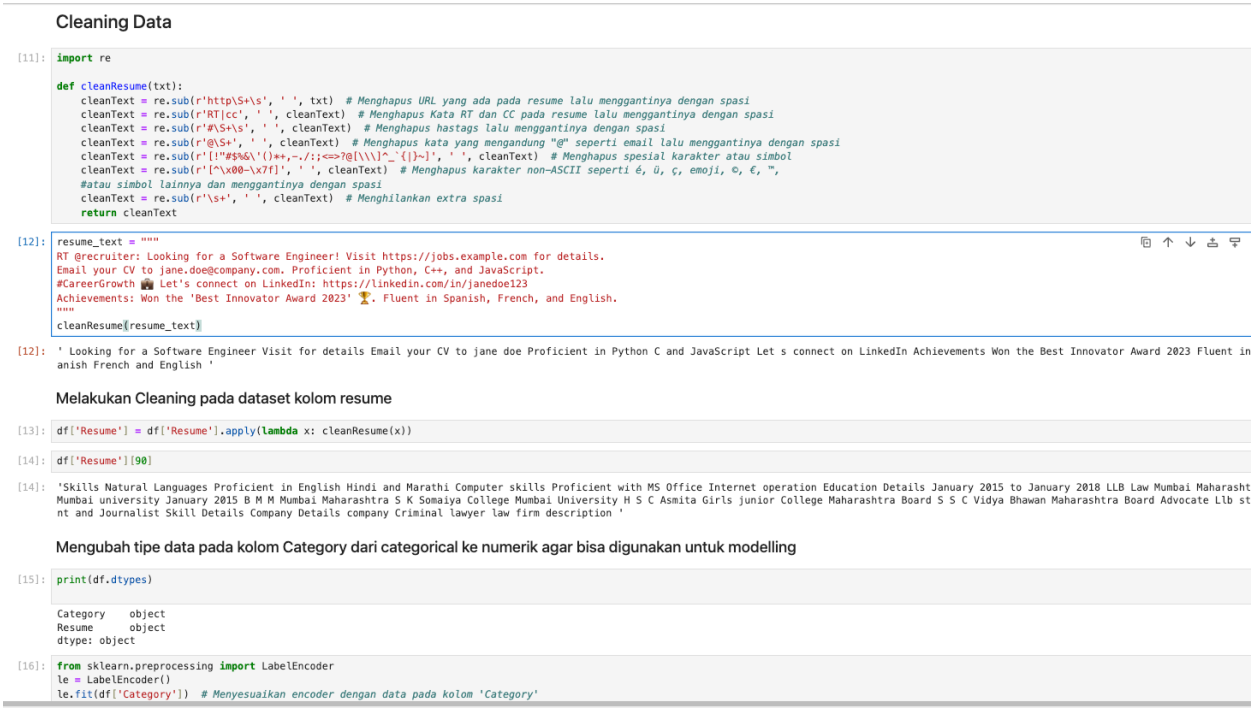
Project Mandiri NLP (Progress Report)

Nama: Muhammad Rizqi
NIM: 0110221267

Topik
Resume Screening
Referensi:

- Resume Screening using Machine Learning: https://www.kaggle.com/code/gauravduttakiit/resume-screening-using-machine-learning
- Resume-Screening-with-NLP: https://www.kaggle.com/code/akashkotal/resume-screening-with-nlp
- Resume Screening App: https://github.com/611noorsaeed/Resume-Screening-App

1. Progress pengerjaan



Cleaning Data

```
[11]: import re

def cleanResume(txt):
    cleanText = re.sub(r'http\S+\s', ' ', txt)  # Menghapus URL yang ada pada resume lalu menggantinya dengan spasi
    cleanText = re.sub(r'RT|cc', ' ', cleanText)  # Menghapus Kata RT dan CC pada resume lalu menggantinya dengan spasi
    cleanText = re.sub(r'#\S+\s', ' ', cleanText)  # Menghapus hastags lalu menggantinya dengan spasi
    cleanText = re.sub(r'@\S+', ' ', cleanText)  # Menghapus kata yang mengandung "@" seperti email lalu menggantinya dengan spasi
    cleanText = re.sub(r'[!"#$%&\'()*+,-./:;<=>?@[\\\]^_`{|}~]', ' ', cleanText)  # Menghapus spesial karakter atau simbol
    cleanText = re.sub(r'[^\x00-\x7f]', ' ', cleanText)  # Menghapus karakter non-ASCII seperti é, ü, ç, emoji, ©, €, ™,
    #atau simbol lainnya dan menggantinya dengan spasi
    cleanText = re.sub(r'\s+', ' ', cleanText)  # Menghilankan extra spasi
    return cleanText

[12]: resume_text = """
RT @recruiter: Looking for a Software Engineer! Visit https://jobs.example.com for details.
Email your CV to jane.doe@company.com. Proficient in Python, C++, and JavaScript.
#CareerGrowth 🚀 Let's connect on LinkedIn: https://linkedin.com/in/janedoe123
Achievements: Won the 'Best Innovator Award 2023' 🏆. Fluent in Spanish, French, and English.
"""
cleanResume(resume_text)

[12]: ' Looking for a Software Engineer Visit for details Email your CV to jane doe Proficient in Python C and JavaScript Let s connect on LinkedIn Achievements Won the Best Innovator Award 2023 Fluent in
anish French and English '
```

Melakukan Cleaning pada dataset kolom resume

```
[13]: df['Resume'] = df['Resume'].apply(lambda x: cleanResume(x))

[14]: df['Resume'][90]

[14]: 'Skills Natural Languages Proficient in English Hindi and Marathi Computer skills Proficient with MS Office Internet operation Education Details January 2015 to January 2018 LLB Law Mumbai Maharasht
Mumbai university January 2015 B M M Mumbai Maharashtra S K Somaiya College Mumbai University H S C Asmita Girls junior College Maharashtra Board S S C Vidya Bhawan Maharashtra Board Advocate Llb st
nt and Journalist Skill Details Company Details company Criminal lawyer law firm description '
```

Mengubah tipe data pada kolom Category dari categorical ke numerik agar bisa digunakan untuk modelling

```
[15]: print(df.dtypes)

Category    object
Resume      object
dtype: object

[16]: from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
le.fit(df['Category'])  # Menyesuaikan encoder dengan data pada kolom 'Category'
```

2. Detail Progress Pengerjaan

Saat ini proyek yang dibuat sudah dilakukan tahap-tahap berikut:

1. EDA
2. Data Cleaning
3. Proses TF-IDF
4. Membuat Model dan Training Model

Kesulitan yang dihadapi dalam mengerjakan proyek ini adalah mepelajari model yang digunakan karena terdapat metode pendekatan model yang baru penulis ketahui yaitu One vs Rest. Disini penulis menggunakan 3 model yaitu KNN, RandomForestClassifier, dan SVC.