BCSE355L - AWS Solutions Architect

SKILL EXTRACTION FROM RESUMES USING CUSTOM ENTITY RECOGNITION

Submitted By

BRINDHA LN 22BAI1433

POOJA SASIKUMAR 22BAI1437

VIJITA MOHANRAAJ 22BAI1452

B. Tech Computer Science and Engineering (AI&ML)

Submitted to

Dr. Ashoka Rajan R

(SCOPE, VIT CHENNAI)

VIT

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

VIT Chennai

Vandalur - Kelambakkam Road, Chennai - 600127

Introduction

This project aims to develop a Custom Entity Recognition Model using Amazon Comprehend, designed to automate the extraction of specific entity, skills, from resumes. In today's fast-paced recruitment landscape, organizations deal with a large volume of resumes daily, making manual parsing both time-consuming and error-prone. By leveraging Natural Language Processing (NLP) and machine learning, this project addresses the need for a scalable and accurate solution for resume analysis.

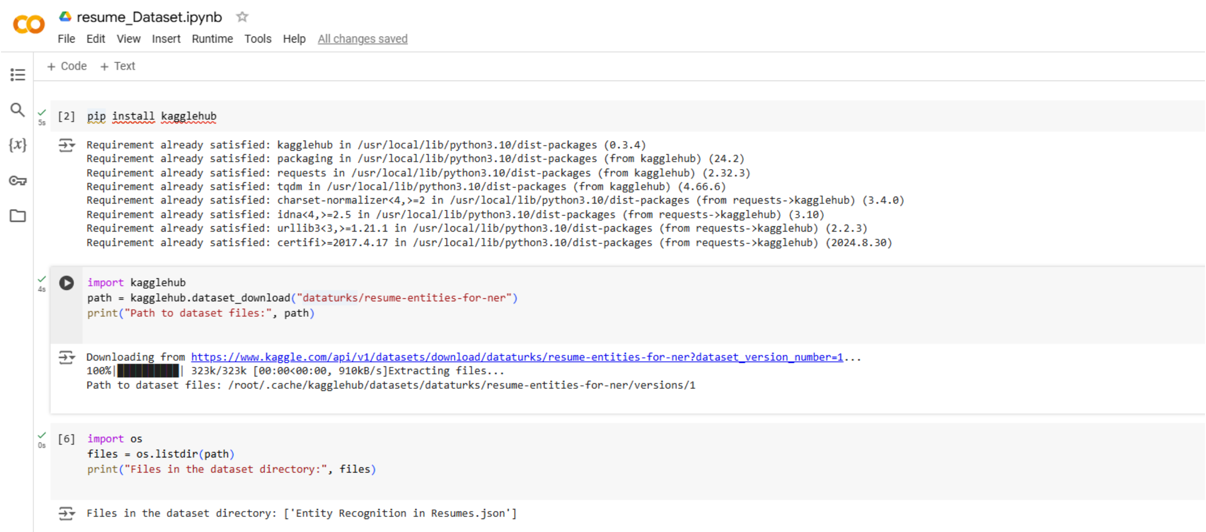
The workflow starts with the training of a custom entity recognizer tailored to identify predefined entities like programming languages, tools, certifications, and soft skills. Training datasets were uploaded to Amazon S3 and used to create and train the entity recognizer model in Amazon Comprehend. Amazon Comprehend, a fully managed NLP service, uses machine learning to find insights and relationships in a text, making it ideal for extracting specific entities like skills from resumes.

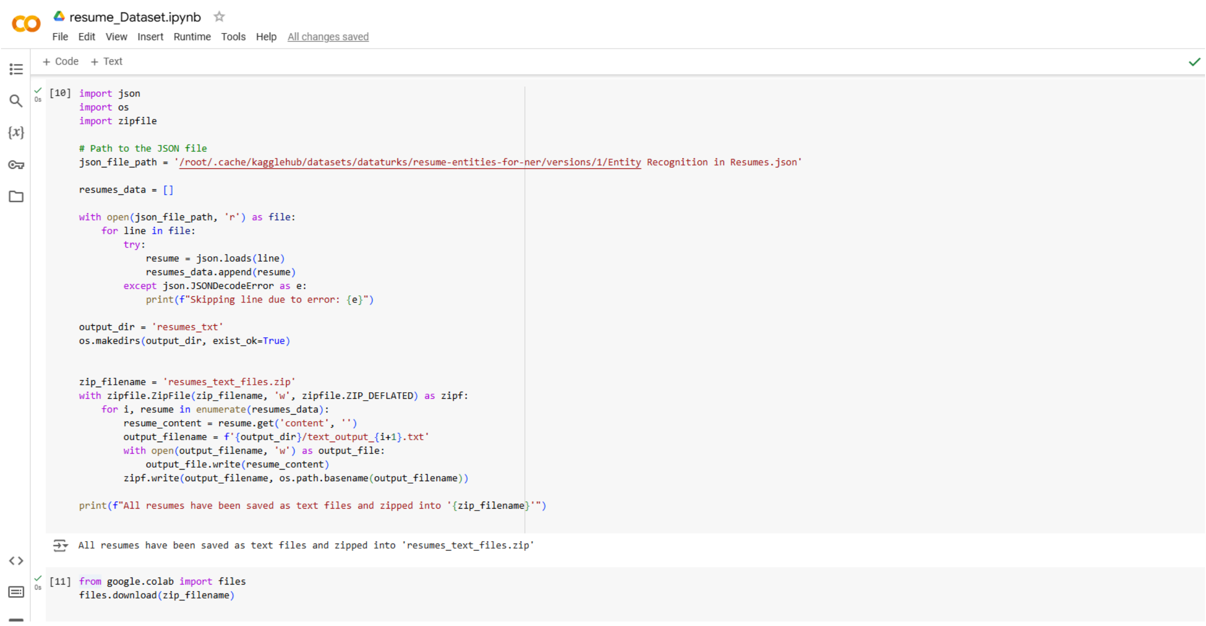
By automating the extraction process, the project significantly reduces human effort, increases efficiency, and ensures consistency in entity recognition. This project showcases the potential of cloud-based NLP tools and machine learning models in solving real-world challenges, particularly in recruitment and talent acquisition workflows. The integration of Amazon Comprehend and S3 for data storage provides a scalable, flexible, and reliable framework that can be adapted for various domains requiring custom entity recognition.

Resume Dataset

Code for Downloading and Converting Dataset to Text Files:

In this code, the "Resume Entities for NER" dataset is downloaded from Kaggle. This dataset is in JSON format, so the data is converted into individual text files. These text files are then compressed into an archive, which can be used for training purposes. Then, the archive containing the text files is downloaded.





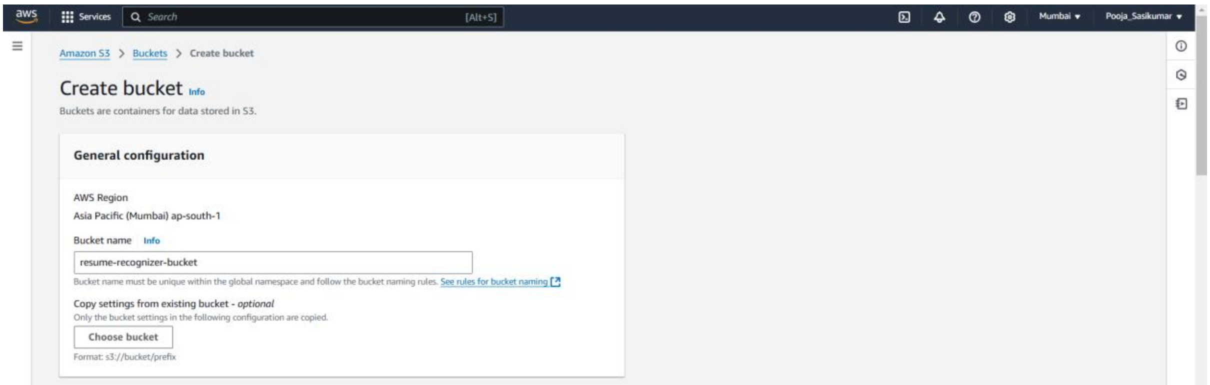
The code in the Python notebook implementation is given below:

```
pip install kagglehub
import kagglehub
path = kagglehub.dataset_download("dataturks/resume-entities-for-ner")
print("Path to dataset files:", path)
import os
files = os.listdir(path)
print("Files in the dataset directory:", files)
import json
import os
import zipfile

json_file_path = '/root/.cache/kagglehub/datasets/dataturks/resume-
entities-for-ner/versions/1/Entity Recognition in Resumes.json'

resumes_data = []

with open(json_file_path, 'r') as file:
    for line in file:
        try:
            resume = json.loads(line)
            resumes_data.append(resume)
        except json.JSONDecodeError as e:
            print(f"Skipping line due to error: {e}")

output_dir = 'resumes_txt'
os.makedirs(output_dir, exist_ok=True)

zip_filename = 'resumes_text_files.zip'
with zipfile.ZipFile(zip_filename, 'w', zipfile.ZIP_DEFLATED) as zipf:
    for i, resume in enumerate(resumes_data):
```

```
        resume_content = resume.get('content', '')
        output_filename = f'{output_dir}/text_output_{i+1}.txt'
        with open(output_filename, 'w') as output_file:
            output_file.write(resume_content)
        zipf.write(output_filename, os.path.basename(output_filename))

print(f"All resumes have been saved as text files and zipped into
'{zip_filename}'")
from google.colab import files
files.download(zip_filename)
```

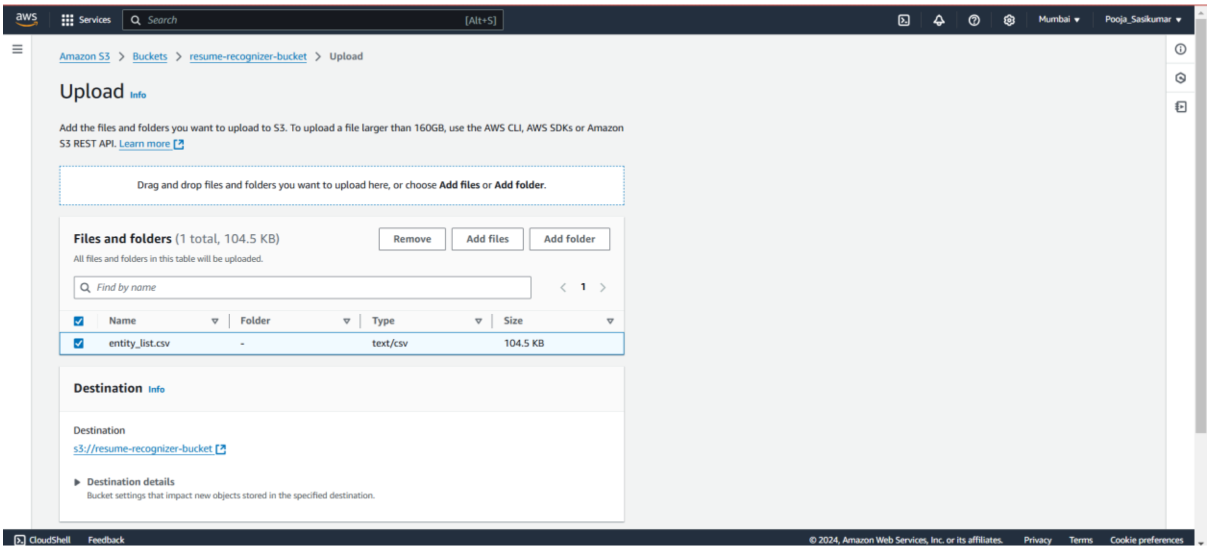
Setting Up S3 Bucket and Uploading Training Data

Create an S3 bucket named resume-recognizer-bucket to store training datasets and test files.

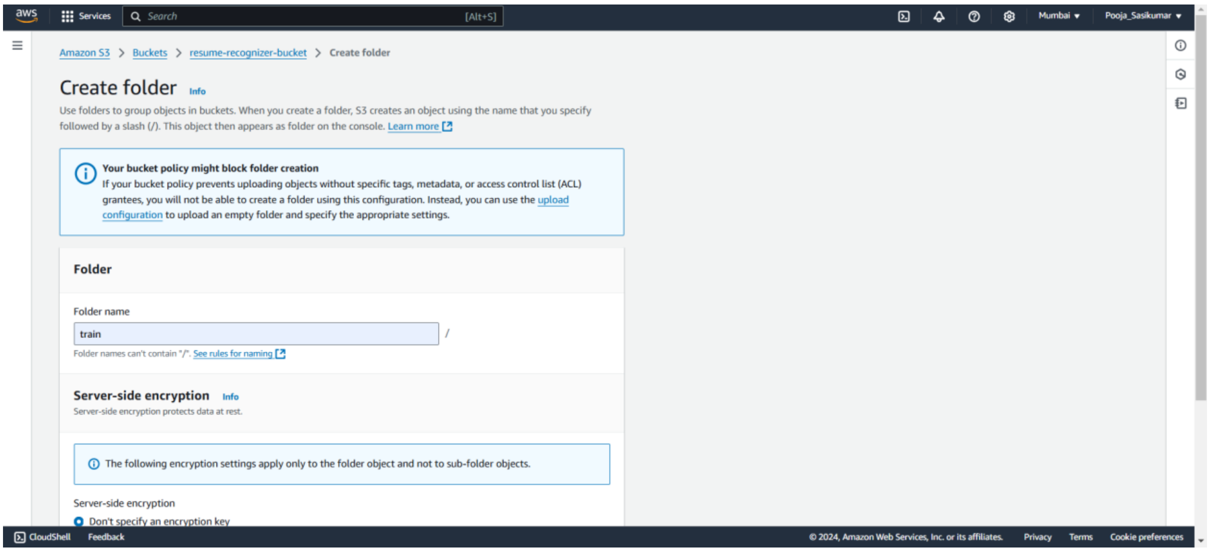


Upload the entity_list.csv file to the root of the S3 bucket.

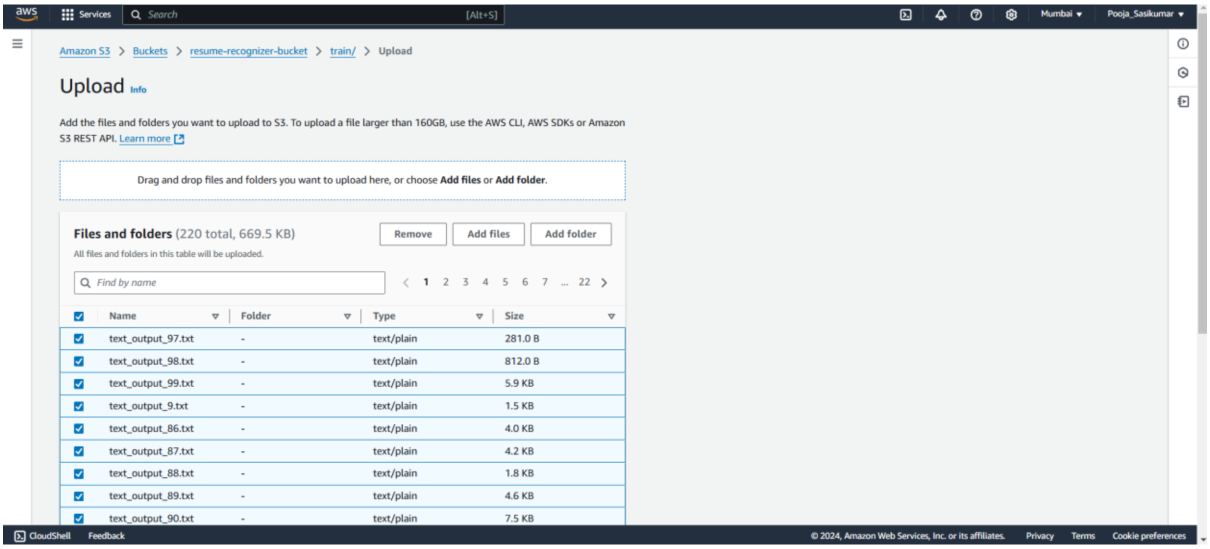
The entity list is a CSV file, with two columns: text and type, where the text column contains examples of skills (e.g., "Java", "Python", "Database Management System," "C++," and "Oracle PeopleSoft.") and the type column contains the label "SKILLS".



Create a folder named train/ within the S3 bucket to store the training data



Upload all training text documents extracted by the Python code to the train folder. These files are resume files used for training the custom entity recognition model.

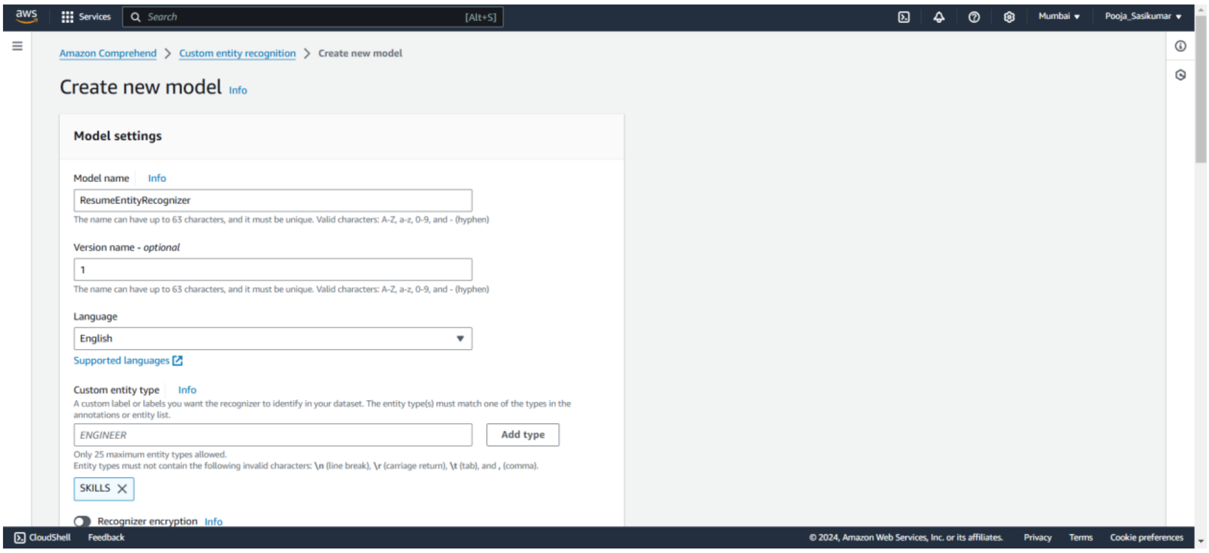


Custom Entity Recognition Model Training

We will create a custom entity recognizer to extract skills from resumes.

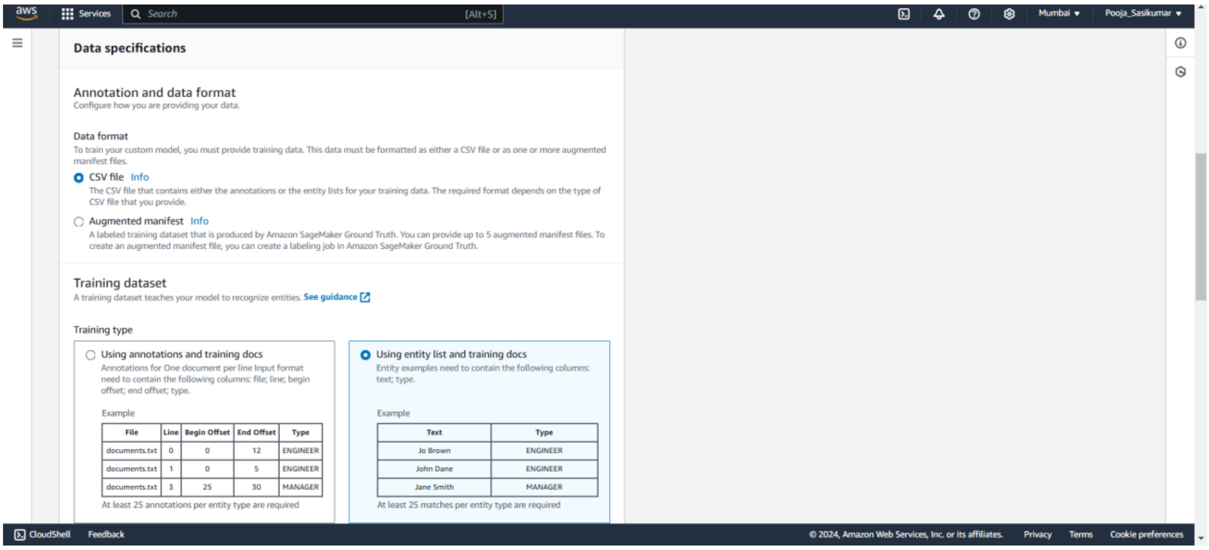
In **AWS Comprehend Console**: Navigate to **Customization** → **Custom entity recognition** → **Create new model.**

The model is named "ResumeEntityRecognizer" and the entity type "SKILLS" is specified to be recognized within the dataset.



To train a custom entity recognition model, data can be provided to Amazon Comprehend using one of two methods: Annotations or Entity lists.

We will use the entity list method.

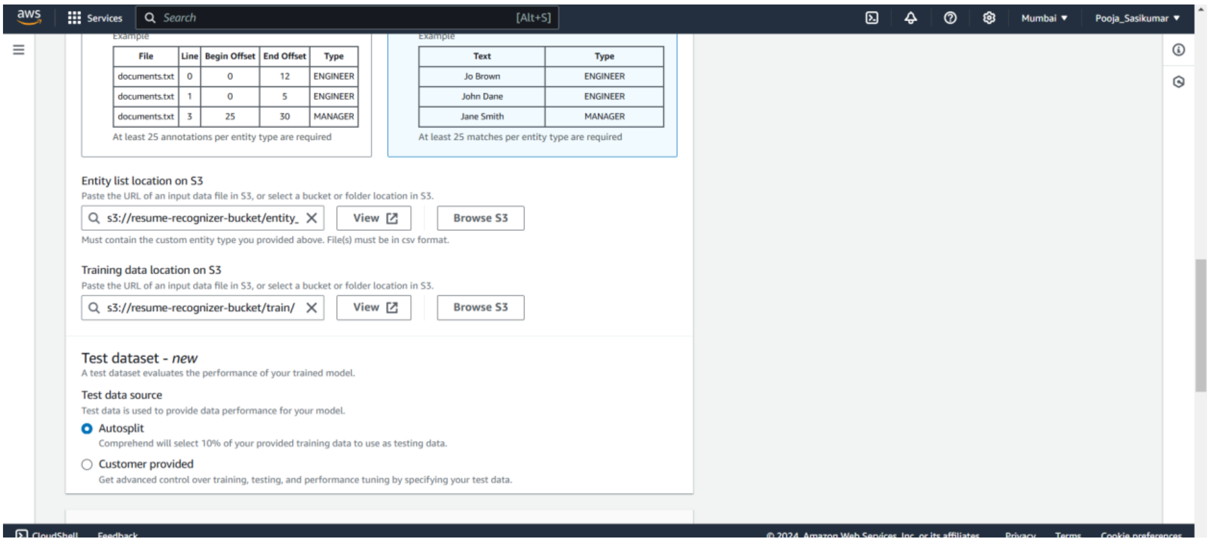


The S3 URLs of the entity list and training data which were uploaded to Amazon S3 are provided.

S3 URL for the entity list file: s3://resume-recognizer-bucket/entity_list.csv

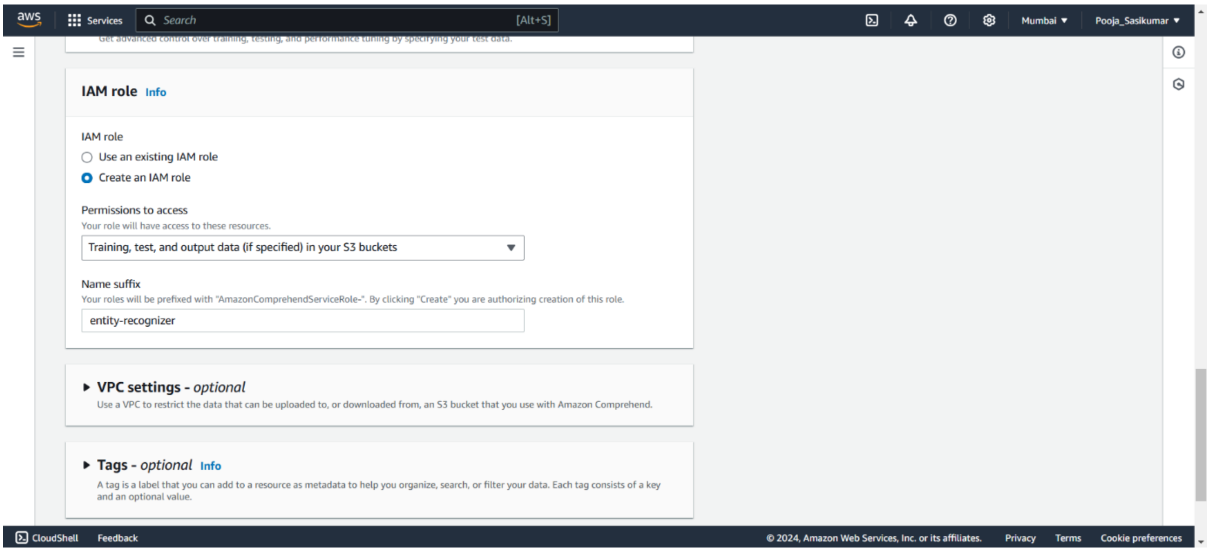
S3 URL for the training data file: s3://resume-recognizer-bucket/train/

Then, the "Autosplit" option is selected, and AWS Comprehend automatically selects 10% of the provided training data for testing.



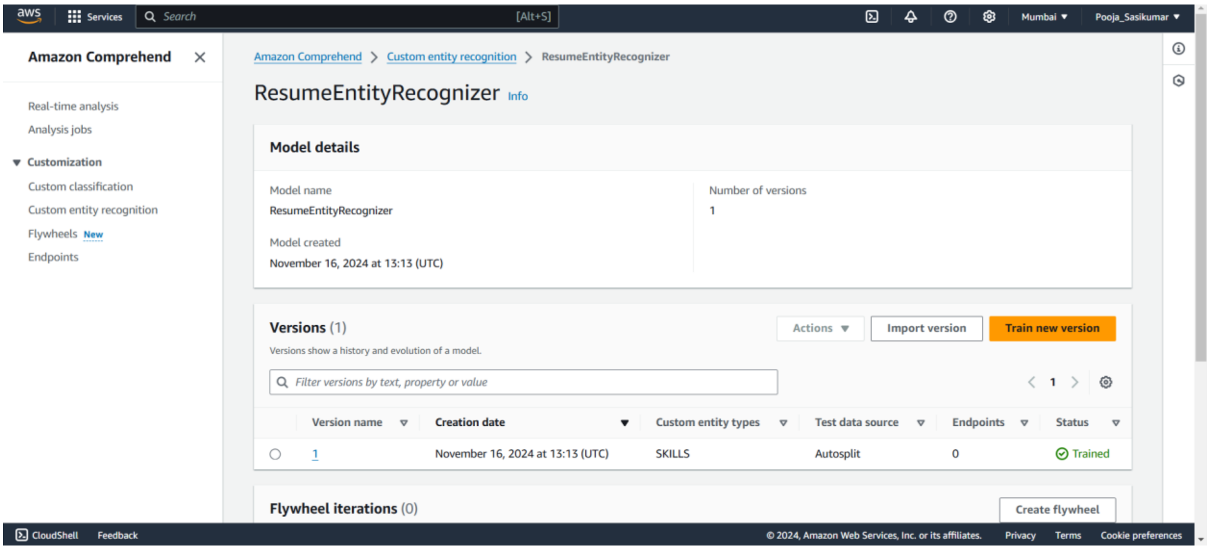
An IAM role named "entity-recognizer" is created to access the training, test, and output data stored in your S3 buckets.

This ensures the role can interact with all necessary resources for the model's training and evaluation.



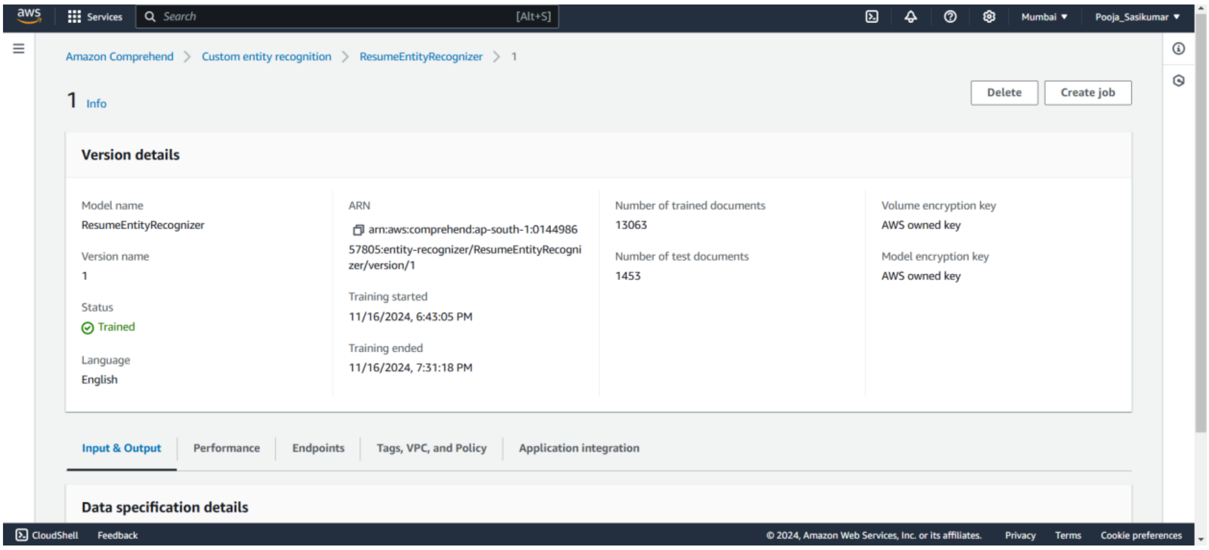
The custom entity recognizer is created, and the training process begins. The training process involves the use of labeled data to teach the model how to recognize and classify entities accurately.

It took 48 minutes to train the model.



The version details of the custom entity recognizer include the model name, version, language, and status, which is shown as "Trained."

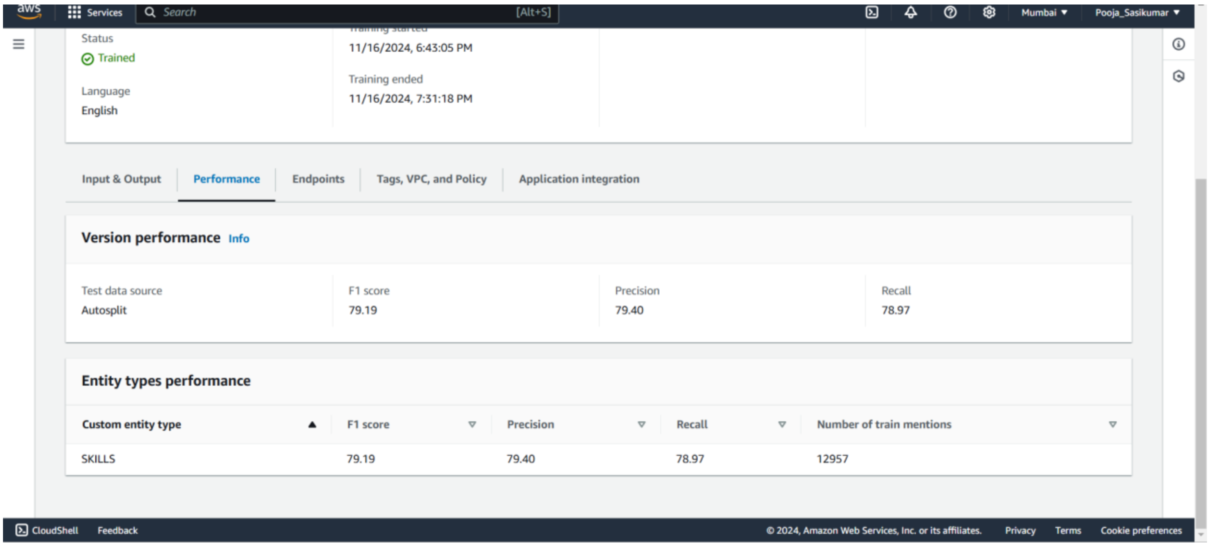
A total of 13,063 documents were used for training, and 1,453 documents were used for testing.



Performance Evaluation Metrics

The performance evaluation of the custom entity recognizer provides key metrics to assess its accuracy and effectiveness. The performance metrics for the custom entity recognizer evaluating the "SKILLS" entity type are as follows:

Metric	Value	Definition
Precision	**79.40**	**Positive predictive value**
Recall	**78.97**	**True positive rate**
F-1 Score	**79.19**	**Harmonic mean of the precision and recall**



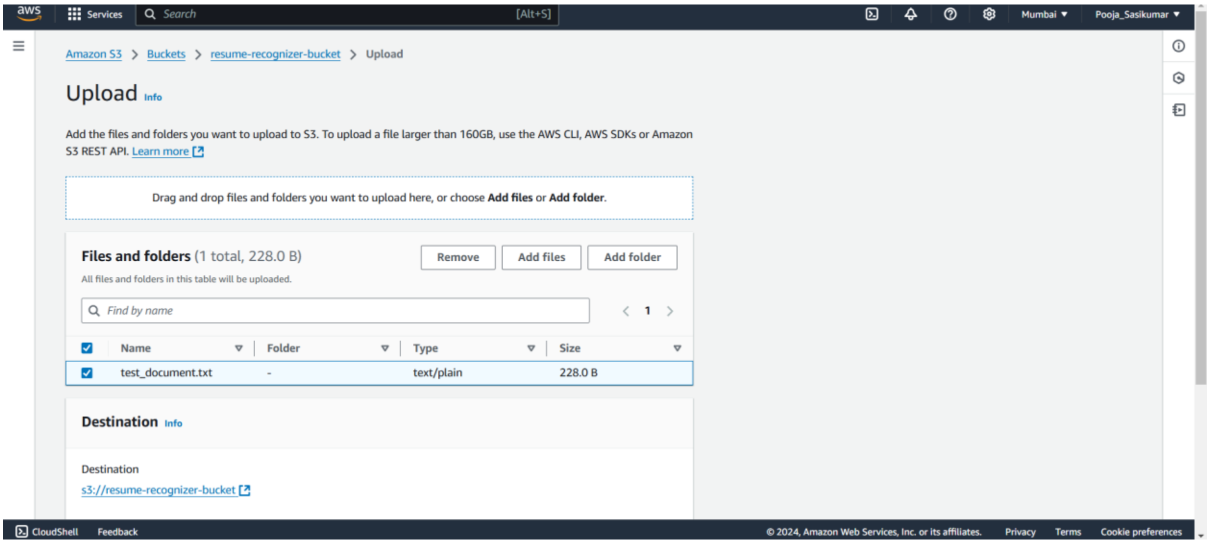
Testing Using Analysis Job

Once the custom entity recognizer is trained, an analysis job is created to test its performance on unseen data.

test_document has the following resume as text:

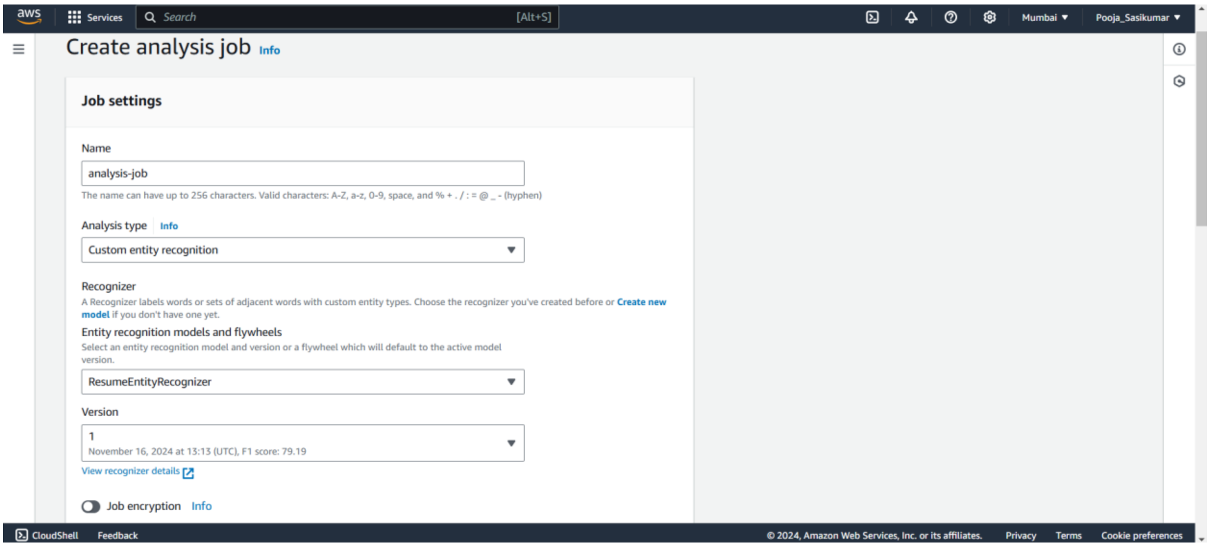
```
Tom Jackson

Skill Summary:
- Strong analytical and problem solving skills
- Holds AWS Certified Associated Solution Architect Certification
- Databases: MySQL, SQL, Oracle
- Programming Languages: C, C++, Java, PHP, JavaScript
```

To perform testing, upload the test document to the resume-recognizer S3 bucket to be used in the analysis job.



In the AWS Comprehend Console: Navigate to ***Analysis jobs → Create job*** to initiate the testing process for the custom entity recognizer.

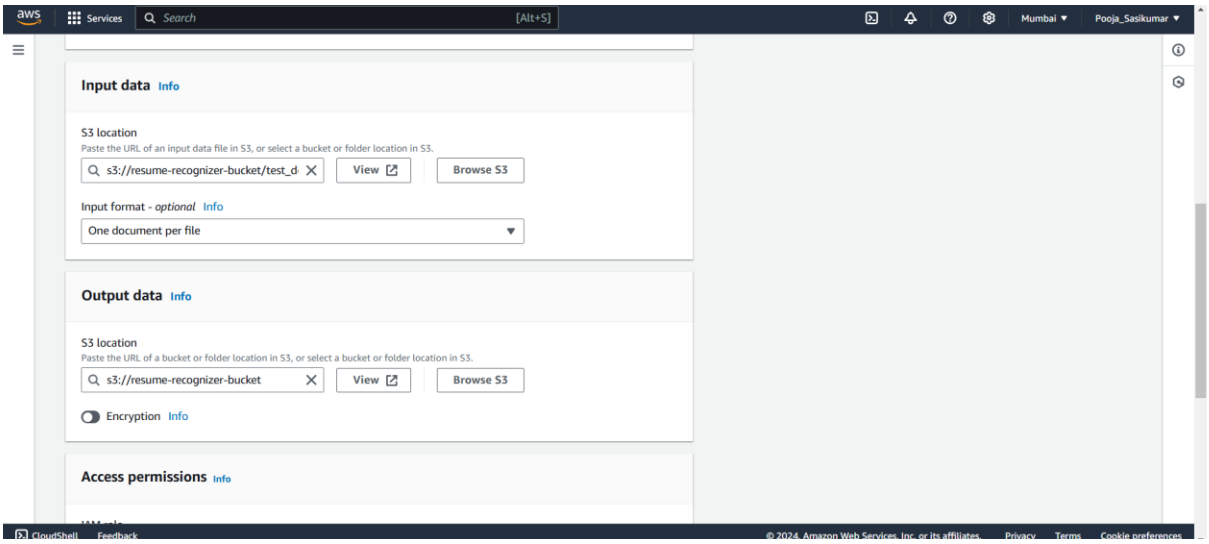
The analysis job is named analysis-job, with the type set to Custom entity recognition. The created recognizer ResumeEntityRecognizer (version 1) is selected for labeling entities in the input data.



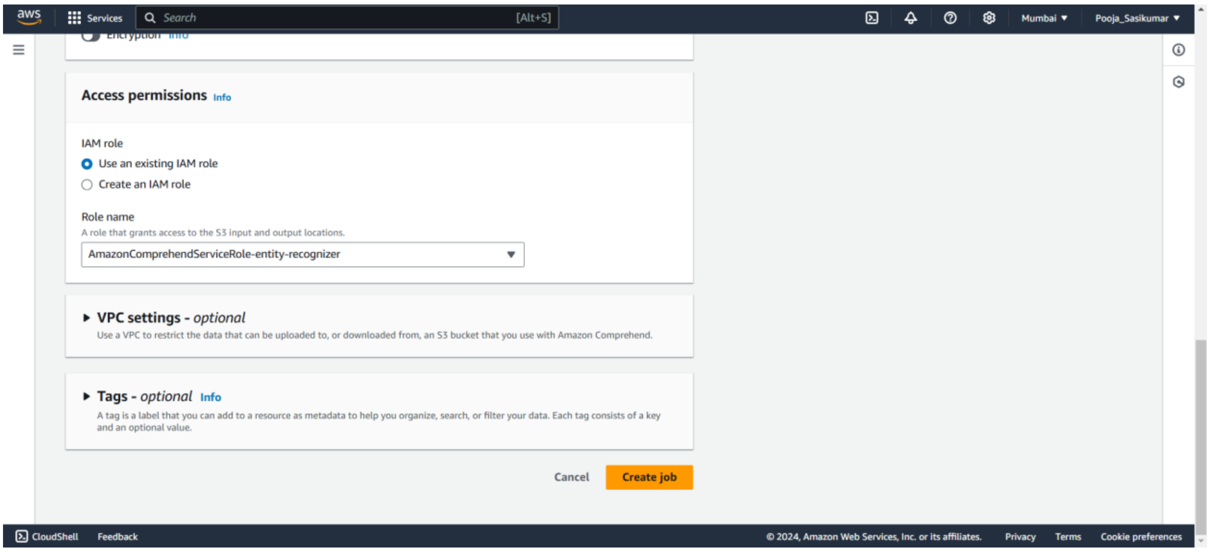
The input data for the analysis job is located at s3://resume-recognizer bucket/test_document.txt.

The test data includes resumes containing text where the model is expected to recognize skills as entities.

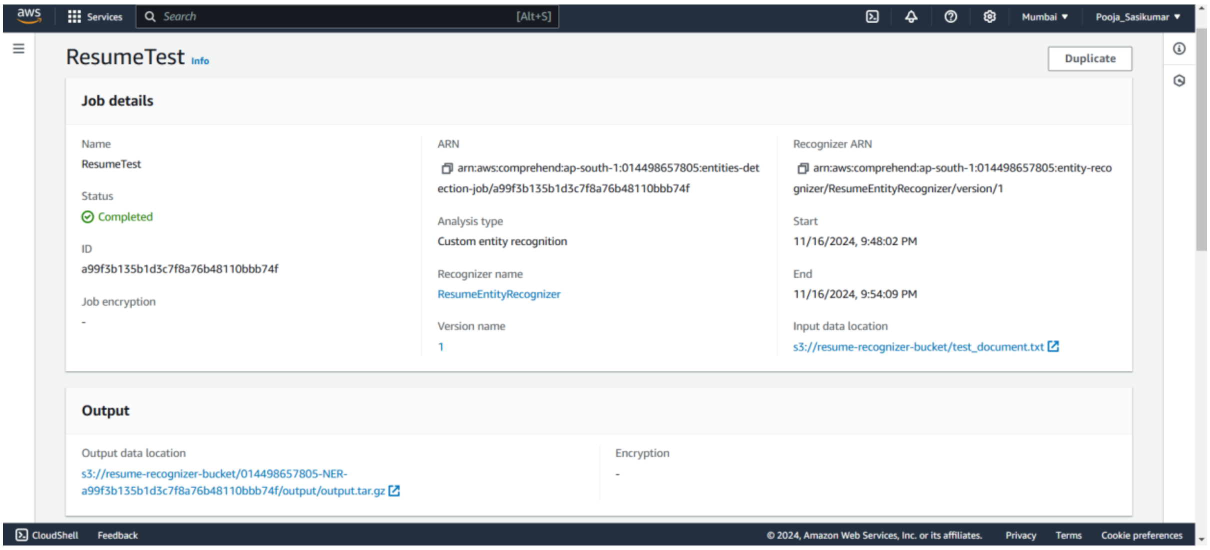
The output data will be saved to the S3 location s3://resume-recognizer-bucket, ensuring proper storage of the analysis results.



The IAM role "AmazonComprehendServiceRole-entity-recognizer," created previously during the model creation, is used for the analysis job. This role grants the necessary permissions to access the S3 input and output locations for the job.

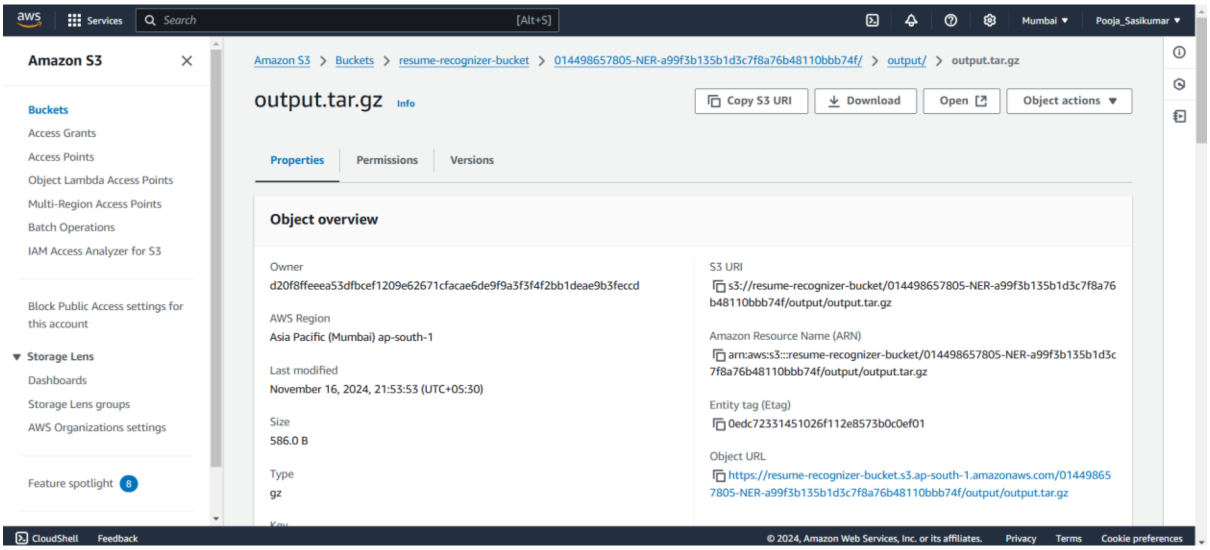


The analysis job processes the test data using the trained recognizer and generates the results. Once the job reaches the "COMPLETED" status, Amazon Comprehend will save the results as JSON files in the specified output S3 bucket.



The output of the analysis job is stored in the S3 bucket as an archive file named `output.tar.gz`. This archive contains the JSON files with the results of testing the `test_document`.

We can download this archive from the S3 bucket and extract the JSON files to review the results.



Test Results of Entity Recognition Model

The output consists of a JSON file containing the recognized skills, including their type, start, and end positions within the text. The result for test document are as follows:


```
{
  "Entities": [
    {
      "BeginOffset": 37,
      "EndOffset": 67,
      "Score": 0.7749637708158428,
      "Text": "analytical and problem solving",
      "Type": "SKILLS"
    },
    {
      "BeginOffset": 83,
      "EndOffset": 86,
      "Score": 0.9999525569523512,
      "Text": "AWS",
      "Type": "SKILLS"
    },
    {
      "BeginOffset": 108,
      "EndOffset": 116,
      "Score": 0.9114705487241651,
      "Text": "Solution",
      "Type": "SKILLS"
    },
    {
      "BeginOffset": 161,
      "EndOffset": 164,
      "Score": 0.9999367038714213,
      "Text": "SQL",
      "Type": "SKILLS"
    },
    {
      "BeginOffset": 166,
      "EndOffset": 172,
      "Score": 0.9997155283789972,
      "Text": "Oracle",
      "Type": "SKILLS"
    },
    {
      "BeginOffset": 175,
      "EndOffset": 186,
      "Score": 0.9948563676923424,
      "Text": "Programming",
      "Type": "SKILLS"
    },
    {
      "BeginOffset": 198,
      "EndOffset": 200,
      "Score": 0.9999510079565681,
      "Text": "C,",
      "Type": "SKILLS"
    },
    {
      "BeginOffset": 201,
      "EndOffset": 205,
      "Score": 0.9999717781772669,
      "Text": "C++,",
      "Type": "SKILLS"
    },
```

```
  {
    "BeginOffset": 206,
    "EndOffset": 210,
    "Score": 0.9999535105382235,
    "Text": "Java",
    "Type": "SKILLS"
  },
  {
    "BeginOffset": 212,
    "EndOffset": 215,
    "Score": 0.999892365596196,
    "Text": "PHP",
    "Type": "SKILLS"
  },
  {
    "BeginOffset": 217,
    "EndOffset": 227,
    "Score": 0.9999289563119074,
    "Text": "JavaScript",
    "Type": "SKILLS"
  }
 ],
 "File": "test_document.txt"
}
```

The above results include the following information:

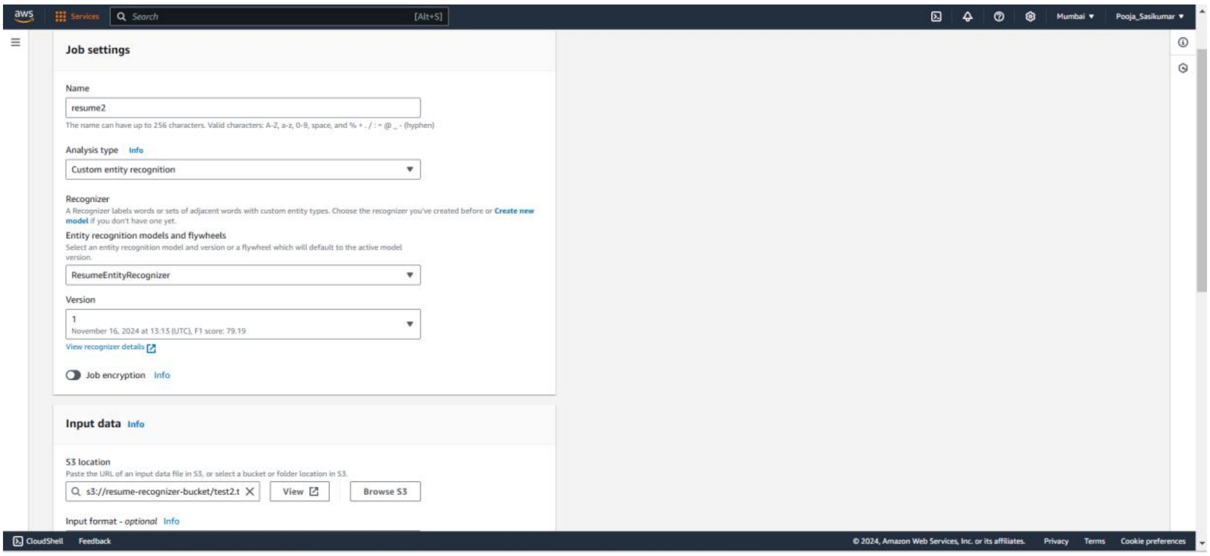
- **Offset**: The position of the entity in the text, indicated by the number of characters from the start of the line.
- **Score**: The confidence level (ranging from 0 to 1) that the model has correctly identified the entity type.
- **Type**: The entity type assigned to the recognized text which is "SKILLS" in this case .

We converted the JSON results into a table format for easier interpretation.

Entities				
BeginOffset	**EndOffset**	**Score**	**Text**	**Type**
37	67	0.7749637708158428	analytical and problem solving	SKILLS
83	86	0.9999525569523512	AWS	SKILLS
108	116	0.9114705487241651	Solution	SKILLS
161	164	0.9999367038714213	SQL	SKILLS
166	172	0.9997155283789972	Oracle	SKILLS
175	186	0.9948563676923424	Programming	SKILLS
198	200	0.9999510079565681	C,	SKILLS
201	205	0.9999717781772669	C++,	SKILLS
206	210	0.9999535105382235	Java	SKILLS
212	215	0.999892365596196	PHP	SKILLS
217	227	0.9999289563119074	JavaScript	SKILLS
File	test_document.txt			

The recognized skills from the resume include **analytical and problem-solving, AWS, Solution, SQL, Oracle, Programming, C, C++, Java, PHP, and JavaScript**, with varying confidence scores ranging from 0.775 to 0.9999.

Similarly, we perform testing with another resume by creating a new analysis job using the trained recognizer.



The new test document is uploaded to S3 and contains the following resume:

```
John Doe
New York, NY | johndoe@email.com | (123) 456-7890

Experienced IT professional skilled in C++, Java, SQL, and PL/SQL. Proficient with SAP Crystal Report and Visual Studio.
Extensive experience in performance testing, test automation, and BI analytics.
Strong background in ERP systems, Oracle, and SQL Server.
Expertise in project management and product management.
Proven ability to deliver high-quality solutions in fast-paced environments.
```

After the completion of the analysis job, the JSON results for the test2.txt document, downloaded from the output S3 location and converted into a tabular format, are as follows:

Entities				
BeginOffset	**EndOffset**	**Score**	**Text**	**Type**
99	103	0.9999929667121773	C++,	SKILLS
104	108	0.9999710329823779	Java	SKILLS
110	113	0.9999187058723578	SQL	SKILLS
119	125	0.9998638815258633	PL/SQL	SKILLS
143	146	0.9999141766761902	SAP	SKILLS
247	249	0.998692911963728	BI	SKILLS
283	286	0.9991567099970742	ERP	SKILLS
296	302	0.9999891520726072	Oracle	SKILLS
308	318	0.9992998643489172	SQL Server	SKILLS
341	351	0.9999339624169533	management	SKILLS
364	374	0.9998909353997704	management	SKILLS
File	test2.txt			

The recognized skills from the above resume are **C++, Java, SQL, PL/SQL, SAP, BI, ERP, Oracle, SQL Server, and management,** each identified with a high confidence score by the custom entity recognizer.

Conclusion

The Skill Extraction from Resumes Using Custom Entity Recognition Project demonstrated the successful application of Amazon Comprehend to extract relevant skills from resumes. By training a custom entity recognizer model and testing it with resume documents, the model efficiently identified entities like programming languages, tools, and management skills with high confidence scores as SKILLS. The output, presented in JSON format, was transformed into tabular formats for easier analysis and interpretation. This approach streamlines the process of resume analysis, enabling automated extraction of key information, which can significantly aid in recruitment workflows and talent management processes.