Title: Scientific Research Papers have 2-3 tables defining various quantitative and non-quantitative aspects of the paper. In this project you are going to be finding out summaries defining these tables to create a dataset of table summarization.

Design a pipeline using Python which does the following:

Section 1: Download and Preprocessing

1. Download 1000 unique scientific papers in pdf format.
2. Write a program in python which converts the papers from pdf to text and XML formats.
3. The papers should be divided into folders by types for eg. Summarization, IOT, Sentiment analysis, translation etc.
4. Rename the text and the xml files in a uniform format and keep them in well defined folders.
   E.g. document100.txt, document100.xml

Section 2: System Design

Summary can be of two types, abstractive and extractive. For each table in each scientific paper you are going to be executing these steps:

1. For each table in each paper, the abstractive summary will be the caption of the table and the extractive summary will be the text in the paper mentioning the table by its number. For eg. Let us say that in a particular paper, this table is given as Table 1.

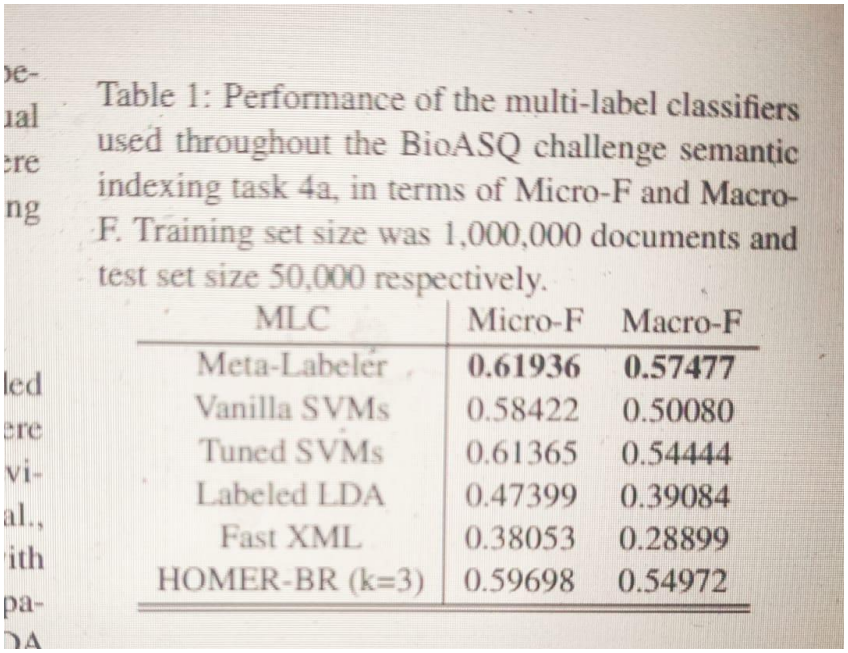
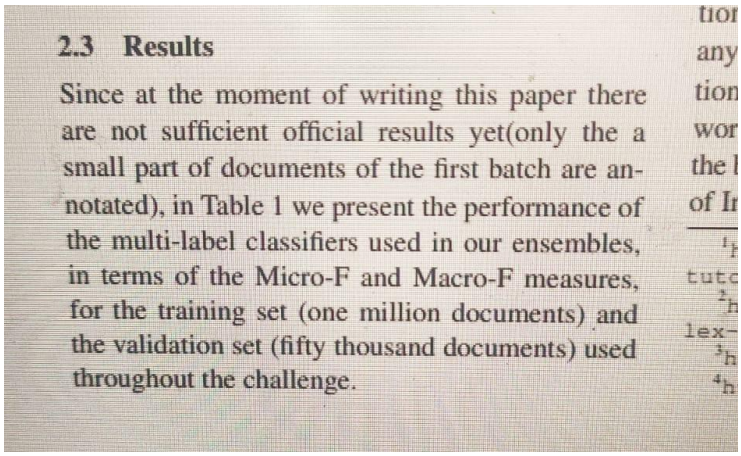


Table 1: Performance of the multi-label classifiers used throughout the BioASQ challenge semantic indexing task 4a, in terms of Micro-F and Macro-F. Training set size was 1,000,000 documents and test set size 50,000 respectively.

MLC	Micro-F	Macro-F
Meta-Labeler	**0.61936**	**0.57477**
Vanilla SVMs	0.58422	0.50080
Tuned SVMs	0.61365	0.54444
Labeled LDA	0.47399	0.39084
Fast XML	0.38053	0.28899
HOMER-BR (k=3)	0.59698	0.54972

And its mentioned in the paper text like this:



2.3 Results

Since at the moment of writing this paper there are not sufficient official results yet(only the a small part of documents of the first batch are annotated), in Table 1 we present the performance of the multi-label classifiers used in our ensembles, in terms of the Micro-F and Macro-F measures, for the training set (one million documents) and the validation set (fifty thousand documents) used throughout the challenge.

So the abstractive summary of table 1 will be the caption of the table: **Performance of the multi-label classifiers used throughout the BioASQ challenge semantic indexing task 4a, in terms of Micro-F and MacroF. Training set size was 1,000,000 documents and test set size 50,000 respectively**

The extractive summary will be the text where the table is mentioned in the paper: In Table 1 we present the performance of the multi-label classifiers used in our ensembles, in terms of the Micro-F and Macro-F measures, for the training set (one million documents) and the validation set (fifty thousand documents) used throughout the challenge

Note: If the table is mentioned multiple times then there will be multiple extractive summary for one table.

Section 3: Output File preparation

You have to prepare a single text file for 1000 papers. The format of the text file will be like this:

<Paper ID =1> <Table ID =1> <Abstractive Summary> =Table 1: Correlations between input features and average system performance for multi-document inputs of DUC 2001-2003, 2004G (generic task), 2004B (biographical task), All data (2002-2004) - UNnormalized and Normalized coverage scores</Abstractive Summary> <Extractive Summary> = NULL</Extractive Summary> </Paper ID =1>

<Paper ID =1> <Table ID =2> <Abstractive Summary> =Table 2: Performance of the multi-label classifiers used throughout the BioASQ challenge semantic indexing task 4a, in terms of Micro-F and MacroF. Training set size was 1,000,000 documents and test set size 50,000 respectively</Abstractive Summary> <Extractive Summary> = In Table 1 we present the performance of the multi-label classifiers used in our ensembles,
in terms of the Micro-F and Macro-F measures, for the training set (one million documents) and the validation set (fifty thousand documents) used throughout the challenge</Extractive Summary> </Paper ID =1>

<Paper ID =1> <Table ID =3> <Abstractive Summary> =Table 3: Multi-document input classification results on UNnormalized and Normalized data from DUC 2002 to 2004</Abstractive Summary> <Extractive Summary> =The classification accuracy for the multidocument inputs is reported in Table 3</Extractive Summary> </Paper ID =1>

<Paper ID =1> <Table ID =4> <Abstractive Summary> =Table 4: Single document input classification Precision (P), Recall (R),and F score (F) for difficult inputs on DUC'01 and '02 (total 432 examples) divided into 2 classes based on the average coverage score (217 difficult and 215 easy inputs)</Abstractive Summary> <Extractive Summary> =Multi-document task From the results in Table 4 it is evident that all three classifiers achieve accuracies higher than those for multi-document summarization</Extractive Summary> </Paper ID =1>

<Paper ID =2> <Table ID =1> <Abstractive Summary> =Table 1: Results As far as heuristic 2 is concerned, it does not cover adequately the information in the abstract</Abstractive Summary> <Extractive Summary> =Table 1 shows the sentences selected along with their scores</Extractive Summary> </Paper ID =2>

<Paper ID =3> <Table ID =1> <Abstractive Summary> =27507 Table 1: Results of ROUGE evaluation compared with other peers in DUC2004</Abstractive Summary> <Extractive Summary> =Table 1 lists the results for the comparison and Table 2 lists all the results for ASRL peers</Extractive Summary> </Paper ID =3>

<Paper ID =3> <Table ID =2> <Abstractive Summary> =33321 Table 2: Results of ROGUE evaluation for each ASRL peer of 10 results in DUC2004</Abstractive Summary> <Extractive Summary> =Table 1 lists the results for the comparison and Table 2 lists all the results for ASRL peers</Extractive Summary> <Extractive Summary> =This can be expected from Table 2, which indicates the results vary although each ASRL solution converged with some solution</Extractive Summary> </Paper ID =3>

<Paper ID =3> <Table ID =3> <Abstractive Summary> =03239 Table 3: Evaluation of combined methods</Abstractive Summary> <Extractive Summary> = NULL</Extractive Summary> </Paper ID =3>

Section 4: Submission checklist for all groups:

1. **The well defined folders containing the 1000 scientific papers in PDF, TXT and XML formats .**
2. **All the python codes used in all sections with clear cut naming conventions.**
3. **The output .txt file in the above mentioned format.**