Data Intake Report

Group Name: <NLP: Resume Extraction>
Name: <Tatiana Moteu Ngoli>
Email: <mtatiana@aimsammi.org>
Country: <Germany>
Data storage location: <https://github.com/TatianaMoteuN/Data-Glacier/tree/master/week9>

Tabular data details:

Total number of observations	<160>
Total number of files	<1>
Total number of features	</>
Base format of the file	<.json>
Size of the data	< 160 B >

Note: Replicate same table with file name if you have more than one file.

Problem description:
Resumes contain surfeit information that is not relevant for the HR/authority, and they have to manually process the resumes to shortlist the promising candidates for them. And, thus making the shortlisting task a herculean task for HR. By making use of the NER(Named Entity Recognition) model of NLP this problem can be solved by finding and classifying the entities that are present in each resume into predefined classes such as person name, college name, academics information, relevant experiences, skill set, etc.

-

Data understanding:

I have got a Json file containing raw resume text. Each raw text is divided into two columns: ‘Content’ and ‘annotation’

- The column ‘Content’: which contain a plain text listing everything in the resume of each raw
- The column ‘Annotation’: which contain annotations of each entities in the content, in the form that the NER library/model can understand . Therefore we can use Spacy for example to process with the training and evaluation.

Problem in the data:

- In the entire dataframe, we can observe the trailed ‘\n’ in each raw text
- We can also observe some noises in the data

Data cleaning and transformation:

- Write a function that will loop over each raw and clean the data by removing those trailed

- Clean the data using regex to remove unwanted characters, stop word, white spaces, url, etc… for the model to better accurate and produce reasonable outputs at the inference step.
- Apply featurization techniques such as :
    - Lowercasing
    - Punctuation and characters removal
    - Tokenization
    - Stop words removal
    - Stemming
    - Lemmatisation
- Apply NER using spacy on the raw text
- Write a function to put all the recognized entities into a new column and create another column to count the number of found entities doe each one