

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, program manager}

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

Solution: Text Classification

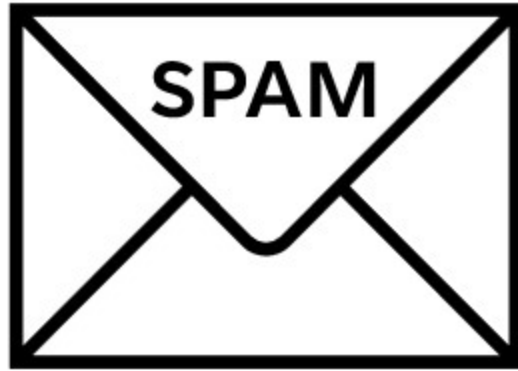
1. We can use LLMs.
2. This can help job seekers....
3. ..and enable internal government analysis and changes.



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

Back to USAJobs Listings

Data Scientist

DEPARTMENT OF THE ARMY
[US Army Civilian Human Resources Agency](#)
[US Army Intelligence and Security Command](#)

[Summary](#) [This job is open to](#) [Duties](#) [Requirements](#) [How you will be evaluated](#) [Required documents](#) [How to apply](#)

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 50, as well as [Department of Defense Instruction 5400.05](#). This position is located at the Enterprise Civilian Talent Acquisition Program (ECTAP), US Army Intelligence and Security Command, Analysis, Modeling and Simulation.

[Learn more about this agency](#)

This job is open to

 **Federal employees - Connected service**
Current or former employees who are 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 (AR 600-950).
- 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.

 Print

Overview

[Review qualifications](#)

Open & closing dates

07/05/2023 to 07/14/2023

Salary

\$68,290 - \$134,721 per year

Pay scale & grade

GS-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)

[1580 Data Science Series](#)

Supervisory status

No

Security clearance

[Sensitive Compartmented Information](#)

Drug test

Yes

Position sensitivity and risk

[Sensitive Security Information \(SSI\)](#)


Trust/declassification concerns

Example 'Duties' Text

Duties

- Work in multiple database environments using variety of programming languages. Recommend improvements. Maintain and review codebase. Collaborate with vendors, clients, and analysts.
- Assess, plan, and maintain data optimized for business and security requirements. Assess requirements and apply relevant techniques to create robust architectures. Maintain databases to prevent disruptions.
- Maintain data sources to extract, transform, and load data from any data source. Construct pipelines that support data mining and data science projects. Optimize large-scale processing systems for performance and scalability.
- Identify opportunities for enhance business and security requirements. Design and implement processes that improve quality, automate tasks and develop standard practices. Identify performance issues.
- Improve the OIG's use of data and analysis. Investigate new technologies and practices where relevant. Dive into data and pinpoint tasks to eliminate manual participation with automation using Cloudera NiFi.
- Recommend new software solutions. Apply a variety of programming languages to meet software and business needs. Understand data warehousing platforms, big data software, and project management applications.

This position is officially titled "Data Scientist"

- My model accurately identified it "Data Engineer"
- Also, this position is misclassified! 
- This is the kind of analysis we can do with this model

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

I'm not going to say who did this

 [Help](#)

Duties

- Collect and analyze data from a wide variety of sources.
- Produce and evaluate statistical data for quality and consistency.
- Communicate economic and statistical information in writing and in person.

My MVP Project Workflow

1. Pulled ~850 1560 Data Scientist job listings.
2. Dropped those with no duty list via length/keyword filter.
3. Used the Marvin library to get structured output from GPT 3.5 for several problems, including the multi-class classification problem.
4. Ad hoc analysis, iteration

The only (Python) code you need

Or why you should use Marvin (even if you're working in R)

```
from marvin import ai_fn

@ai_fn
def generate_job_title(duties: str) -> str:
    """Given `duties`, generates a specific job title based on the content,
    not based on any titles contained in the text."""

class JobAnalyzer:
    def __init__(self, duties: str):
        self.duties = duties
        self.job_title = generate_job_title(self.duties)
```

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, challenging to hire for, or misclassified
- A variable that might be predictive for research and analysis

This is not the one "true" label 

What Did We Get?

1. Overall differences in top titles
2. Job titles that never appear in the official title field
3. Spot checks

Comparing Top 10 Job Titles -- LLM vs. Official

LLM-Generated Titles

Data Scientist

Data Analyst

Senior Data Scientist

Data Analytics Specialist

Data Science Analyst

Data Science Specialist

Supervisory Data Scientist

Chief Data Scientist

Chief Data Officer

Data Science Manager

Official Titles

Data Scientist

Interdisciplinary

Supervisory Data Scientist

Engineer/Scientist

Health Scientist (Data Science)/Physical Scientist (Data Science)/Data Scientist

Supervisory Interdisciplinary

Manager

Interdisciplinary Health Scientist (Data Science)/Physical Scientist (Data Science)/Data Scientist

Interdisciplinary Data Scientist

Lead Data Scientist

New Job Titles

Some of these are descriptive 🌟🌟🌟



Examples of big title discrepancies

Generated Title: Data Analyst

Data Scientist

 [Help](#)

Duties

- Develops standards for analyzing, testing, and assessing emerging data analytics technologies and methods.
- Works with large and complex data sets and performing extensive analysis of volumes of data through macros, pivot tables, etc.
- Designs and delivers frameworks to facilitate the data development lifecycle.
- Designs innovative approaches to complex data analytics projects utilizing various mathematical, statistical and other scientific methodologies, programs, processes, and methods.
- Develops standards for analyzing, testing, and assessing emerging data analytics technologies.
- Provides information regarding data analytics projects, studies, solutions, technologies, deliverables and challenges to Senior Leadership and Technical Organizations.

Generated Title: Technical Program Manager



Data Scientist

 [Help](#)

Duties

- Provide technical leadership to a highly technical professional staff, supported by policy and data analytic experience.
- Conduct special assignments involving sensitive and/or confidential technical program related issues that have significant impact on models, initiatives, and CMS programs.
- Report orally or prepares written/visual reports of findings, actions taken, or recommendations for the Division Director's attention.
- Participate in, and makes substantive contributions to top level discussions and planning sessions on management program policy development

What's better than ad hoc assessment?

 **Systematic assessment!** 

I tried a lot of other things, too

- Named entity recognition for finding tools and software 🚀
- Grouping by category (BI intelligence, management, etc.) 🧑
- Comparing actual titles and generated titles 👎

What's next?

- If you're a federal job seeker or researcher/analyst, let's talk!
- If you have a text classification problem you might like LLMs (and Marvin)