# **Qatar Content Classification**

Client

Tarek Kanan

tarekk@vt.edu

Presenter

Mohamed Handosa

handosa@vt.edu

VT, CS6604

March 6, 2014

### About The Project

- Funded by QNRF (<a href="http://elisq.qu.edu.qa">http://elisq.qu.edu.qa</a>)
- Started at VT in 1/1/2013, and running through 12/31/2015.
- A project to advance digital libraries in the country of Qatar.
- Collaborating institutes: Penn State, Texas A&M, and Qatar University.

#### Project Plan

- Build Arabic collections using Heritrix crawler
- Build a universal taxonomy for Arabic newspapers
- Use different classifiers to classify Arabic documents
- Use Apache Solr to index and search Arabic collections
- Evaluate the performance of the classifiers on Arabic data

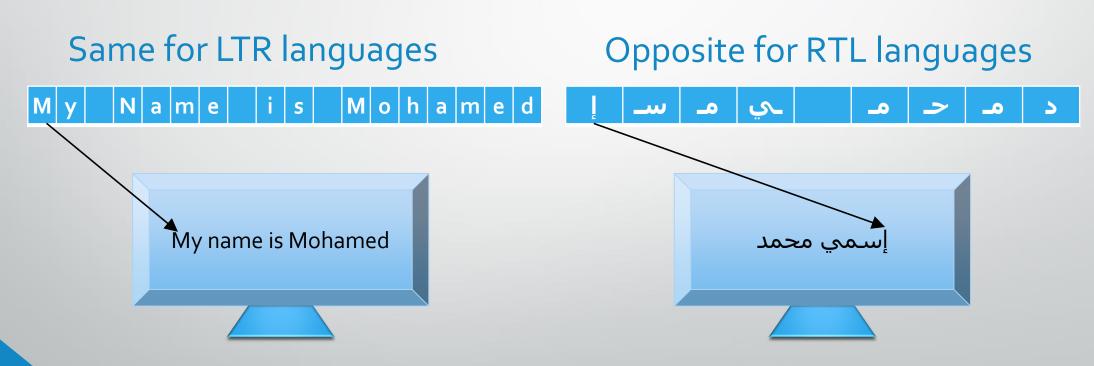
### Accomplished

- Helped building the Arabic newspaper taxonomy.
- Helped developing a tool to convert Arabic PDF files to TXT files.
- Helped installing and running Solr with Tomcat as a web container.
- Helped uploading, indexing and testing (querying) the Arabic collection.

#### PDF to Text Conversion

- Converting PDF to TXT makes files easier to transfer and process.
- Converting Arabic PDF can be challenging because it is a RTL language.
- Generally, text is stored in logical order, but displayed in presentation order.

### Logical and Presentation Orders



#### Conversion Tool (PDF2TXT-A)

- PDF stores data in presentation order.
- Need to convert from presentation to logical order.
- After decoding each line, reverse the order of the Arabic text.



#### Preparing the Dataset

#### Procedure (for each PDF file)

Extract and clean Arabic Text

Create an XML file

id = file name c

content = text

#### XML file format

```
<add>
<doc>
<id>file-name</id>
<class>initially-empty</class>
<content>Arabic-text</content>
</doc>
</add>
```

#### Classification

- Split the dataset into a training and a testing set.
- Classify the training set (fill the class tag) manually.
- For each of the classifiers to be tested
  - Train the classifier using the training set.
  - Run the classifier on its own copy of the testing set (fill the class tag).

## Uploading to Solr

- Create a core R on Solr.
- Upload the manually classified training set to R.
- For each classifier  $X_i$ ,
  - Create a core  $R_i$  on Solr.
  - Upload the testing set copy classified by  $X_i$  to  $R_i$ .

## Planning to Accomplish

- Building more collections of Arabic documents.
- Preparing manually classified training set and upload it to Solr.
- Training and running different classifiers on the unclassified testing set.
- For each classifier, uploading classified documents to a different Solr core.
- Running different queries on Solr for classifiers cores and training set core.
- Compare the query results of each classifier core with the training set core.

# **Thank You**