Brandon Cuadrado, 109237297

Pravani Venkata Changamma Meda, 111492602

Zenab Bhinderwala, 109897840

**Automatically Building Book Indices: Final Project**

**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 is a command line application developed using Python 2.7.

**Reading LaTeX Files**

As discussed in the Progress Report, Pylatexenc is used to parse text from a LaTeX file and for cleaning and processing. To refine the process of determining terms, the Parser has been updated to filter out words with less than 2 characters, words containing numbers, and words containing algebraic symbols. While this shortens the scope of the Parser to files with English indices, it allows for a more accurate analysis of the words and phrases making up a document.

**Parsing Entire Directories**

**Google Ngram Phrase Finder**

**Scoring Function**

**Model**

**Evaluation**

**Linear Regression**

**Model**

**Evaluation**

**Logistic Regression**

**Model**

**Evaluation**

**K-Means Clustering**

**Model**

**Evaluation**

**K-Means Clustering + Scoring Function**

**Model**

**Evaluation**

**Output Index**

**Final Index Creation**