

Project Title: Enhance Arabic Search

Project Description:

Search Engines (SE's) should consider **Arabic rich morphology**. SE's in even famous Arabic sites (like Aljazeera.net) don't fully support Arabic characteristics! (Search by stem or lemma, root, derivatives, etc.)

Solution: As an example for a **morphological SE**, we will:

- ✓ 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).
- ✓ Call **QCRI's Farasa** tools to extract for each word: its root, stem, etc.
- ✓ Use **Solr** search engine to index words and their morphological info.
- ✓ Build **interface!**
- ✓ Search for **complex queries** (ex: What are the adjectives that follow a certain noun?) which are important for linguistic study and language learning.

The image shows a screenshot of the corpus.byu.edu website. The main header reads "corpus.byu.edu" and "corpora, size, queries = better resources, more insight". Below this is a table listing various corpora with columns for language, number of words, language/dialect, time period, and a compare button. The table includes corpora like NOW Corpus (2.8 billion words), Global Web-Based English (1.9 billion words), Wikipedia Corpus (1.9 billion words), Hansard Corpus (1.6 billion words), and others. Below the table, there are two search result snippets. The first snippet is for the query "the-day the-sportive of-qatar" (Sports Day of Qatar) and shows three search results with Arabic text. The second snippet is for the query "the-day the-sportive qatar" (Sports Day Qatar) and shows three search results with Arabic text. A yellow starburst graphic with the text "Different Results!" is overlaid on the second snippet. At the bottom left, there is a small table with the heading "CLICK FOR MORE CONTEXT" and three rows of text related to the word "Anarchism".

English	# words	language/dialect	time period	compare
NOW Corpus	2.8 billion*	20 countries / Web	2010-yesterday	
Global Web-Based English (GloWbE)	1.9 billion	20 countries / Web	2012-13	
Wikipedia Corpus	1.9 billion	English	2016	Info
Hansard Corpus (British Parliament)	1.6 billion	British	1803-2005	Info
Corpus of Contemporary American English (COCA)	520 million	American	1990-2015	* * * * *
Corpus of Historical American English (COHA)	400 million	American	1810-2009	* *
Corpus of US Supreme Court Opinions	130 million	American	1790s-present	
TIME Magazine Corpus	100 million	American	1923-2006	
Corpus of American Soap Operas	100 million	American	2001-2012	*
British National Corpus (BYU-BNC)*	100 million	British	1980s-1993	* *
Srathy Corpus (Canada)	50 million	Canadian	1970s-2000s	
CORE Corpus	50 million	Web registers	-2014	
Other languages				
Corpus del Español (see also...)	2.1 billion	Spanish	1200s-1900s	*
Corpus do Português (see also...)	1.1 billion	Portuguese	1300s-1900s	
N-grams				
Google Books: American English	155 billion	American	1500s-2000s	*
Google Books: British English	34 billion	British	1500s-2000s	
Google Books: Spanish	45 billion	Spanish	1500s-2000s	

CLICK FOR MORE CONTEXT			
1	Anarchism	A B C	only, referring to free-market anarchism as" libertarian anarchism". # History # # Origins # The
2	Anarchism	A B C	and... the first anarchist society was that of the apostles. In early Islamic history, some manifest
3	Anarchism	A B C	year in" Le Rvloit" that a structure based on centuries of history can not be destroyed with a few

Required Skills:

- Good programming skills (Python/Java/C#/C++).
- Understanding Arabic is a plus but not a must!

Expected Team Size:

One to two students

Mentors: Hamdy Mubarak, Kareem Darwish, and Ahmed Abdelali

Name:

Hamdy Mubarak

Kareem Darwish

Ahmed Abdelali

e-mail:

hmubarak@hbku.edu.qa

kdarwish@hbku.edu.qa

aabdelali@hbku.edu.qa