Brandon Cuadrado, 109237297

Pranavi Venkata Changamma Meda, 111492602

Zenab Bhinderwala, 109897840

Automatically Building Book Indices: Project Proposal

Introduction

An 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 a LaTeX document and the desired index size as input and outputs an index in a new LaTeX document. The application will use a model learned from existing LaTeX indices to predict the appropriate content for the generated index. The automatic index builder will be a command line application developed using Python 3.

Dataset

Resources to create the prediction model will be long papers, documents, and scanned books with indices that have LaTeX source documents. Currently, we have 17 LaTeX documents found on https://arxiv.org which have LaTeX source files and indices. The LaTeX documents vary in subject and feature indices containing both English words and mathematical symbols/concepts.

An example document can be found at the following link: https://arxiv.org/abs/1208.4948

Toolset

The LaTeX source documents need to be parsed so that the file can be split into words/phrases which can be used for indexing. There are tools available online which will take accept LaTeX files as input and provide parsed content (words/phrases) from the document as output. We are currently using the “Tex2py” python tool to parse and analyze our LaTeX documents. This tool reads the LaTeX document, including tags and rendered words, and provides the parsed content as a tree structure.

Tex2py uses a library called TexSoup in generating when reading from the LaTeX file. This tool will also be used for lower-level analysis of LaTeX files.

For determining a higher-level understanding of parsed content, a Python Natural Language Toolkit (NLTK) will be used to English keywords in a file and determine their usage as a noun, verb, or other linguistic classifier.

Preliminary Analysis

Subtasks

1. Parse the input file and split the complete file into a tree of tags and words/phrases
2. Identify the index-worthy terms from the input document
3. Determine how and why index phrases are selected
4. Apply this process to LaTeX files without an existing index

Parsing the file

As in the Toolset section above, the Tex2py tool will parse a LaTeX file into a tree in python.

Identifying index-worthy terms

Once the key words and phrases are collected, they will be classified based on their usage in the document. This will be based on their linguistic classifiers derived using the NLTK. After grouping the terms, our model will rank the words using a scoring function to determine index-worthiness.

This prediction model will be developed based upon the existing LaTeX documents with indices that have been gathered from the arxiv database. The model will consider linguistic variables, frequency in the document, and context of use. Based on the desired index size, key terms with the highest ranking will be rendered to the index.

Determine how index phrases are selected

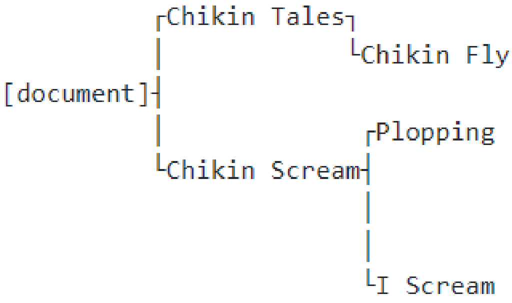
Adding indices to a LaTeX document can be done by the following steps:

1. Add \usepackage{makeidx} to the header section of the LaTeX document.
2. Add \makeindex command before the document begins.
3. Add \index{word to be indexed}
4. Add the command \printindex to the end of the document

This process requires editing a temporary copy of the inputted LaTeX document, as well as generating a new LaTeX file with the index. The model used to determine indices will also be used to determine which usage of the term will be indexed. For example, a word used 30 times will only need to be indexed for a small fraction of its use.

Proposed Model

The proposed prediction model of this automatic index builder will use Tex2py to parse a LaTeX file into a tree structure of tags and content. The parsing of LaTeX files will occur in a separate python module from analyzing the csv of vocabulary data. This python tree will be outputted in the following form:



```
          ┌Chikin Tales┐
          │            └Chikin Fly
[document]┤
          │             ┌Plopping
          └Chikin Scream┤
                        │
                        │
                        └I Scream
```

The leaves of the outputted tree will then be classified using the NLTK tool. This linguistical insight will be used as the basis for clustering and analyzing key words and phrases.

The type of scoring model, and prediction details beyond our hypotheses, remain to be determined; as the full analysis of the existing LaTeX indexed documents begins following this proposal.

Evaluating Success

To determine the success of the automatic index builder, the model will be trained using the documents of existing LaTeX files with indices. Output indices from existing documents should reflect the true index developed by the original author. For determining the success of documents without existing indices, our team will use intuition to determine whether terms are relevant to the source material. For example, if we observe that the use of the word “source” was generated as an index, the word “source” may simply be a word that was used frequently in important sentences.

The evaluation of success will be updated as we further analyze the existing LaTeX documents. This analysis will result in a better quantifiable understanding of the indexing process.