Northeastern University, Boston
01.10.2017

Project Proposal: Plagiarism Detector

Background

Plagiarism detection is the process of locating instances of plagiarism within a work or document. The widespread use of computers and the advent of the Internet has made it easier to plagiarize the work of others. Most cases of plagiarism are found in academia, where documents are typically essays or reports. However, plagiarism can be found in virtually any field, including scientific papers, art designs, and source code. According to HEC less than 19% plagiarism is acceptable in a document.

Objectives

Existing system detect Plagiarism using vector space model (VSM) technique. Vector space model (or term vector model) is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms. It is used in information filtering, information retrieval, indexing and relevancy rankings.The aim of the project is to use **Frequent Pattern growth** algorithm.

Algorithm

FP-growth algorithm is an efficient method of mining all frequent item sets without candidate's generation. The algorithm mine the frequent item sets by using a divide-and conquer strategy as follows: FP-growth first compresses the data representing frequent item set into a frequent pattern tree, or FP-tree, which retains the item set association information as well. The next step is to divide a compressed data into set of conditional data, each associated with one frequent item. Finally, mine each such data separately. Particularly, the construction of FP-tree and the mining of FP-tree are the main steps in FP-growth algorithm.

Data

Link: https://goo.gl/apF8PA

The data set consists of students answers to about 5 questions and Wikipedia answers to those questions in separate .txt files. There are a total of 100 text files.

Approach

We plan on cleaning, merging and arranging the data in the following way:

#documentID ---- document id for a question

#question ---- Question asked

#name ---- Student name

#answer----Answer provided

Thus, the way we will approach this problem is as follows:

1) For each question, we will have various transactions which are represented as students.
2) Each transaction will consist of item sets which are words in the answer. Of course, these words will be pruned by removing common words like 'a', 'is' and so on and so forth.
3) Then we will find frequent item sets of words
4) After that we will compare those frequent item sets in our transaction which is students in this case thus giving us an answer on group of students that have plagiarized.

We will use support threshold for finding plagiarism but we might also use other factors like lift or conviction.

Northeastern University, Boston
01.10.2017

Project Proposal: Plagiarism Detector

Challenges

The main challenges we will face are which is the best threshold factor to use from support, lift and conviction. Thus, we might modify FP-growth algorithm based on that. The next challenge might be the optimum way to arrange each item sets which are words in the answer. What methods to use for pruning the words and on what basis should we prune them?

Results

Our results will be depicting which students have submitted similar work and which students have copied from Wikipedia.

Technology Stack

Programming Language: Python
Tools: Jupyter notebook
Version Control: Git

Team:

Varun Nandu: nandu.v@husky.neu.edu
Suraj Nair: nair.sur@husky.neu.edu