GenAI & NLP Project Ideas with HF models (Unit 1 Concepts)

Name: Neha Rajkumar Patil
SRN: PES2UG23CS379
Sec: F

Category 3: Analysis & Extraction (The "Smart" Agents)

Topic: Smart Resume Parser

Goal: Upload a resume text and automatically extract specific fields: **Name**, **University**, **Company**.
Tech: pipeline('ner') to find PER (Person) and ORG (Organization) entities.

Abstract:

This project implements a Smart Resume Parser that automatically extracts essential information from unstructured resume text. The system focuses on identifying the candidate's Name, University, and Company details. A pre-trained BERT-based Named Entity Recognition (NER) model is used to detect semantic entities such as Person and Organization. The solution is designed to handle variations in resume formatting and produces structured outputs from raw textual data, making it suitable for automated resume screening applications.

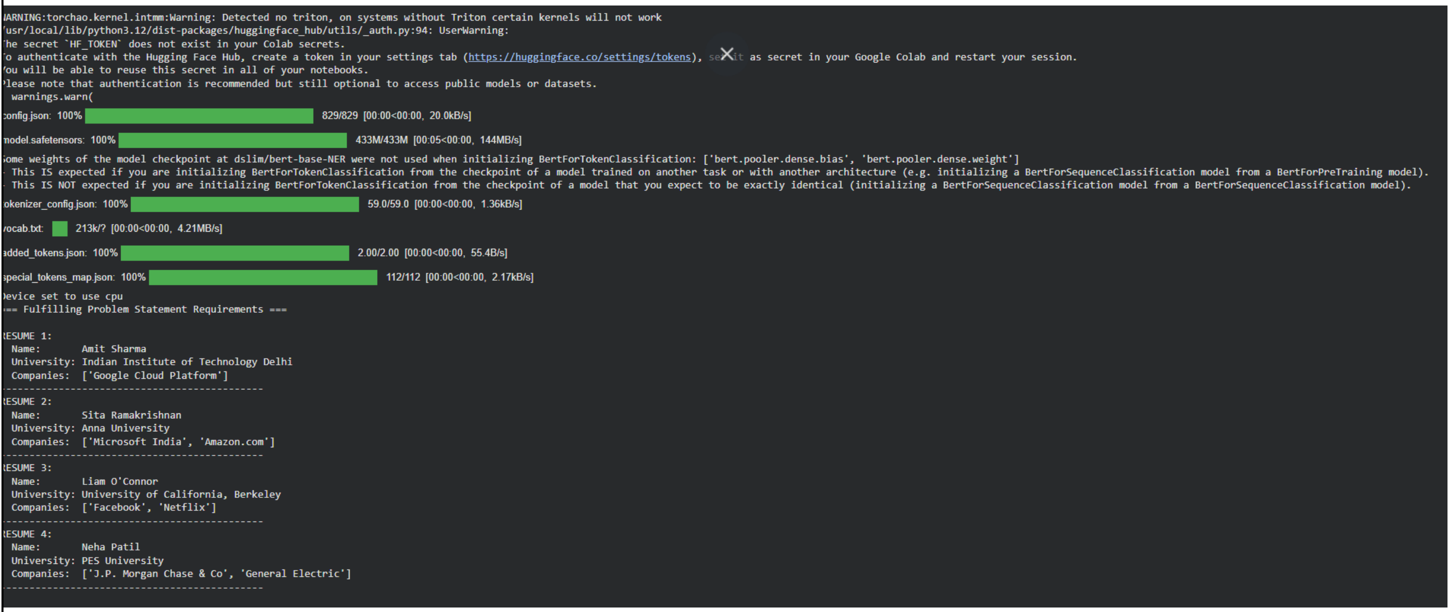
Short Documentation:

The objective of this project is to understand how Natural Language Processing (NLP) techniques can be applied to extract meaningful information from unstructured text documents such as resumes. Resumes often vary in format and structure, making manual extraction inefficient and error-prone.

To solve this problem, a Named Entity Recognition (NER) approach was adopted using the Hugging Face pipeline('ner') with the dslim/bert-base-NER model. The resume text is processed line by line, and the NER model is used to identify Person (PER) entities for name extraction and Organization (ORG) entities for company extraction. University details are extracted using pattern matching with regular expressions to improve reliability, as educational institutions are not consistently identified by NER models.

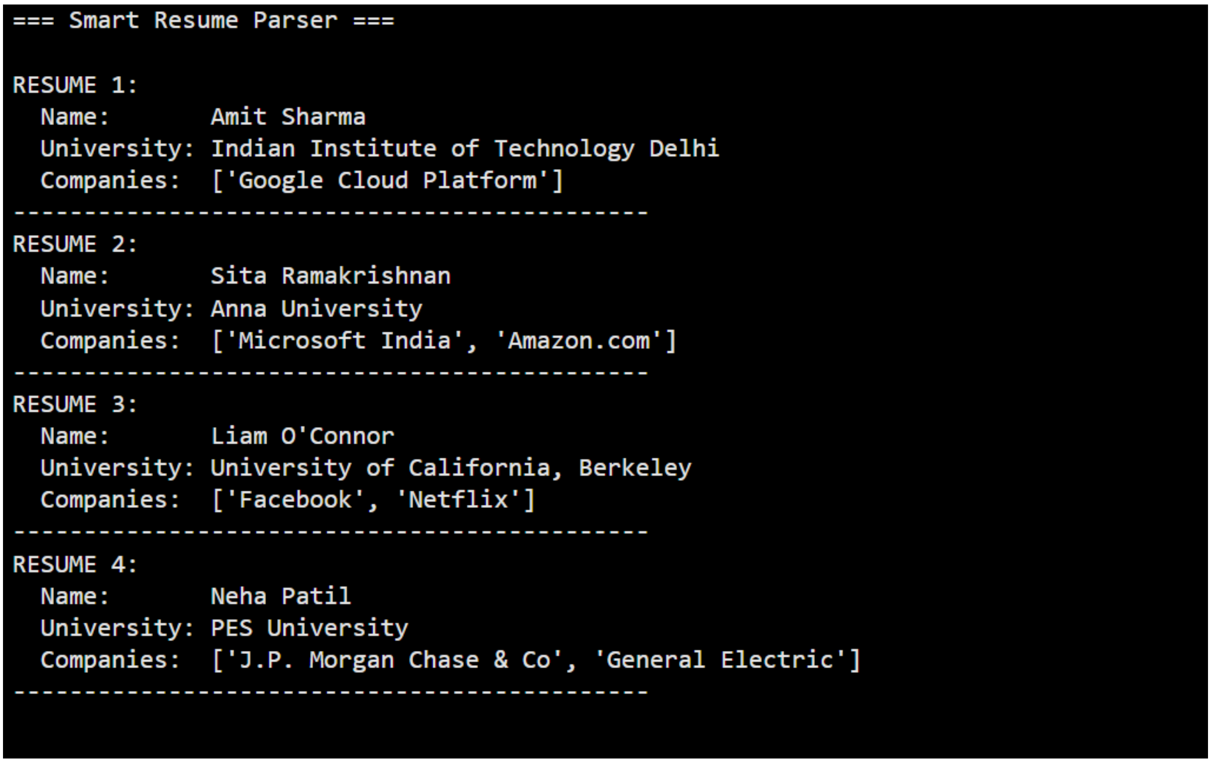
Additional post-processing is applied to handle token fragmentation and remove irrelevant role or designation information from organization names. Duplicate company entries are filtered, and the final extracted information is presented in a structured format. The implemented system successfully converts raw resume text into meaningful structured data, demonstrating the practical application of NLP in real-world document analysis.

Output Screenshot:



```
WARNING:torchao.kernel.intmm:Warning: Detected no triton, on systems without Triton certain kernels will not work
/usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set it as secret in your Google Colab and restart your session.
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
  warnings.warn(
config.json: 100%  829/829 [00:00<00:00, 20.0kB/s]
model.safetensors: 100%  433M/433M [00:05<00:00, 144MB/s]
Some weights of the model checkpoint at dslim/bert-base-NER were not used when initializing BertForTokenClassification: ['bert.pooler.dense.bias', 'bert.pooler.dense.weight']
- This IS expected if you are initializing BertForTokenClassification from the checkpoint of a model trained on another task or with another architecture (e.g. initializing a BertForSequenceClassification model from a BertForPreTraining model).
- This IS NOT expected if you are initializing BertForTokenClassification from the checkpoint of a model that you expect to be exactly identical (initializing a BertForSequenceClassification model from a BertForSequenceClassification model).
tokenizer_config.json: 100%  59.0/59.0 [00:00<00:00, 1.36kB/s]
vocab.txt:  213k/? [00:00<00:00, 4.21MB/s]
added_tokens.json: 100%  2.00/2.00 [00:00<00:00, 55.4B/s]
special_tokens_map.json: 100%  112/112 [00:00<00:00, 2.17kB/s]
Device set to use cpu
=== Fulfilling Problem Statement Requirements ===

RESUME 1:
  Name:       Amit Sharma
  University: Indian Institute of Technology Delhi
  Companies:  ['Google Cloud Platform']
---------------------------------------------
RESUME 2:
  Name:       Sita Ramakrishnan
  University: Anna University
  Companies:  ['Microsoft India', 'Amazon.com']
---------------------------------------------
RESUME 3:
  Name:       Liam O'Connor
  University: University of California, Berkeley
  Companies:  ['Facebook', 'Netflix']
---------------------------------------------
RESUME 4:
  Name:       Neha Patil
  University: PES University
  Companies:  ['J.P. Morgan Chase & Co', 'General Electric']
---------------------------------------------
```



```
=== Smart Resume Parser ===

RESUME 1:
  Name:       Amit Sharma
  University: Indian Institute of Technology Delhi
  Companies:  ['Google Cloud Platform']
---------------------------------------------
RESUME 2:
  Name:       Sita Ramakrishnan
  University: Anna University
  Companies:  ['Microsoft India', 'Amazon.com']
---------------------------------------------
RESUME 3:
  Name:       Liam O'Connor
  University: University of California, Berkeley
  Companies:  ['Facebook', 'Netflix']
---------------------------------------------
RESUME 4:
  Name:       Neha Patil
  University: PES University
  Companies:  ['J.P. Morgan Chase & Co', 'General Electric']
---------------------------------------------
```