CALIFORNIA STATE SCIENCE FAIR
2012 PROJECT SUMMARY

Name(s)

Samyukta Yagati

Project Number

S1429

Project Title

Detecting Duplicate Content in Text Documents Using N-Gram Indexing

Abstract

Objectives/Goals

The goal was to develop a novel test for accurately identifying duplicated content in a large text corpus and build an efficient program to detect duplicate content based on the test.

Methods/Materials

The basic algorithm is to break each document in the collection into n-grams (which are consecutive runs of n words), compile an index of these n-grams, and search for clusters of documents that share a large number of n-grams. Then, a second, smaller paragraph-level index is assembled from these clusters to ascertain the proximity of shared n-grams. I experimented with different n-gram sizes, document types, and document sizes to maximize the effectiveness of the program.

The program was implemented in the Java language using DrJava IDE and JDK 6.0 on an iMac and a MacBookPro.

Results

The program successfully identified duplicate paragraphs in large documents (70 to 150 pages in length). It also pinpointed duplicate content in hundreds of news articles from the web and identified duplicate content within a single, large document through the paragraph indexing analysis. Through my experiments, I discovered that an n-gram size of three provides the best balance between storage space and accuracy.

Conclusions/Discussion

I developed a simple, effective criterion for finding duplicate content in document sets of moderate size, which I implemented into a fast, easy-to-use, accurate stand-alone program that allows the user to check for duplicate content in a group of documents.

Summary Statement

I built a duplicate content detector for large text document corpora using a simple trigram overlap test.

Help Received

My father helped me to learn to use the IDE and find relevant information about Java language constructs on the Internet.