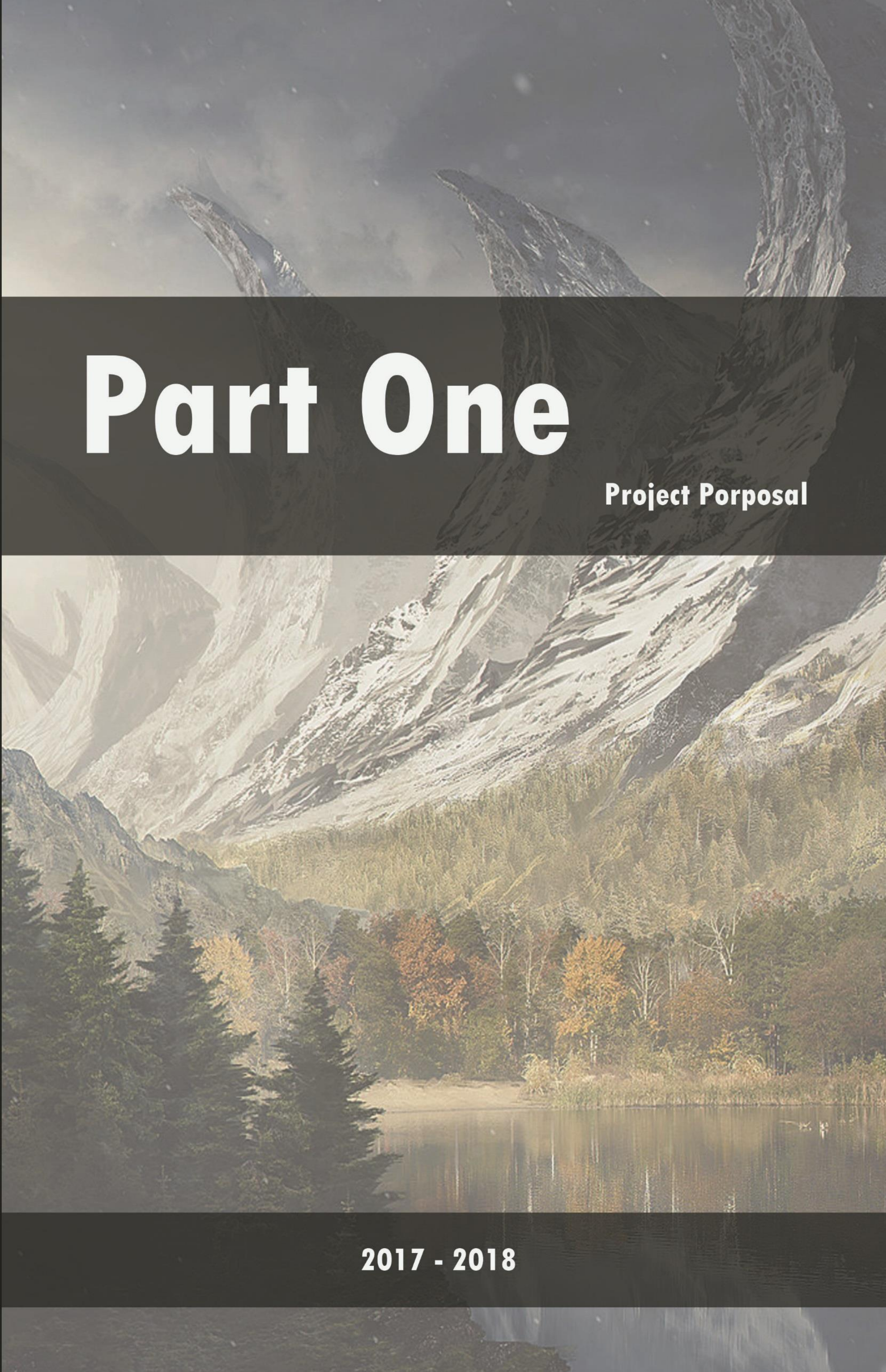
Enhance Arabic Search

Student's Name	Feryal Abdullah Al-khalaqy Khawlah Mohammed Shaiban Somia Homadi Al-khadhamy
Semester	**Semester 7 & 8**
Supervisor's Name	Dr. Asma'a Al-sharjabi



Part One

Project Porposal

2017 - 2018

CONTENTS

Glossary

1- **Search Engine:** A type of software program or script available through the Internet that seeks out and indexes documents from the World Wide Web and USENET groups based on specific criteria and searches these documents and files using keywords and returns the results of any files containing those keywords EXAMPLES: Google, Excite, Lycos, AltaVista, Infoseek, Yahoo.

2- **Web site:** A group of related pages, images, and files on a Web server.

3- **Graph link:** An electronic pathway that may be displayed in the form of highlighted text, graphics or a button that connects one web page with another web page address. Think of a hyperlink as a request to visit another place. A simple click on the link will take you there.

4- **Browser (Web browser):** software program (either text-based or graphical) that allows a person to access and browse (surf) pages on the Internet in an easy-to-use way. EXAMPLES: Firefox, Safari, Internet Explorer, Chrome, Opera.

5- **Surf:** To look at or search for web pages, usually when you are browsing from one page to another quickly by following links.

6- **Web:** Abbreviation for World Wide Web.

7- **QCRI:** Abbreviation for Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar, QCRI is an advanced research center focused on modern ICT areas.

8- **Morphology:** the study of the internal structure of words and their semantic building blocks.

9- **Corpus:** is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.

10- **Farasa:** (“insight” in Arabic), a fast and accurate Arabic SVM-based segmenter that uses a variety of features and lexicons to rank possible segmentations of a word.

PROJECT TITLE

Enhance Arabic Search

INTRODUCTION

The web creates new defies for information retrieval. The amount of information on the web is growing fast, as well as the number of new users inexperienced in the art of web research. People are likely to surf the web using its link graph, often starting with high quality human maintained or with search engines [1] that are indexing almost a thousand times as much data and between them providing reliable sub second responses to around a billion queries a day in a different of languages [2].

Automated search engines that depend on keyword matching usually return too many low quality matches, especially for Arabic language. Because Arabic language is very different and difficult structure than other languages, that's because it is a very rich language with complex morphology. [3] The morphology of Arabic poses special challenges to computational natural language processing systems. [4] Many SE's doesn't fully support Arabic morphology but still there are many weakness and problems in returning too many low quality results. Starting from this point, we have built a morphological search engine using existing components (Farasa segmenter & Solr search engine) which addresses many of the problems of existing systems. It available for users in Internet and fully supports Arabic morphology to provide much higher quality search results.

PROBLEM STATEMENT

SE's (search engines) don't consider Arabic rich morphology and in even famous Arabic sites (like Aljazeera.net) don't fully support Arabic characteristics (search by stem or lemma, root, derivatives, etc.), (e.g. typing same sentence with same key word and different stop words such as (حروف الجر - prepositions) the result almost be different), there are some search engine that try to solve this problem, such as Yamli, Eiktub, Yoolki, Bing and so on. But still have some weakness with get high quality results, because of the complex morphology that Arabic Language has.

PROJECT OBJECTIVES

Build a website that undersell the search using Arabic language, by provide the following:

1. **Enhance Accuracy of Arabic Search:** Support the search using the different morphological forms for words and synonyms.
2. **Design a search mechanism with flexible structure:** that letting addition of new linguistic resources to enrich the search process.

3. **Availability of search using complex queries.**

PROJECT SCOPE

1. Develop and configure Farasa tools and Solr SE to identify the processes of search that should be provided.
2. The search engine should integrate, with all QCRI resources.

PROJECT METHODOLOGY

As a software engineering, the methodology that will be used here is Reuse-oriented Software Engineering methodology with waterfall methodology: because the project will use existing components and should build in high quality to ensure the quality, reliability, and maintainability of the project.

The methodology that will be used in this project, as the following:

1. **Requirement specification:** Build a site to search in Arabic Wikipedia (similar to corpus.byu.edu, from Brigham Young University, which has texts in English, Spanish, and Portuguese).
2. **Component Analysis:** Call **QCRI's Farasa** tools to extract for each word: its root, stem, etc. And Use Solr search engine to index words and their morphological info.
3. **System design with reuse:** Build interface!
4. **Development and Integration:** Integrate Solr, Farasa and Search for complex queries (ex: What are the adjectives that follow a certain noun?) which are important for linguistic study and language learning.
5. **System validation:** Test the system.

Tools

- ✓ Apache Solr 6.6 or 7.1 for index the files.
- ✓ Farasa Arabic corpus.
- ✓ Visual studio 2015 or 2012 with .Net framework 4.5.
- ✓ Python Program.
- ✓ Dev C++.
- ✓ Rapid PHP 2015.

LIMITATION

1. The SE will not provide the search in any language except Arabic.
2. If the user writes the Arabic word with English characteristics, then the search engine

will not expect what is that word.

REFERENCES

[1] Brin, S. and Lawrence, **The Anatomy of a Large-Scale Hyper-Textual Web Search Engine,** Computer Science Department, Stanford University, Stanford, CA 94305, USA

[2] Hawking, D. (June 2006), **Web Search Engines,** CSIRO ICT Centre.

[3, 4] Khalid, A., Hussain Z., Anwarullah, M. 2016. **Arabic Stemmer for Search Engines Information Retrieval,** International Journal of Advanced Computer Science and Applications, Vol. 7, KSA, No. 1.