Extraction from pdf file:

**Problem Statement -**

**Given a particular PDF document ,How to extract particular keywords**

* You will need below mentioned libraries installed on your machine for the task
* **PyPDF2**(To convert simple, text-based PDF files into text readable by Python)
* **textract**(To convert non-trivial, scanned PDF files into text readable by Python)
* **re** (To find keywords)

**Steps I followed:**

**Step 1**: Import all libraries.

**Step 2**: Convert PDF file to txt format and read data.

**Step 3**: Use “.findall()” function of regular expressions to extract keywords.

**Step4**: Apply concept of TF-IDF for calculating weights of each keyword.

**Note:**

* **TF: Term Frequency**, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:

**TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).**

* **IDF: Inverse Document Frequency**, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

**IDF(t) = log\_e(Total number of documents / Number of documents with term t in it).**