Brandon Cuadrado, 109237297

Pranavi Venkata Changamma Meda, 111492602

Zenab Bhinderwala, #########

**Automatically Building Book Indices**

**Introduction**

Index is an alphabetical listing of words or phrases (usually key words) with references to the places/page numbers where they occur. The goal of this project is to develop an automatic index builder, which takes as input a text(document/book) and an index size and builds the index automatically. An application/user interface which allows the user to provide a text file (in latex format) as an input and outputs a new file along with the index.

**Dataset**

Resources would be long papers, documents, scanned books which have latex source and indices.

We have found papers on <https://arxiv.org/> which have latex source and indices.

\*\*\*\*\*\* to add sample tex files\*\*\*\*\*\*

**Toolset**

The latex source files for the papers/documents need to be parsed so that the file can be split into words/phrases which can be used for indexing.

There are many tools available online which will take the latex file as input and give the parsed content ( words/phrases) from the document as output.

We are currently using “**TexSoup**” tool, as a part of this project. This reads the data word by word and gets all the parsed content into a tree(python tree) structure.

\*\*\*\* should we add about the final application platform\*\*\*\*

**Preliminary Analysis**

**Subtasks**

1. To parse the input file and split the complete file into words/phrases
2. Identifying the index worthy terms from the input document
3. Where and How to put the pointers
4. Develop an application to use this automatic book index builder.

**To parse the input file and split the complete file into words/phrases**

As mentioned above we are using “TexSoup” tool to parse the latex file.

**Identifying the index worthy terms from the input document**

Approach : Once the words/terms are collected, these words should be grouped based on the usage. This could be on the base of parts of speech using a classifier. After grouping the words, a model has to be developed to rank the words which are index worthy. Based on the index size the words with the better rank are chosen and the index file is generated.

**Where and How to put the pointers**

Adding indices to a latex document can be done by the following steps :

1. Add \usepackage{makeidx} in the header section of the latex document.
2. Add \makeindex command before the document begins.
3. Through the text add \index{word to be indexed}
4. To print the indices, at the end command \printindex has to be written

**Next Set of work:**

To process the data and classify the words.