Information Retrieval and Processing-Rainy report

Dickson Sareddy Rain

July 3, 2017

1 Introduction

1.1 Aim

For this year's project, we as team Rainy need to implement full text search function in an article database and get a presentation of extracts of the most similar parts in decreasing parts in decreasing order with the source mentioned on Harvard citation format with the title of the publication, its type and link to the publisher's full text PDF and my library's full text PDF.

2 Background

3 Methods

3.1 Latent Semantic Analysis

Measuring the similarity between words, sentences, paragraphs and documents is an important component in various tasks such as information retrieval, document clustering, word-sense disambiguation, automatic essay scoring, short answer grading, machine translation and text summarization. This survey discusses the existing works on text similarity through partitioning them into three approaches; String-based, Corpus-based and Knowledge-based similarities. Furthermore, samples of combination between these similarities are presented. Text similarity measures play an increasingly important role in text related research and applications in tasks such as information retrieval, text classification, document clustering, topic detection, topic tracking, questions generation, question answering, essay scoring, short answer scoring, machine translation, text summarization and others. Finding similarity between words is a fundamental part of text similarity which is then used as a primary stage for sentence, paragraph and document similarities. Words can be similar in two ways lexically and semantically.

Words are similar lexically if they have a similar character sequence. Words are similar semantically if they have the same thing, are opposite of each other, used in the same way, used in the same context and one is a type of another. Lexical similarity is introduced in this survey though different String-Based algorithms, Semantic similarity is introduced through Corpus-Based and Knowledge-Based algorithms. String-Based measures operate on string sequences and character composition. A string metric is a metric that measures similarity or dissimilarity (distance) between two text strings for approximate string matching or comparison. Corpus-Based similarity is a semantic similarity measure that determines the similarity between words according to information gained from large corpora. Knowledge-Based similarity is a semantic similarity measure that determines the degree of similarity between words using information derived from semantic networks. The most popular for each type will be presented briefly

3.2 Metadata extraction

Metadata Extractor is the module which can extract metadata in PDF files such as title, subheading, doi, etc. We use python package pyPdf to extract metadata directly and use metadata to do couple things. First, we extract titles of articles returning those to the web page as search results. Second, we use metadata extractor to extract subheadings which are treated as the definition of break point to split PDF file i.e. one article will be separated into small pieces based on subheading. The small pieces will convert into .txt files by the .txt converter we built so that we can take .txt files into similarity comparison and get the most relative parts in the article.Third, We extract doi [explain what is doi] which can link to the original article source.

3.3 Similarity comparison

The Sequence diagram of text similarity comparison of functionality is presented following $f$igure 1.

4 Results and discussion

We already have an interesting and hard working course. In the course, we build Django, PostgreSQL database, and develop search engine, after that we combine them together. Then, we can get the search result based on the search query. In order to have a user-friendly web page, we add some CSS programming files into it! We extend our database to 89 articles by manually downloading them from the Internet. Finally, we can display the similarity percentage and build a good-looking result page.

In brief, we have a functional web page implemented with search bar that you can copy a bunch of texts and return the percentage of similarity between the texts and the articles in the database.

5 Conclusions and future work

The results of this year are described below. We built a functional web server which contains full-text search function, similarity comparison, and the article hyperlink which can connect to the source website.
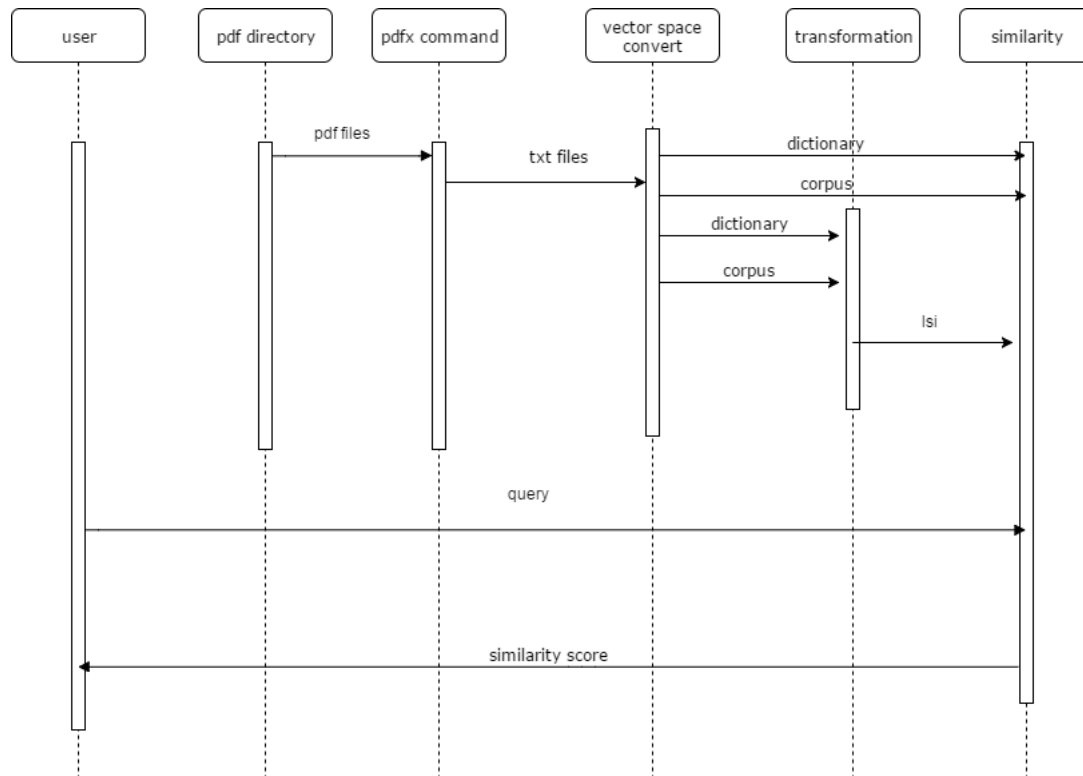
Figure 1: Sequence diagram of text similarity compare functionality.