# A perfect FIT Hackerearth ML (August 2021)

Q : This is the File that shows the approach for solving this problem:

Problem Statement- Given a Set of resumes in PDF format with their Score(Training Data) , We have to determine the score for the set of Resumes for which the PDFs are given.

Tools Used:

* Xpdf (An open source tool to extract the text from PDFs [XPDF (Download and Doc) Click here](https://www.xpdfreader.com/download.html)
* Windows Shell (Used to Call the XPDFs from Python Code)
* Python 3 (Main Programming Language)
* Keras (ANN)
* Pandas ( Used for making features in tabular format)

Steps:-

1. Copy all PDFs in Separate folder dataset in trainResume and testResume respectively.
2. Copy XPDF to these two folders
3. Make a Windows batch file that can extract the data from this files(Open processing.py to look at it)
4. Call this batch file from processing.py Using Subprocess1/2/3.bat file (You cant call a bat file in different location from current python dir.Right click and view the code inside the subprocess1/2/3 and tmp.bat to get a better understanding
5. All the text files are generated in respective folder from their pdfs
6. Open Featureextractor.ipynb.
7. Take out all the words using separators such as (,. \n)
8. Make words from this individual words having 1 2 and 3 words such as from “The indo european technology was state of art’ This can help get some composite words such as ‘indo european’ (len2) and state of art (len3)
9. Find all such words from documents and iterate over all of them again and find the frequency.
10. Choose a threshold of frequency and choose top say having fq >20
11. Add all these features to Myfeatures.txt file
12. Read this Myfeatures.txt file from Regressor ipynb
13. Make a table of this words as features(columns) and rows as the document name with the fequency of that word in it.

Ex: good thing happens to people who do good

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| good | things | happen | to | people | who | do | indo | stateofart |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |

Assign the score of these rows (training data) from train.csv

1. Design a ANN using Keras.dense and tun the parameters manually
2. Train the model
3. Test the model on the generated table from test data
4. Make a table of doc name and Score and write it to a CSV file as myres.csv

CONTACT:

RAHUL KUMAR

[rkr.deoghar@gmail.com](mailto:rkr.deoghar@gmail.com)

[Linkedin(click here)](https://www.linkedin.com/in/rahul-kumar-62778715b/)