

**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

<b>1</b>	<b>Implementation</b>	<b>1</b>
1.1	Web Crawlers . . . . .	1
1.1.1	Concordia Crawler . . . . .	1
1.1.2	AITopics Crawler . . . . .	2
1.2	Search Engine . . . . .	2
1.2.1	Tf-idf Ranking . . . . .	3
1.2.2	BM25 Ranking . . . . .	3
<b>2</b>	<b>Queries Analysis</b>	<b>4</b>
2.1	Which departments have AI research? . . . . .	4
2.1.1	department deep learning research . . . . .	4
2.1.2	department carrying out machine learning research . . . . .	6
2.1.3	department carrying out ai research . . . . .	7
2.2	Which researchers are working on AI research? . . . . .	9
2.2.1	ai researchers . . . . .	9
2.2.2	concordia chair machine learning . . . . .	11
2.2.3	thesis machine learning . . . . .	13
2.3	What AI research is being conducted at Concordia? . . . . .	15
2.3.1	nlp research at concordia . . . . .	15
2.3.2	neural network at concordia . . . . .	16
2.3.3	pattern recognition research at concordia . . . . .	17
<b>3</b>	<b>Answers</b>	<b>19</b>
3.1	Which departments have AI research? . . . . .	19
3.1.1	Mine . . . . .	19
3.1.2	Daniel Stroppolo . . . . .	19
3.1.3	Comparaison . . . . .	20
3.2	Which researchers are working on AI research? . . . . .	20

3.2.1	Mine . . . . .	20
3.2.2	Daniel Stroppolo . . . . .	21
3.2.3	Comparison . . . . .	21
3.3	What AI research is being conducted at Concordia? . . . . .	21
3.3.1	Mine . . . . .	21
3.3.2	Daniel Stroppolo . . . . .	22
3.3.3	Comparison . . . . .	22
<b>4</b>	<b>Difficulties</b>	<b>23</b>
	<b>References</b>	<b>24</b>

# 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.

Scrapy is run by executing the following command adding

```
scrapy crawl name_of_spider [-s CLOSESPIDER_PAGECOUNT = #]
```

### 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.

The df values were compiled using the SPIMI then traversed through the entire index and save the df for each term to be used later in the ConcordiaAI index. To see the compiled df value please refer to the document bellow

```
/indices/AIindex/documents_length.txt
```

## 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 wildy 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 than 2.1.1. It seems that the term machine learning is used more often than 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 than 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**

Similarly to the BM25 from 2.1.2 the results are identical to the original BM25 from 2.1.3. This further demonstrate that BM25 is impervious to the change of the document frequencies.

### **Original Tf-Idf**

Results @10 shows a precision of  $8/10 = 0.8$ . 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.

## **Tf-Idf with AITopics Document Frequencies**

The use of the document frequencies from AITopics did contribute to improve the usefulness results @10. Even though the results are less precise than BM25 from 2.1.3 they are still relevant to the query.  $P@10 = 8/10 = 0.8$  which is a more useful result than from Tf-idf from 2.1.3.

While analyzing the results @50 a decrease of usefulness can be observed when the results go 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 tends to sort relevant documents first before non-relevant. But this technique has 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.

## **2.2 Which researchers are working on AI research?**

### **2.2.1 ai researchers**

#### **Original BM25**

Results @10 shows a precision of  $10/10 = 1$ . The results seem to indicate that using unique terms such as ai tend to improve the relevance of the documents returned 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 used less than other news outlets.

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. Furthermore, simpler queries tends to have better result. This might be due to the lack of words for the results to drift on.

### **BM25 with AITopics Document Frequencies**

Similarly to the BM25 from 2.1.2 the results are identical to the original BM25 from 2.2.1. This further demonstrate that BM25 is impervious to the change of the document frequencies.

### **Original Tf-Idf**

Results @10 shows a precision of  $10/10 = 1$ . 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.

### **Tf-Idf with AITopics Document Frequencies**

The use of the document frequencies from AITopics did contribute to did not improve the usefulness results @10. Since both the results for tf-idf and

BM25 from 2.2.1 they are still relevant to the query.  $P@10 = 10/10 = 1$  which is a more useful result then from Tf-idf from 2.2.1.

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 2.2.1 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.

## 2.2.2 concordia chair machine learning

### Original BM25

Results @10 shows a precision of  $7/10 = 0.7$ . The results seem to indicate that using more specific terms such as 'research chair' 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 other term then research chair could be used. Since no query expansion is being done then this limit the results being returned to only the ones with research chair present in them.

Results ranking between @50 and @100 tends to be completely non-relevant to the information for which the query is trying to satisfy. Unfortunately the term research as a wide use in the english language and tends to drift the results from the informaton need.

In conclusion, the query tends to satisfy the information @10 but tends to drift as we increase in rank. This query helps searching for results where the term 'ai' or 'researchers' was not used in the document.

## BM25 with AITopics Document Frequencies

Similarly to the BM25 from 2.1.2 the results are identical to the original BM25 from 2.1.3. This further demonstrate that BM25 is impervious to the change of the document frequencies.

## Original Tf-Idf

Results @10 shows a precision of  $0/10 = 0$ . The results seem to indicate that using the tf-idf ranking scheme made the results drift away from the information need. It seems that tf-idf perform better on small simple queries. 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, using this ranking scheme with this query does not produce useful results.

## Tf-Idf with AITopics Document Frequencies

The use of the document frequencies from AITopics did contribute to improve the usefulness results @10. Even though the results are less precise then BM25 from 2.2.2 they are still relevant to the query.  $P@10 = 10/10 = 0.8$  which is a more useful result then from Tf-idf from 2.2.2.

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 2.2.2 but consequently move back the non-relevant document.

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 2.2.2 for Tf-Idf without the AITopics document frequencies.

## **2.2.3 thesis machine learning**

### **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**

Similarly to the BM25 from 2.2.2 the results are identical to the original BM25 from 2.1.3. This further demonstrate that BM25 is impervious to the change of the document frequencies.

### **Original Tf-Idf**

Results @10 shows a precision of  $0/10 = 0$ . The results seem to indicate that using using this query with tf-idf on this corpus makes the results drift from the information need. Note that the issue with duplicate is still present.



Results @50 shows similar result to @10. As hypothesize earlier tf-idf tends to perform better on simple and short query. This query have terms that are wildly used in an academic. Hence, this ranking scheme drifts on the term 'chair', 'concordia' and 'learning'.

Results ranking between @50 and @100 tends to be completely non-relevant to the information for which the query is trying to satisfy.

In conclusion, the original tf-idf ranking scheme does not fare well with the query. The query as terms that will make the ranking drifts from the information need. As explained, using specific term and short queries makes the tf-idf not drift.

### **Tf-Idf with AITopics Document Frequencies**

The use of the document frequencies from AITopics did contribute to improve the usefulness results @10. Even though the results are less precise then BM25 from 2.2.2 they are still relevant to the query.  $P@10 = 8/10 = 0.8$  which is a more useful result then from Tf-idf from 2.1.3.

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.

## 2.3 What AI research is being conducted at Concordia?

### 2.3.1 nlp research at concordia

#### Original BM25

Results @10 shows a precision of  $9/10 = 0.9$ . The results seem to further demonstrate that using unique terms such as nlp 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 nlp could be less used then its counter part 'natural language processing'. Since Concordia website target audience in higher education the use of abbreviation such as nlp 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. The last results tend to only contain one of the terms from the query making it less likely to satisfy the information need.

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 'language' and 'natural' two widely use term in academia. Yet, this comes at the cost of numbers of relevant documents retrieved.

#### BM25 with AITopics Document Frequencies

Similarly to the BM25 from 2.2.3 the results are identical to the original BM25 from 2.3.1. This further demonstrate that BM25 is impervious to the change of the document frequencies.

#### Original Tf-Idf

The ranking scheme tf-idf shows a  $P@10 = 0/10 = 0$ . This precision fore-shadows the remaining result return by this scheme. In brief, this ranking scheme does not return results that satisfies the information need.

### **Tf-Idf with AITopics Document Frequencies**

Similarly to the Tf-idf ranking for 2.3.1 replacing the document frequencies for the ones calculated from AITopics resulted in no improvement to the usefulness of the results.

## **2.3.2 neural network at concordia**

### **Original BM25**

Results @10 shows a precision of  $9/10 = 0.9$ . 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 term 'neural' network is widely use in neuroscience. Yet, BM25 was able to rank the relevant documents first.

Results ranking between @50 and @100 tends to be completely non-relevant to the information for which the query is trying to satisfy. As mentioned neural network is a specific topic in artificial intelligence and thus as the rank increase more and more documents are returned concerning neuroscience.

In conclusion, the query tends to satisfy the information @10 but tends to drift as we increase in rank. The BM25 ranking scheme does a efficient outcome in ranking the relevant document first.

### **BM25 with AITopics Document Frequencies**

Similarly to the BM25 from 2.3.1 the results are identical to the original BM25 from 2.3.1. This further demonstrate that BM25 is impervious to the change of the document frequencies.

### **Original Tf-Idf**

The tf-idf ranking scheme did not fare well with this query. With a  $P@10 = 0$  and the usefulness of the results up to @110 is low we had to conclude that the query is to intricate for this ranking scheme.

### **Tf-Idf with AITopics Document Frequencies**

The use of the document frequencies from AITopics did contribute to improve the usefulness results @10. Even though the results are less precise than BM25 from 2.3.1 they are still relevant to the query.  $P@10 = 3/10 = 0.3$  which is a more useful result than from Tf-idf from 2.3.2.

While analyzing the results @50 a decrease of usefulness can be observed when the results go beyond @10. It seems that the use of document frequencies from AITopics moved the more relevant document found in the @100 from 2.3.2 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 tends to sort relevant documents first before non-relevant. But this technique has limits that starts to show around @50 where the results start to be similar in relevance to the one found in 2.3.2 for Tf-Idf without the AITopics document frequencies.

### **2.3.3 pattern recognition research at concordia**

#### **Original BM25**

Results @10 shows a precision of  $0/10 = 0.0$ . The results seem to indicate that this query did not return any relevant results for the first 10 results. Note that the duplicate issue is still a problem.

Results @50 shows no increase in usefulness. Strangely even the CEN-PARMI appears less than expected in the results.

Results ranking between @50 and @100 are as completely non-relevant to the information for which the query is trying to satisfy. This further demonstrates that this query drifts.

In conclusion, this query should not be considered as a useful query.

#### **BM25 with AITopics Document Frequencies**

Similarly to the BM25 from 2.1.2 the results are identical to the original BM25 from 2.1.3. This further demonstrates that BM25 is impervious to the

change of the document frequencies.

### **Original Tf-Idf**

These results are as scattered as for the BM25 from 2.3.3. The results were scattered to a certain degree that the user would have a hard time to retrieve any useful information from the result set.

### **Tf-Idf with AITopics Document Frequencies**

Particularly, this was the only query that returned the same result for Tf-idf with AITopics Document Frequencies and the original Tf-idf. This is a peculiar situation that might be explained by none of the terms being found in the concept tags of AITopics.

# Chapter 3: Answers

## 3.1 Which departments have AI research?

### 3.1.1 Mine

- Department of Computer Science and Software Engineering (CSSE)
- Concordia Institute for Information Systems Engineering (CIISE)
- Computational Linguistics at Concordia (CLaC) Laboratory
- Centre for Pattern Recognition and Machine Intelligence
- Department of Mathematics & Statistics
- Department of Studio Arts
- Concordia Institute of Aerospace Design and Innovation
- Design and Computation Arts

### 3.1.2 Daniel Stroppolo

- Building Engineering
- Computer Engineering
- Information Systems Engineering
- Computer Science

### **3.1.3 Comparaison**

Seems that Daniel's answers were mostly related to mostly Engineering. This might be due to the use of only the term 'artificial intelligence' in his query. When the term 'machine learning' in queries helped to broaden the results to other STEM departments and even fine arts.

## **3.2 Which researchers are working on AI research?**

### **3.2.1 Mine**

- Xavier-Henri Hervé
- Sydney Swaine-Simon
- Charles Onu
- Sabine Bergler
- Anne Martel
- Jason Edward Lewis
- Dr. C. Y. Suen
- Hassan Rivaz
- Michael Hallett
- Eromonsele Samuel Oboh
- Nizar Bouguila
- Dr. T. Fevens
- Dr. O. Ormandieva
- Dr. L. Kosseim
- Dr. Marie-Jean Meurs

- Dr. W. Lucia
- Dr. J. Bentahar
- Dr. C. Poullis
- Dr. P. Rigby

Many more

### **3.2.2 Daniel Stroppolo**

- Dr. Leila Kosseim

### **3.2.3 Comparison**

Seems that Daniel's answers were mostly related to mostly one individual. This might be due to the use of only the term 'artificial intelligence' in his queries. When the term 'machine learning' and 'neural network' in queries helped to broaden the results to other STEM departments and even fine arts.

## **3.3 What AI research is being conducted at Concordia?**

### **3.3.1 Mine**

- Tool that analyzes informal online writing
- Software architecture that accelerates research and boosts quality of writing
- Neural Networks as Pseudorandom Number Generators
- Neural Network Approaches to Implicit Discourse Relation Recognition



### **3.3.2 Daniel Stroppolo**

- Pattern recognition (various theses)
- Bioinformatics and software supporting it
- Smart buildings
- Linguistics
- Cyber Security

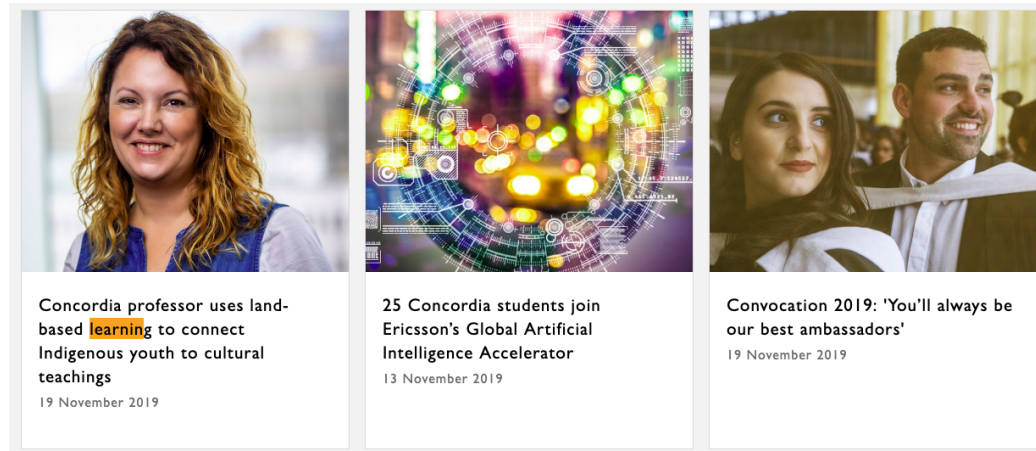
### **3.3.3 Comparison**

Daniel was able to have broader results on the research conducted at Concordia that uses artificial intelligence.

## Chapter 4: 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'.



In conclusion, lack of crawling/scraping proved to be any issue that both I and my partner encountered in the project. If something should be improved upon for any next iteration it would be to make specialized rule for each department and make sure that only the right information is collected from the website.

# 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].