variety of different roles that are under 1560
- A prototype job labeling system for applicants
- A way to highlight jobs that might be mislabeled or challenging to hire for
- A variable that might be predictive for research and analysis

What this can't be

- The one (or more) "true" label
- A solution to ambiguous duties sections

Machine Learning Engineer
Policy Analyst
Systems Analyst
Artificial Intelligence Engineer
Test Engineer
Data Governance Specialist
Supervisory Personnel Management
Senior Manager
Supervisor
Data Analyst
Remote Sensing Specialist
Human Resources Specialist
Data Management Division
Data Transformation Agent
Senior Data Scientist
Assist On Project Planning and Development
Health Data Analyst
Cybersecurity Analyst
Strategic Planner
Hr Analyst
Database Developer
Branch Chief
Technical Expert
Intelligence Analyst
Mathematical Statistician
Statistician
Computer Scientist
Operations Research Analyst
Artificial Intelligence Specialist
Data Engineer
Natural Language Processing Specialist
Environmental Scientist
Database Administrator
Geospatial Analyst
Chief Data Officer
Test And Evaluation Analyst
Program Analyst
Information Technology Specialist
Scientist
Business Intelligence Analyst
Data Mining Specialist
Risk Analyst
Operations Manager
Data Architect
Statistical Analyst
Financial Analyst
Data Manager
Business Analyst
Data Strategist
Software Engineer
Systems Engineer
Supervisory Data Scientist
Data Management Specialist
Epidemiologist
Data Governance Manager
Healthcare Analyst
Chief Data Scientist
Programmer
Data Scientist
Management Analyst
Data Science Expert
Subject Matter Expert
Statistical Data Analyst
It Project Manager
Data Visualization Specialist
Economist
Research Scientist
Public Health Analyst
Gis Analyst
Operational Analyst
Project Manager
Research Analyst
Researcher
Software Developer
Quantitative Analyst
Program Manager
Technical Advisor
Emergency Management Specialist
Computational Scientist

Big range of job title labels!

Cluster	%	Top 5 Titles
Data & Project Management	34	Data Scientist, Project Manager, Data Analyst, Program Manager, Supervisor
Core Data Science & Engineering	22	Data Scientist, Data Analyst, Machine Learning Engineer, Data Engineer, Statistical Analyst
Data Science & Research	17	Data Scientist, Data Analyst, Research Scientist, Data Engineer, Database Administrator
Data Strategy & Specialization	16	Data Analyst, Data Strategist, Geospatial Analyst, Supervisory Data Scientist, Chief Data Scientist
Data Analysis & Research Intelligence	10	Data Analyst, Data Scientist, Business Intelligence Analyst, Statistical Analyst, Research Analyst

Spot Checking Title Sets

Data Scientist, Data Analyst, Software Engineer

center

Data Scientist, Data Analyst, Business Intelligence Analyst, Machine Learning Engineer, Artificial Intelligence Specialist, Data Engineer

center

Data Analyst, Data Architect, Data Scientist, Data

What's better than ad hoc assessment?

Systematic assessment!

Derivative Metrics Revisited

Metric Name	Acronym	Definition
Recall	TPR	The percentage of actual positives that the model correctly identified.
Specificity	TNR	The percentage of actual negatives that the model correctly identified.
Precision	PPV	Out of all the instances labeled as positives by the model, the percentage that are actual positives.
Negative Predictive Value	NPV	Out of all the instances labeled as negatives by the model, the percentage that are actual negatives.

Assessment Example with Synthetic Data

Label	TPR (Recall)	TNR (Specificity)	PPV (Precision)	NPV
Data Scientist	23%	90%	26%	88%
Project Manager	33%	91%	35%	90%
Data Analyst	32%	90%	33%	90%
Program Manager	24%	90%	25%	90%
Supervisor	28%	90%	27%	90%

But also, are current labels perfect?

What Can We Do With This?

- If you're a federal job seeker or researcher/analyst, let's talk!
- You might have a text classification problem -- and it might be multi-label.
- Importance of structured outputs and assessment.

Necessary Technical Improvements

- Labeling/assessment
- Revisit clustering
- More structured LLM responses (possibly with the Marvin library)