Zulu information retrieval

Authors: Nkosana Malumba and Katlego Moukangwe
University of Cape Town

Research Proposal

Introduction

Isizulu is one of South Africa’s eleven official languages. According to ethnologue, it is the most widely spoken home language in South Africa and it is understood by more than 50% of South Africa’s population of about 50 million people. There is little ability to search through Web-documents in IsiZulu and not much research has been done on searching Zulu Web-documents.

The goal of the Com honours project is to build a Zulu information retrieval system. The system should be able to perform simple search and retrieval. . The core of the project will be the development of a focused crawler to gather, and subsequently index, content in isiZulu from the Web. A Web interface will be used to search through indexed content. The system will be able to understand English and , Zulu search queries. The results will be returned in Zulu.

Problem statement

The main aim of the project is to investigate the feasibility of harvesting Zulu Web-documents, indexing and subsequently retrieving the harvested documents. This can be achieved by being able to automate the harvesting, indexing, searching and retrieval processes into one coherent process. One of the research tasks is to determine if it is possible to use a focused crawler to find Zulu documents on the Web. One of the major problems is recognizing isiZulu text. There isn’t a well-developed corpus for isiZulu at the moment. Working with poor corpora for language identification is one of the major topics of research in information retrieval. Using poor Zulu corpora for Zulu identification becomes a research problem. Solving the problem of Zulu Web-documents identification enables one to find the content to index when building a Zulu information retrieval system.

Zulu is morphologically different from English. In order to determine the most effective way to search through Zulu data, different searching techniques will be investigated. Stemming and removing stop-words are the main techniques to be researched . Being able to effectively and efficiently search Zulu text is one step closer to eliminating social and digital divides amongst South Africans. The research question to address the above mentioned problem is: Is it possible to bridge the digital divide through African language information retrieval? Can African culture and language be preserved through African language search engines?

A morphological parser will be used to perform stemming and to find stop-words. The system is made specifically for people who are looking for information written/stored in isiZulu. The users will primarily be people who understand Zulu and are searching for information currently stored in isiZulu. People who do not understand Zulu, however need access to the indigenous knowledge. Individuals who do not understand Zulu will be aided by a translation service.

Another research task is assessing the effectiveness of using a visitor based crawling model and how effective it is when building a Zulu corpus. This method will be more beneficial because Zulu related links are found on Zulu related Web-documents.

Methodology

In the previous section, the problems and difficulties that will be faced during the project development phase were outlined. Each problem will be dealt with according to specific methods or routines.

Recognizing Zulu text and harvesting the Web-content are the first problems. These two problems will be split up and worked on separately. To find Zulu data, there will be an initial set of pages where the crawler will start. It will consider a page relevant if it contains Zulu text. The crawler will then iteratively visit the links at a distance 2 of links from the relevant page to determine if the page is relevant. Any crawled files other than Web-documents will be stored. The strategy that will be employed to recognize Zulu text is through a language model. A language model will consist of a Zulu corpus and use n-gram similarity calculations to measure the relevance of a Web-page.

Solr will be the primary Toolkit used to handle the indexing, stop-words and stemming procedures. Solr requires the Web documents to be in Dublin Core format. The conversion from HTML to XML:dc will be achieved via a Python script. The Python script will parse HTML files, find metadata and then write the content to an XML file for indexing. A morphological parser together with Solr will be used to perform the stemming and removing stop-words. Using stemming and finding stop-words are very similar procedures with regard to IsiZulu. Stop-words will be prefixes and suffixes to words and ignoring those prefixes and suffixes will obtain the stem or root word. Lets consider an example of using stemming on the words *m-fund-isi(teacher) and aba-fund-isi(teachers)*. The system will ignore the stop-words and search for the root word *-fund-*, which will match *-funda-(learn)* in our inverted document lists.

The development cycle will consists of a preliminary design or prototype phase followed by system development and evaluation phase. The evaluation will be through direct feedback. Twenty-five Zulu speaking individuals will be acquired and requested to assess the quality and effectiveness of the system. They will also be required to fill in a questionnaire to evaluate specific sections of our system and provide suggestions for improvement. The suggestions and evaluations will be iteratively used to improve the quality of the system. The will be done by incorporating their suggestions and then re-evaluating the system.

Ethical, Professional and Legal Issues

This project is offered by the University of Cape Town's(UCT) Computer Science department and will be carried out by UCT students. UCT asserts legal and beneficial ownership of Intellectual Property(IP) arising from work by Employees and Students except as otherwise agreed in writing by an authorised officer of UCT(UCT intellectual property).

The system requires evaluation from human subjects. The University of Cape Town requires ethical clearance when performing any research that uses human or animal subjects. This is to ensure that research takes place in a scientifically and ethically justifiable environment.
For this project, the ethical clearance sought is to ensure anonymity of subjects and to ensure they are fully cognisant of the aims and outcomes of the research. The Principal Investigator and students involved are responsible for applying for ethical clearance. Students conducting the experiment or research should be listed as participants in the document to apply for ethical clearance.

Related Works

This is the one of the first attempts at developing a search engine for the Nguni languages, however, other research projects have been completed that are focused on rare languages. From these papers, potential problems have been outlined that may present some hurdles that may need to be overcome while building the IsiZulu search engine.

According to Tune(2007), Afaan Oromo is a language that is similar to the IsiZulu language that was used in an attempt to build a cross language information retrieval system. It is one of the semitic languages that convey their meaning through the language structures

The use of stop words and stemming resulted in the Afaan Oromo information retrieval system increased efficiency of the search engine in terms of higher recall(Tune et. al, 2007). In the stemming phase, words that were not found in the language corpus had to be added manually into the dictionary that was being used for translation.

Anticipated Outcomes

The theme of the African Web Languages (Zulu Information Retrieval) project is a blend of experimental design and real world software development. The most important outcome of this project will be software artefact that will be used to store and retrieve language sensitive Web pages. The language of choice in this case is the IsiZulu Language, which is a subset of the Nguni Languages of Southern Africa. The language is spoken by approximately 11.58 million people in South Africa (Statistics South Africa, 2001).

The details of the main components of the system are discussed below:

a) Focused Crawler - a crawler is an Internet Robot that systematically browses Web pages, paying attention to the content and outgoing links that are part of a Web page. The content is downloaded and indexed in a search engine to be made available for searching purposes. Any outgoing links that are found in this case are used as inputs when determining the next set of pages to crawl.

In this case the crawler will have to be specialized in order to recognize the target language when crawling Web pages. A language model will be required in order to ensure this sensitivity is employed into the crawler. The language model will be derived from the language corpus and used as the focused crawler's classifier input. This will ensure that the downloaded pages are classified in terms of their language content.

b) Information Retrieval System - An IR system will be needed to store, index and retrieve the data that is downloaded by the focused Web crawler. The chosen IR system for this project is the open source Apache Solr IR system.

Apache Solr will require customizations in order to leverage of the characteristics of the Zulu Language. Additionally, the Apache Solr system will allow the addition of stemming and stopping algorithms, which aim to improve the quality and recall of the retrieved data.

c) Web Interface - A Web interface will be developed for the purpose of allowing the user to interact with the Apache Solr IR system. User queries will be submitted and their results viewed via this interface.

The interface will make use of a language translation API such as Google Translate or Bing in order to accommodate users who can read and write the Zulu language as well as other user groups that are not fluent in the language. The latter will interact with the interface through the English Language, which will be provided by the translation APIs implemented in the Web interface.

Expected Outcomes

To this date, there hasn't been any information retrieval system that has been built specifically for South African languages. Therefore, this project is set to be an experiment that will investigate the viability of such a project. This will also expose some aspects such as the availability of Zulu texts on the internet. Additionally, this will also serve as a basis upon which other research projects on any of the 9 official languages of South Africa.

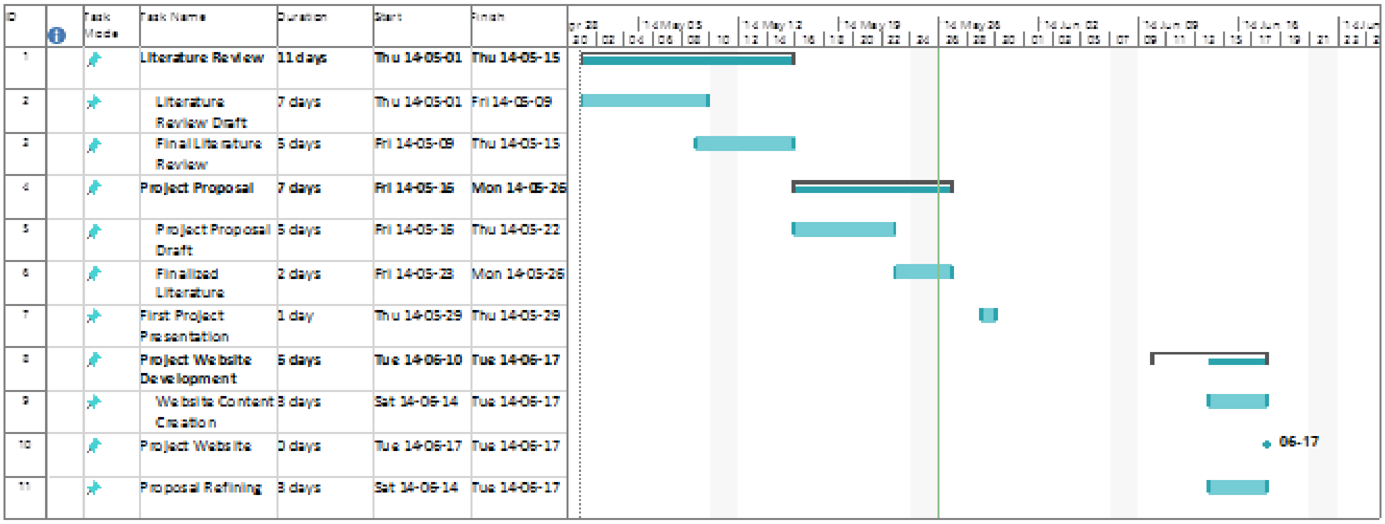
Project Plan

Risks and Management

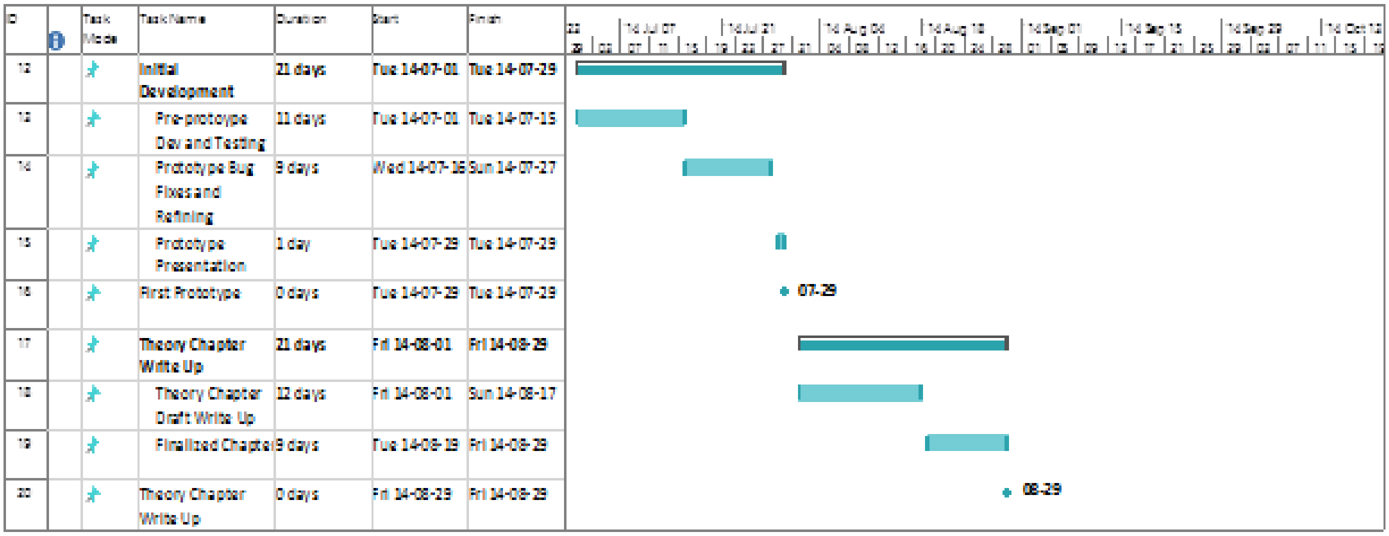
Risks	Details	Mitigation Strategy
Insufficient Zulu language sources on the Web	● Currently unclear about the amount of Web pages or documents that are written in the IsiZulu language. ● The effectiveness of the IR system depends on there being a large number of the resources that are written in the language being indexed.	● Perform the Web crawling in several iterations to determine the number of documents available. ● Explore other languages such as IsiXhosa if the documents are found to be insufficient. ● Additionally, consider including documents that are in the language but not necessarily Web pages.
Language corpus deficiencies	● The language corpus may be incomplete, thus affecting the language model. ● An ineffective language model may affect the effectiveness of the crawler and thus affect the documents that are indexed.	● Explore other languages with more complete corpora. ● Manually add in the words that have been omitted by the language corpus.
Translation API costs	● In order for the system to be usable for groups that do not speak the IsiZulu language, there has to be a method of translating the documents into the English language. ● As the documents increase in number, the costs begin to increase exponentially, which creates a monetary barrier to the effectiveness of the solution.	● Investigate other open sources that are just as effective as the paid APIs(e.g Google Translate). ● Consider translating a small segment such as the document summary or the description.
Access to individuals that can read and write IsiZulu	● In order to test the system, there needs to be Zulu	● Firstly, it is possible to advertise about the

	speakers that can test the relevance of the documents retrieved subject to the query that has been given to the search engine. ● It is highly possible to not find a large enough group of Zulu speak to provide enough test results to give an accurate approximation of the search engine's performance.	system's testing well in advance in order to raise awareness and hopefully reach a larger audience. ● Consider writing to other universities such as UWC and UKZN to provision students to test the system. ● Possibly include an incentive for users to step up and test the system.
Focused Crawler ineffectiveness	● The crawler can prove ineffective in the task of crawling for a specific language. ● This can because by either the crawler's design or the ineffectiveness of the language model that is used as the topical classifier for the crawler	● The crawler can be reconfigured in several cycles to test it's feasibility in the task at hand. ● If this proves too complicated, then it's possible that a crawler will have to be written from scratch. However, the time needs to be allocated sufficiently to ensure that this avenue can be explored.

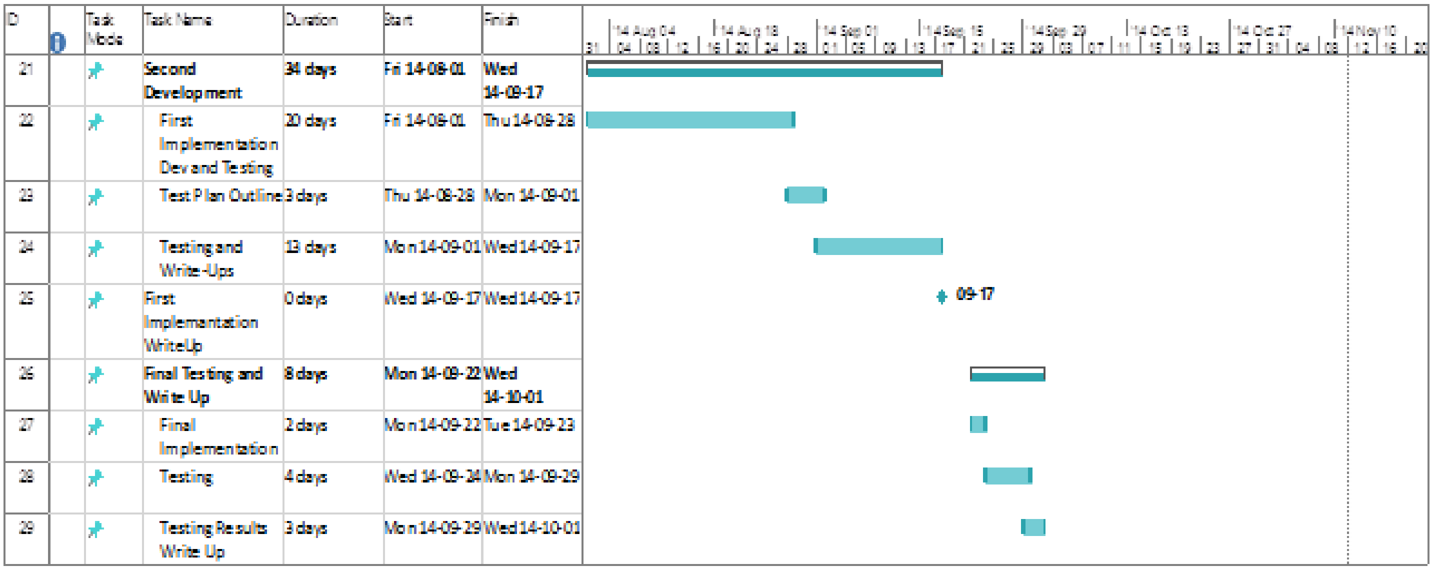
Timeline



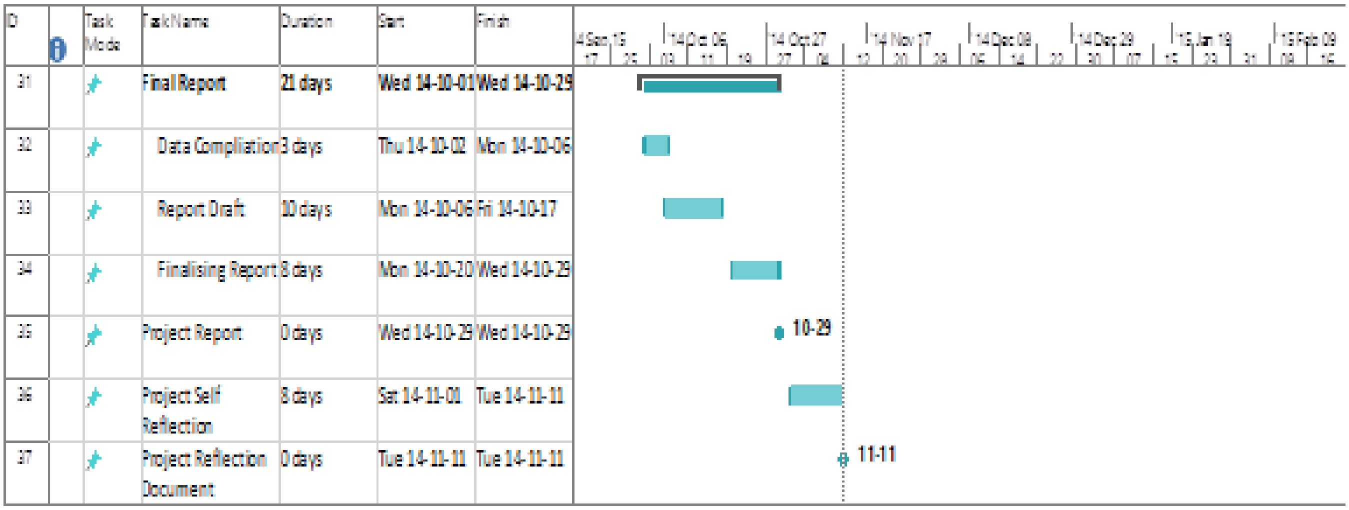
ID		Task Mode	Task Name	Duration	Start	Finish
1			**Literature Review**	**11 days**	**Thu 14-05-01**	**Thu 14-05-15**
2			Literature Review Draft	7 days	Thu 14-05-01	Fri 14-05-09
3			Final Literature Review	5 days	Fri 14-05-09	Thu 14-05-15
4			**Project Proposal**	**7 days**	**Fri 14-05-16**	**Mon 14-05-26**
5			Project Proposal Draft	5 days	Fri 14-05-16	Thu 14-05-22
6			Finalized Literature	2 days	Fri 14-05-23	Mon 14-05-26
7			First Project Presentation	1 day	Thu 14-05-29	Thu 14-05-29
8			**Project Website Development**	**5 days**	**Tue 14-06-10**	**Tue 14-06-17**
9			Website Content Creation	3 days	Sat 14-06-14	Tue 14-06-17
10			Project Website	0 days	Tue 14-06-17	Tue 14-06-17
11			Proposal Refining	3 days	Sat 14-06-14	Tue 14-06-17



ID		Task Mode	Task Name	Duration	Start	Finish
12			**Initial Development**	**21 days**	**Tue 14-07-01**	**Tue 14-07-29**
13			Pre-prototype Dev and Testing	11 days	Tue 14-07-01	Tue 14-07-15
14			Prototype Bug Fixes and Refining	9 days	Wed 14-07-16	Sun 14-07-27
15			Prototype Presentation	1 day	Tue 14-07-29	Tue 14-07-29
16			First Prototype	0 days	Tue 14-07-29	Tue 14-07-29
17			**Theory Chapter Write Up**	**21 days**	**Fri 14-08-01**	**Fri 14-08-29**
18			Theory Chapter Draft Write Up	12 days	Fri 14-08-01	Sun 14-08-17
19			Finalized Chapter	9 days	Tue 14-08-19	Fri 14-08-29
20			Theory Chapter Write Up	0 days	Fri 14-08-29	Fri 14-08-29



ID		Task Mode	Task Name	Duration	Start	Finish
21			**Second Development**	**34 days**	**Fri 14-08-01**	**Wed 14-09-17**
22			First Implementation Dev and Testing	20 days	Fri 14-08-01	Thu 14-08-28
23			Test Plan Outline	3 days	Thu 14-08-28	Mon 14-09-01
24			Testing and Write-Ups	13 days	Mon 14-09-01	Wed 14-09-17
25			First Implementation WriteUp	0 days	Wed 14-09-17	Wed 14-09-17
26			**Final Testing and Write Up**	**8 days**	**Mon 14-09-22**	**Wed 14-10-01**
27			Final Implementation	2 days	Mon 14-09-22	Tue 14-09-23
28			Testing	4 days	Wed 14-09-24	Mon 14-09-29
29			Testing Results Write Up	3 days	Mon 14-09-29	Wed 14-10-01



ID		Task Mode	Task Name	Duration	Start	Finish
31			Final Report	21 days	Wed 14-10-01	Wed 14-10-29
32			Data Compliation	3 days	Thu 14-10-02	Mon 14-10-06
33			Report Draft	10 days	Mon 14-10-06	Fri 14-10-17
34			Finalising Report	8 days	Mon 14-10-20	Wed 14-10-29
35			Project Report	0 days	Wed 14-10-29	Wed 14-10-29
36			Project Self Reflection	8 days	Sat 14-11-01	Tue 14-11-11
37			Project Reflection Document	0 days	Tue 14-11-11	Tue 14-11-11

Milestones and Deliverables

Milestone	**Deliverables**	**Date Due**
1. Literature Review	1. Proposed Literature 2. Literature Review Draft - 8 May 3. Refined literature review	15 May 2014
2. Project Proposal	1. Project Proposal Draft - 23 May	26 May 2014

	2. Finalized first draft - 26 May	
3. Project Presentations	1. Presentation Preparation - 28 May	29 - 30 May 2014
4. Finalized project proposals	1. Project Proposal Refining - 13 June	16 June 2014
5. Project Web Presence	1. Website design and development - 9 June 2. Content Creation - 12 June	17 June 2014
6. Initial Project Feasibility Demonstration	1. Initial Development Cycle Commences - 1 July 2. Pre-prototype - 15 July 3. Completed and Tested Prototype - 27 July 4. Presentation preparation - 29 July	29 July 2014
7. Background Theory Chapter	1. Theory Chapter Content - 7 August 2. Theory Chapter Draft - 21 August 3. Finalized Theory Chapter - 29 August	29 August 2014
8. First implementation test and write ups	1. Prototype refinement - 5 September 2. Preliminary Tests - 10 September	17 September 2014
9. Final implementation and testing write ups	1. Test plan - 20 September 2. Implementation and testing draft - 23 September 3. Finalized Implementation Write up - 1 October	1 October 2014
10. Final report outline	1. Report Data Collection and	22 October 2014

	Compilation - 8 October 2. Report Draft - 15 October 3. Finalized Outline - 22 October	
11. Finalized Report Handin	1. Finalized Report - 29 October	29 October 2014
11. Project Reflection	1. Reflection Draft - 5 November 2. Finalized Reflection - 11 November	11 November 2014
12. Final Project Presentations	1. Presentation preparation - 12 November	14 November 2014

Work allocation

The project will be split into two segments that will allow for parallel development and iterative cycles to be experienced on both sides of the project.

The split per student has been allocated in the following way:

Nkosana Malumba

1. Web Interface
   a. Develop a Web interface for the user to interact with the search engine
   b. Implement query translation through the use of an API
2. Search Engine
   a. Setup and configure a search engine to suit the data
   b. Structure the data so that it can be indexed and search effectively
3. Indexed Data Manipulation
   a. Build a stopword list from the indexed data
   b. Implement a stemming algorithm to increase the amount of recall.

Katlego Moukangwe

1. Web Crawling
    a. Setup and configure a Web crawler
    b. Implement a language model to ensure that the right content is crawled

2. Morphological Parser(related to stemming and stopping)
    a. Implement a morphological analysis algorithm
    b. Integrate morphological analysis parser with solr for stop-words and stemming implementations

References

1. Ethnologue. (2011). Zulu - A language of South Africa. Available from http://archive.ethnologue.com/16/show_language.asp?code=zul [Accessed on: 24 May 2014]

2. University of Cape Town.(2011). UCT intellectual property policy. Available from https://www.google.co.za/url?sa=t&rct=j&q=&esrc=s&source=Web&cd=1&cad=rja&uact=8&ved=0CCwQFjAA&url=http%3A%2F%2Fwww.uct.ac.za%2Fdownloads%2Fuct.ac.za%2Fabout%2Fpolicies%2Fintellect_property.pdf&ei=UIOCU9aoIqes7Qak-4DQAg&usg=AFQjCNH2_Q8DzQqV1GrhihwNL4OwqKRbXw&sig2=HWZp73m_dhNNdfDdY49prQ&bvm=bv.67720277,d.ZGU [Accessed on: 24 May 2014]

3. University of Cape Town. (2012). General ethics of being a scientist. Available from http://www.science.uct.ac.za/usr/science/research/responsibleresearch.pdf. [Accessed on: 24 May 2014]

4. Statistics South Africa(2001). The languages of South Africa. Available from http://www.southafrica.info/about/people/language.htm#isizulu [Accessed on: 26 May 2014]