COMP479 Project 3
Report

Project 3

Submitted by

Student ID	Name of Student
40058103	Guillaume Rochefort-Mathieu

Department of Computer Science and Engineering
Concordia University
Montreal

Fall 2019

Contents

Chapter 1: Implementation

The implementation of the project required a new tool to be made that is the web crawler. The SPIMI program that was used in the previous was mostly left unchanged except for the addition of the tf-idf ranking technique.

1.1 Web Crawlers

The framework that was used to build the web crawler to scrape the required domain was Scrapy [3]. Scrapy is Python library which takes care of crawling websites following the proper etiquette. Scrapy also offers a powerful scraping module which allows us to extract with ease the text from the HTML page that we want.

1.1.1 Concordia Crawler

The crawler was made to start from the `https://concordia.ca/research.html` link. Since these resources are HTML files I've decided to consider the content of the paragraphs and headings tags. The crawler was implemented to save the document to a file with the number of the link. The number and the link was then saved in a separate file to be able to retrieve the link for queries.

Unfortunately, after my first attempt to crawl `https://concordia.ca/research.html` the crawl was stuck in some kind of loop when crawling the `https://www.concordia.ca/research/lifestyle-addiction/tools/scientific-monitoring.html`. It ended up making roughly 10 000 requests with only changing the query string. So I decided to deny the crawler to crawl this link in hopes to make the crawl proceeding in crawling other resources.

With the updated crawler, I was able to crawl 83 690 documents where

their URL either contained news, research or next-gen. The previous requirements of the URL for it to be crawled was manually determined to limit the information that would not be related to artificial intelligence.

The content of the page that were scrape are the headings and the paragraphs inside the main-content of the site. This made sure that navigation section, footer and other parts of the website which do not change from page to page were ignore as the user information need would not be answered by the content found in those parts.

1.1.2 AITopics Crawler

The crawler was made to start from the `https://concordia.ca/search` link and extract artificial research related keywords. This website includes a section which categorized all articles under a certain technology. Since the since index needed to be related to artificial intelligence I decided to limit the crawling to the artificial intelligence articles.

Since we were only to extract keywords related to artificial intelligence research the crawling of the articles content was not constructive in the objective of the index. This is where the concept tags that AITopics assigns to articles are useful. The crawler was programed in extracting the concept tags from all articles and these would be used as the documents for the index.

1.2 Search Engine

Some changes were introduced in the search engine to adapt to the requirements of the project. First, a new ranking system, tf-idf, was introduced to run along side the BM25 that was introduced in the previous project. The search function was modify to accept the number of results to return for the query.

1.2.1 Tf-idf Ranking

The tf-idf ranking is simply ranking the documents by the results of their term frequency with respect to a document times the inverse document frequency of the term. This result in less useful results as pages where a certain terms appear for the once tends to be skewing its position in the top. BM25 seemed to be more impervious to this influence by repeating terms.

With AITopics Document Frequencies

The tf-idf ranking is the ranking scheme that was the most affected by the change of the document frequency of a term from the ConcordiaAI index to the AITopic index. Even though the project seems to hint in an improvement of the results, later in the report the opposite will be discussed. This might be due to the corpus upon which this technique was applied on.

1.2.2 BM25 Ranking

The BM25 was the ranking scheme that returned the overall most useful results.

With AITopics Document Frequencies

The BM25 ranking scheme seems to be impervious to the of the term document frequencies to the one collected from the AITopics website. The comparison between the original and the scheme modified with the document frequencies of AITopics is discussed in chapter.

Chapter 2: Queries Analysis

Using the index compiled for ConcordiaAI and AIindex queries were designed and tested to try satisfy the following information needs.

2.1 Which departments have AI research?

2.1.1 department deep learning research

In this sub-section the usefulness of the result @10, @50 and @100 will be discussed.

Original BM25

First, $P@10 = 4/10 = 0.4$. This result is suboptimal for any practical use as the user will be return more non-relevant results. After the analysis of the relevancy of the first 10 returns a relevant problem with the corpus was discover. It seems that the same is accessible by multiple URL which lead in duplication of documents.

Second, the usefulness of results @50 is similar to the previous results. A lot of duplicates are present in the result set. Hence, a lot of documents with the same content is rank with the same score.

Third, the results seems to be better @100 as the duplicates are found in the first 50 and the remaining 50 documents seem to be more relevant and diversified.

In conclusion, the word learning seems to be wildly used especially on a university website such as Concordia. It would be wise to search more toward machine learning and artificial intelligence.

BM25 with AITopics Document Frequencies

It seems that the results @10, @50 and @100 are identical to 2.1.1.

In conclusion, it seems that BM25 is impervious to the change in the document frequencies from the original corpus.

Original Tf-Idf

First, $P@10 = 5/10 = 0.5$. This result is suboptimal for any practical use as the user will be return more non-relevant results. After the analysis of the relevancy of the first 10 returns a relevant problem with the corpus was discover. It seems that the same is accessible by multiple URL which lead in duplication of documents.

Second, the usefulness of results @50 is similar to the previous results. A lot of duplicates are present in the result set. Hence, a lot of documents with the same content is rank with the same score.

Third, the results seems to be better @100 as the duplicates are found in the first 50 and the remaining 50 documents seem to be more relevant and diversified.

In conclusion, the word learning seems to be wildly used especially on a university website such as Concordia. It would be wise to search more toward machine learning and artificial intelligence.

Tf-Idf with AITopics Document Frequencies

First, $P@10 = 5/10 = 0.5$. This result is suboptimal for any practical use as the user will be return more non-relevant results. After the analysis of the relevancy of the first 10 returns a relevant problem with the corpus was discover. It seems that the same is accessible by multiple URL which lead in duplication of documents.

Second, the usefulness of results @50 is similar to the previous results. A lot of duplicates are present in the result set. Hence, a lot of documents with the same content is rank with the same score.

Third, the results seems to be better @100 as the duplicates are found in the first 50 and the remaining 50 documents seem to be more relevant and diversified.

In conclusion, for this query the use of AITopics document frequencies does not seem to outweigh the cost of crawling the keywords. It would be better to clean the corpus and use the original document frequencies.

2.1.2 department carrying out machine learning research

Original BM25

First, $P@10 = 10/10 = 1$. This query fares better on the corpus then 2.1.1. It seems that the term machine learning is used more often then deep learning. This might indicate that using less specific term in queries will result in better returns. The issue with duplicate documents is still present in the results, but it is less noticeable.

Second, results @50 shows less signs of the duplicate documents issue. The results are promising, but some documents are more about experts in the field then departments that are carrying out machine learning research. It's worth noting that usually these expert profiles have the department in which the

Third, as results tends closer to @100 the become less relevant. This might be due to BM25 has ranked all the documents that was considered more relevant first and the remaining are document that fills partially the query. Most of the last results in the @100 only are relevant to the terms research and department.

In conclusion, this query shows promise for the corpus and last query should be built upon this one to be sure to have good results. The duplication seems to have less influence, but still compromise the first 10 to 25 results.

BM25 with AITopics Document Frequencies

Similarly to 2.1.1 the results are identical to 2.1.2. This further demonstrate that BM25 is impervious to the change of the document frequencies.

Original Tf-Idf

Similarly to 2.1.1 the $P@10$ is not good. In this case, $P@10 = 3/10 = 0.3$ which is also a sub-optimal result. The word learning is producing good result for all tf-idf ranking scheme explored in this project.

After skimming through the results @50 the results are not useful. Some document are relevant but since the majority are non-relevant it makes it impractical for users to search through the results to determine which documents are relevant.

Results @100 are similar to the other results. It seems this query satisfy more an information need about pedagogical research.

In conclusion, this query does not fare well with Tf-idf ranking scheme. The corpus does contain results that is relevant to this query as seen in 2.1.2 which might indicate that tf-idf is not well suited for ranking complex queries.

Tf-Idf with AITopics Document Frequencies

Astonishingly, the use of the document frequencies from AITopics did contribute to improve the usefulness results @10. Even though the results are less precise then 2.1.1 they are still relevant to the query. $P@10 = 8/10 = 0.8$ which is a more useful result then 2.1.2.

While analyzing the results @50 a decrease of usefulness can be observed when the results goes beyond @10. It seems that the use of document frequencies from AITopics moved the more relevant document found in the @100 from 2.1.2 but consequently move back the non-relevant document back.

Results @100 are similar to the results @50, that is some documents past the 50th rank are relevant but most are not relevant to the query, consequently not satisfying the information need. As stated previously, the use of AITopics document frequencies tends to rank the most relevant documents at higher rank.

The use of AITopics document frequencies is sensible in this scenario as it tend to sort relevant documents first before non-relevant. But this technique have limits that starts to show around @50 where the results start to be similar in relevance to the one found in 2.1.2 for Tf-Idf without the AITopics document frequencies.

2.1.3 department carrying out ai research

Original BM25

Results @10 shows a precision of $6/10 = 0.6$. The results seem to indicate that using unique terms such as ai tend to improve the relevance of the documents return from this corpus. Note that the issue with duplicate is still present.

Results @50 shows a decrease in usefulness crossed the 10th rank. This could be understandable as the abbreviation of the term ai is a short form of artificial intelligence. Since Concordia website target audience in higher

education the use of abbreviation such as ai tends to be use less then other news outlet.

Results ranking between @50 and @100 tends to be completely non-relevant to the information for which the query is trying to satisfy. This further demonstrate that the use of the term ai use not as common as artificial intelligence.

In conclusion, the query tends to satisfy the information @10 but tends to drift as we increase in rank. This query helps filter out the documents that contains the term 'artificial intelligence' at the cost of numbers of relevant documents retrieved.

BM25 with AITopics Document Frequencies

Original Tf-Idf

Tf-Idf with AITopics Document Frequencies

2.2 Which researchers are working on AI research?

2.2.1 ai researchers

BM25 with AITopics Document Frequencies

Original Tf-Idf

Tf-Idf with AITopics Document Frequencies

2.2.2 concordia chair machine learning

2.2.3 thesis machine learning

2.3 What AI research is being conducted at Concordia?

2.3.1 nlp research at concordia

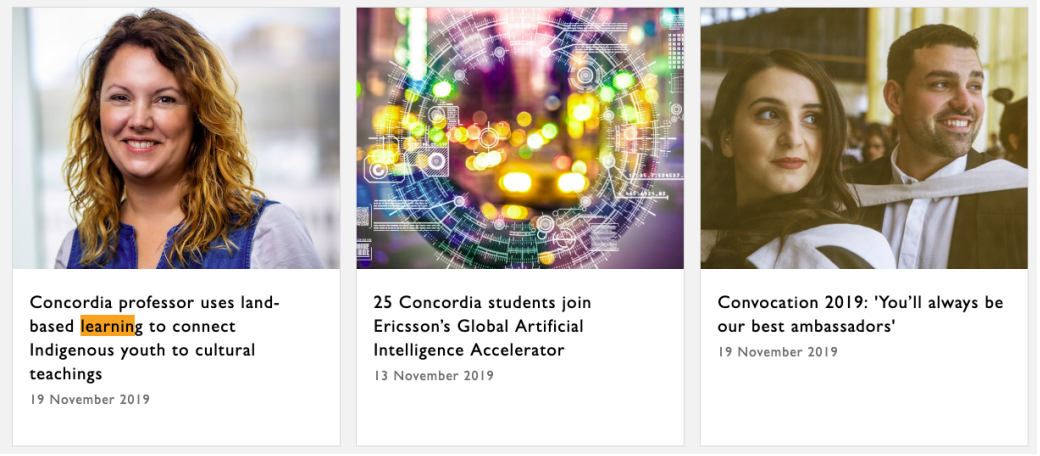
2.3.2 neural network at concordia

2.3.3 pattern recognition research at concordia

Chapter 3: Difficulties

One of the difficulty that was encountered in the project was the loop found in the crawling of the Concordia website

Another difficulty that was encountered that some of the documents accidentally contains the term artificial intelligence without being relevant to these term. This is due to how the Concordia website was scraped and the title of news article that are found on multiple page was wrongly assumed to be part of the document. See below an example of news headlines that skewed some documents for the term 'artificial intelligence'.



References

[1] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008, p. 233.

[2] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008, p. 233.

[3] "A Fast and Powerful Scraping and Web Crawling Framework," Scrapy. [Online]. Available: https://scrapy.org/. [Accessed: 27-Nov-2019].