2015

IsiXhosa Search Engine Development Report

DEVELOPING INFORMATION RETRIEVAL SYSTEMS FOR AFRICAN LANGAUGES

MICHAEL KYEYUNE – KYYMIC001@MYUCT.AC.ZA

Table of Contents

ABSTRACT

With Internet access becoming more widespread, opportunities for cultures to contribute to content on the World Wide Web have began to increase. This report describes the development of an information retrieval system for one of these cultures: the AmaXhosa people.

The system developed is able to crawl the Internet and index documents that are in the IsiXhosa language, then handle queries from a user with regards to the documents indexed. This IsiXhosa search engine is still in need of further development such as the inclusion of stemming in the querying and indexing of documents, which should be implemented in the future to make it more effective.

1.Introduction

As access to the Internet becomes much more wide spread in both developed and developing countries, much more diverse communities are being created online that reflect those that exist in the real world be it based on sex, colour, nation or language. This access provides an opportunity for different cultures to migrate their heritage that already existed in a physical form to digital form where they can last longer and be much more accessible to the world.

One of the components of a culture is its language and it is this aspect that the IsiXhosa search engine project seeks to address. In order to preserve IsiXhosa digitally, documents in the language have to be found, crawled, indexed and utilised through retrieval for research or any other actions and it is exactly for this purpose that the search engine is set up. The capabilities of the search such engine are :

- Crawl the web looking for documents in IsiXhosa
- Index/store any documents in IsiXhosa
- Retrieve results to a given query by a user

This report delves into the process of developing this search engine by first describing the search engine overview, then how it is designed and implemented. Evaluation of the project is then considered as well as final conclusions and recommendations as to how the project can be improved.

2.Project Overview

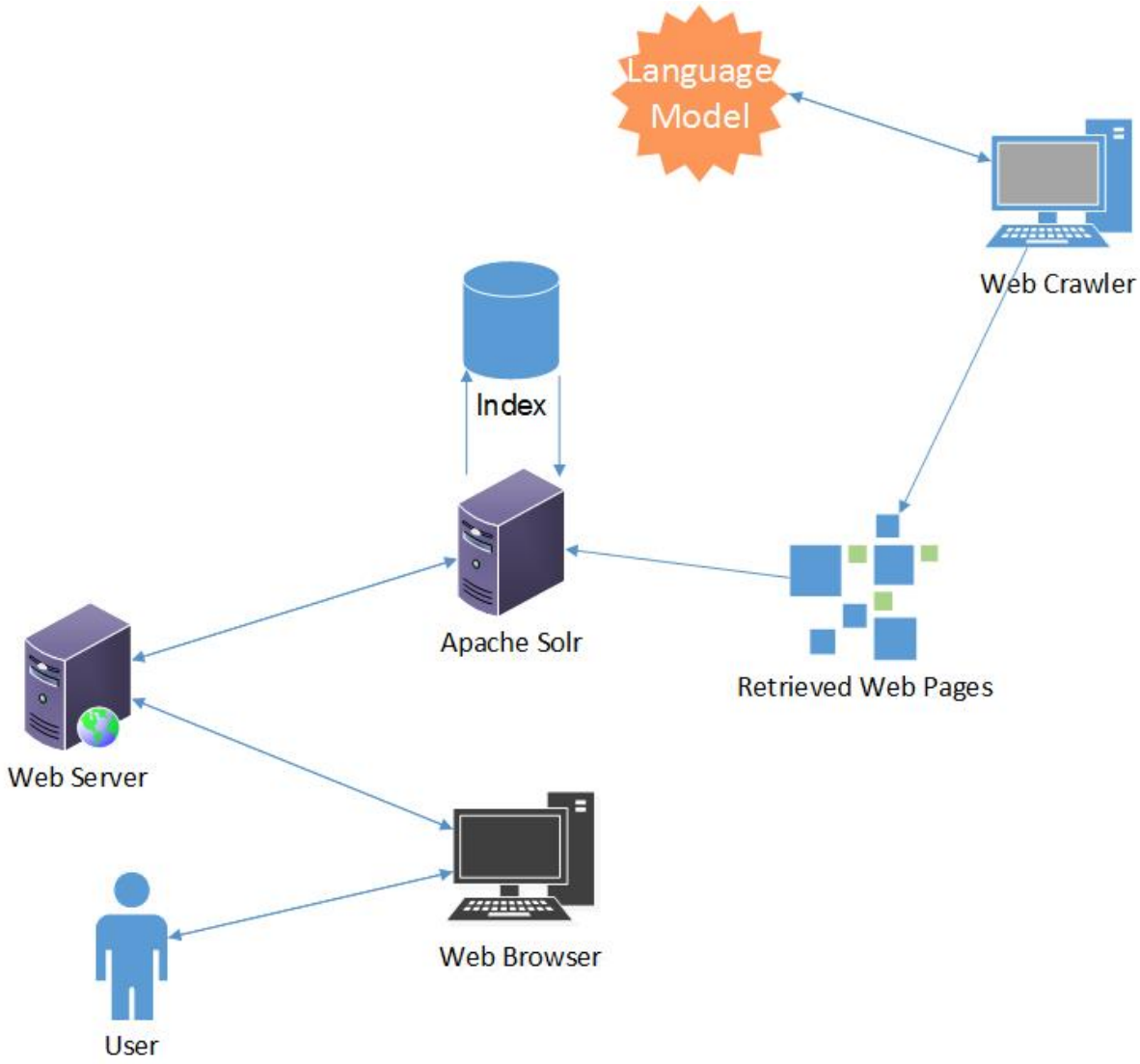

Figure 1 – An overview of the IsiXhosa Search Engine

The project involved two major parts; harvesting of IsiXhosa documents from the Web and indexing harvested documents for later retrieval.

2.1.Harvesting IsiXhosa Documents

In order to find documents in IsiXhosa, the Web is crawled with a Web crawler that, with the aid of a language model, is able to distinguish IsiXhosa from other languages, as shown in Figure 1.

2.2.Indexing And Retrieval Of IsiXhosa Documents

The Web pages obtained by the Web crawler are then indexed for later retrieval using Apache Solr, an open source search platform. Through a Web application, a user is then able to provide a query to a Web server that in turn queries the search platform for results that fit the provided query, as shown Figure 1.

3.Project Design And Implementation

In order to have the project fully functioning as expected, the following artefacts had to be designed and/or implemented:

- A statistical language model
- A Web crawler
- Index (utilising Apache Solr)
- Search engine Web application

3.1.Statistical Language Model

A statistical language model can be defined as a probabilistic distribution of a sequence of characters in that language. Each language has a language model that shows how often a set of characters or a word shows up in that language in its sentence structures and this is used to differentiate one language from another.

For this project the n-gram distribution of IsiXhosa was calculated which is the probabilistic occurrence of an n character string in the language. For example *"abc"* is a 3-gram and the probability of that string occurring in a language would be considered as the 3-gram distribution of *"abc"* in that language. To obtain an n-gram distribution for IsiXhosa a language corpus was utilized in order to traverse those documents and generate the distribution of different n-grams from the documents. The n-grams that were considered for this project were 1-gram up to and including 6-gram. All corpora utilised were obtained from (Language Resource Management Agency).

The Java Text Categorizing Library (JTCL) (Knallgrau New Media Solutions) was utilised to obtain the n-gram distribution of IsiXhosa.

3.2.Web Crawler

A Web crawler is an application that systematically browses the Web to index the documents that it encounters. An open source Web crawler known as *Crawler4J* (Repository Home Page : Crawler4J) was utilised in this project. However instead of indexing every document that was found, it only indexed those that were determined to be in IsiXhosa.

In order to determine whether or not the text in a document is in IsiXhosa, the Web crawler retrieves the text from the document using the *Jsoup* library (Hedley) and then, utilising JTCL, generates the n-gram distribution of the text which is in turn compared with that generated from the IsiXhosa corpus to see how far off the text distribution is from the standard one.

This is illustrated by the pseudo-code in Figure 3. The closer the given text is to IsiXhosa, the smaller the distance value generated.

Some languages such as IsiZulu are also similar in nature to IsiXhosa so text in that language produce a low distance value that might at times be lower than the set acceptable distance. To counter this, a Zulu corpus is also utilised to generate an n-gram distribution for that language and that distribution is compared to that of the text to obtain another distance value. The same is done for English as well as shown in Figure 2.

```
procedure CreateStandardNgramDistribution(corpus)
  standardDistribution = (all possible 1-gram, 2-gram,…, 6-grams generated from corpus)
  return standardDistribution

procedure CreateNgramDistribution(text)
  textDistribution = (all possible 1-gram, 2-gram,…, 6-grams generated from text)
  return textDistribution

procedure CalculateDistance(standard distribution , text distribution)
  distance = 0
  for every ngram in text distribution
    if ngram in standard distribution
      distance = distance + abs(position of n-gram in standard distribution – position in text
distribution)
    else
      distance = distance + number of n-grams in standard distribution
  return distance

main
  standard distribution = CreateStandardNgramDistribution(xhosa corpus)
  text distribution = CreateNgramDistribution(text)
  if CalculateDistance(standard distribution, text distribution) <= acceptable distance
    return isXhosa
  else
    return notXhosa
```

Figure 2

```
main
  standard distribution = CreateStandardNgramDistribution(xhosa corpus)
  english distribution = CreateNgramDistribution(english corpus)
  zulu distribution = CreateNgramDistribution(zulu corpus)
  text distribution = CreateNgramDistribution
  xhosaDistance = CalculateDistance(standard distribution, text distribution)
  englishDistance = CalculateDistance(english distribution, text distribution)
  zuluDistance = CalculateDistance(zulu distribution ,text distribution)
  if xhosaDistance < zuluDistance && xhosaDistance < englishDistance
    return isXhosa
  else
    return notXhosa
```

Figure 3

3.3.Index

The documents collected by the Web crawler are indexed using Apache Solr, an open source search engine platform. This platform also handles retrieval of documents given a query using standard Information Retrieval algorithms.

3.4.Search Engine Web Application

This Web application could be considered as the front end of the entire project; it provides a UI for the user to provide a query and presents the results to the user once getting feedback from Apache Solr. Figure 4 shows this UI.

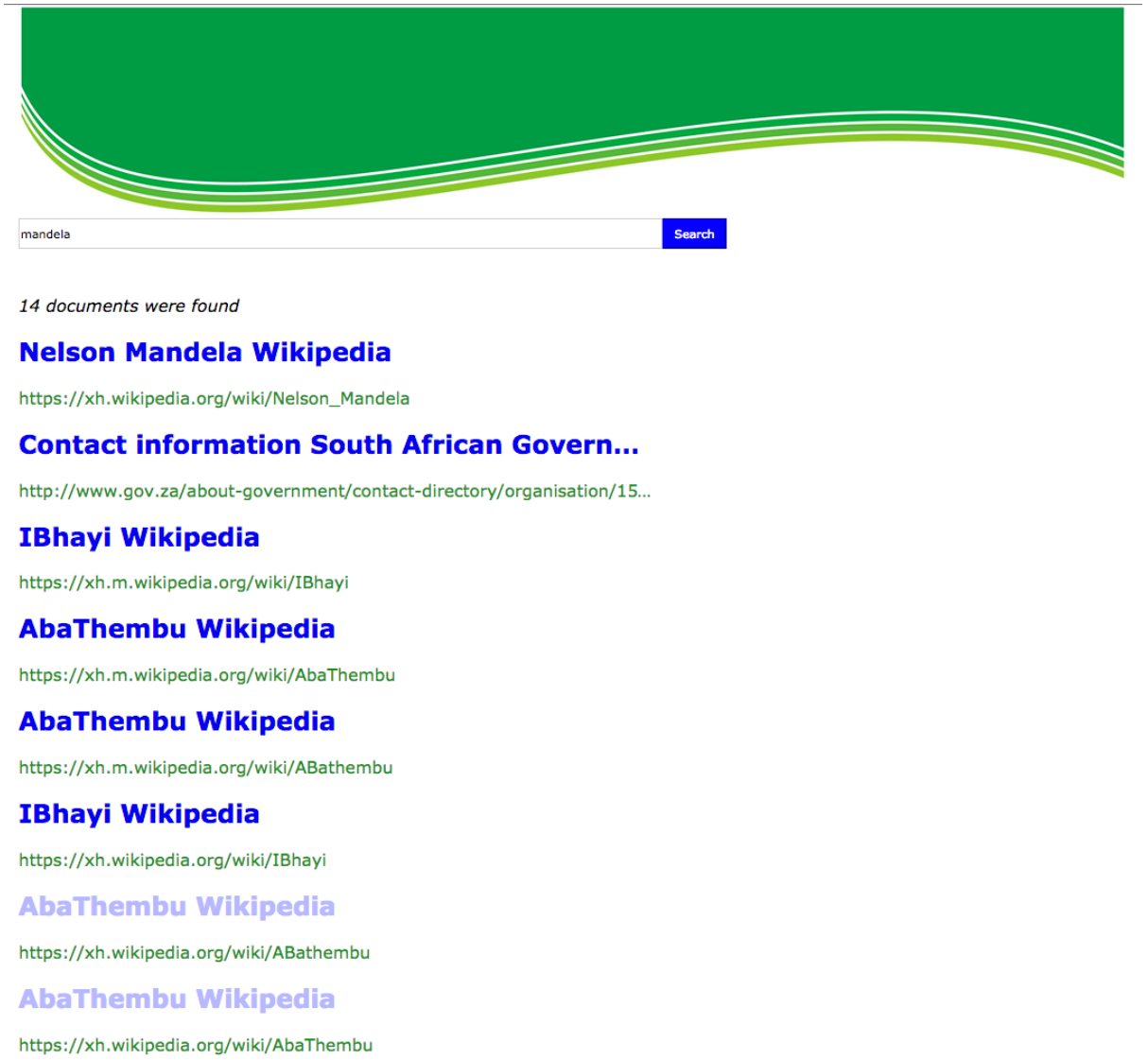

Figure 2 – Sample UI for the Web application

4.Project Future Plans

As it stands the project is functional, however there is room for improvement in the form of:

- Seed websites that are continuously updated for the Web crawler to traverse and add new documents. At the moment only the Wikipedia IsiXhosa page is a viable candidate.
- Introducing stemming of IsiXhosa in Solr to widen the range of results that can be attained from a query.
- Evaluation of the system. As it stands, the system is yet to be formally evaluated to confirm that it meets the expectations of such a system.

5. References

Hedley, J. (n.d.). *Home Page : Jsoup*. Retrieved September 6, 2015, from Jsoup Web site: http://jsoup.org/

Knallgrau New Media Solutions. (n.d.). *Home Page : Java Text Categorizing Library*. Retrieved September 6, 2015, from Java Text Categorizing Library Web Site: http://textcat.sourceforge.net/

Repository Home Page : Crawler4J. (n.d.). Retrieved September 6, 2015, from Crawler4J Github Repository Home Page: https://github.com/yasserg/crawler4j

Language Resource Management Agency. (n.d.). *Home Page : Language Resource Management Agency*. Retrieved September 6, 2015, from Language Resoucrce Management Agency Web site: http://rma.nwu.ac.za/