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/week7>

Tabular data details:

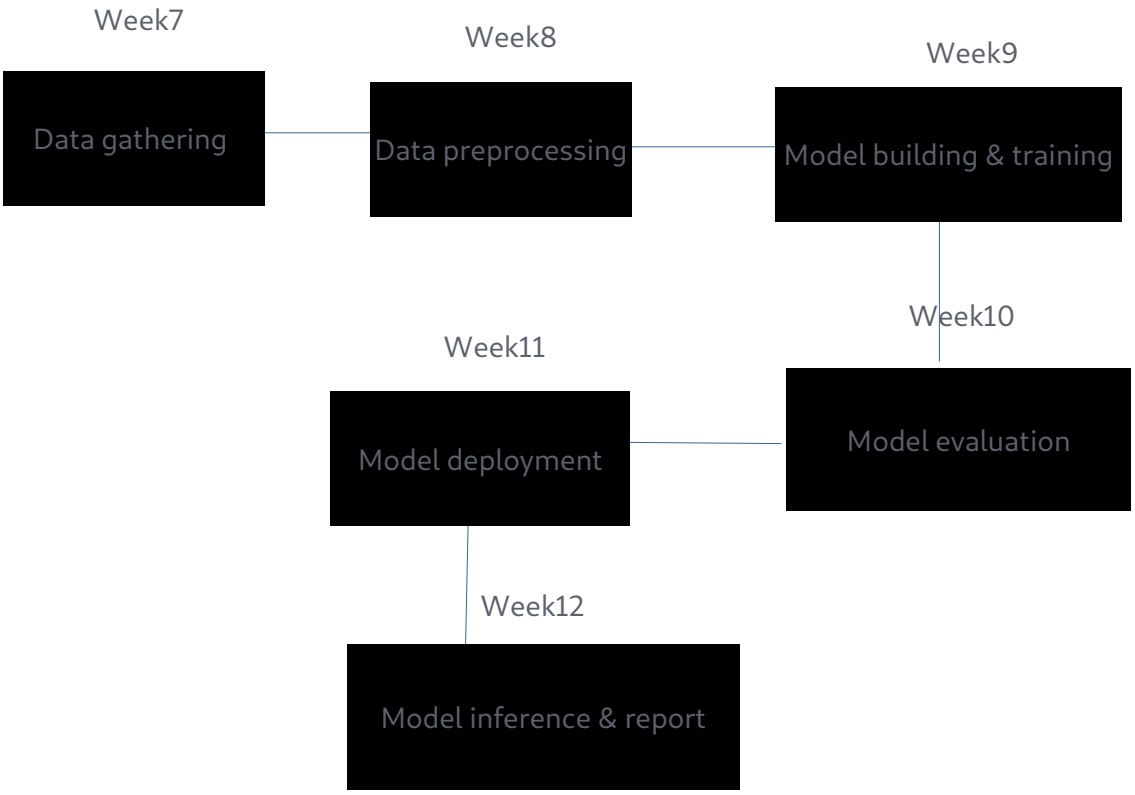
Total number of observations	<160>
Total number of files	<1>
Total number of features	</>
Base format of the file	<.json>
Size of the data	< 160 B >

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.

Business understanding:
Given different types of files as unstructured data to the HR, he need to be able to pass it through the system we will build to get an output that will help him clearly shortlist the promising candidates for the role. Our system will be able to analyze the raw text, classify and map down each entity in the text using NER into a predefined class, extract those entities as a structure data.

Project Lifecycle:
- First collect and analyze the dataset
- Preprocess and transform the dataset to the required format which will include classify each entities in the text into a predefined class
- Apply model selection and proceed with the training
- Evaluate the model and improve the model
- Deploy the model using one of the method of deployment seen before (Flask/Heruko)
- Model inference


Week7
Data gathering
Week8
Data preprocessing
Week9
Model building & training
Week10
Model evaluation
Week11
Model deployment
Week12
Model inference & report