NLP for Arabic Curriculums

Mentors: Hamdy Mubarak, Ahmed Abdelali, Kareem Darwish {hmubarak, aabdelali, kdarwish}@hbku.edu.qa

Motivation: Few studies on Arabic curriculums. These studies help in social studies and understanding differences between cultures, improve language skills for language learners, proficiency tests, text simplification, etc.

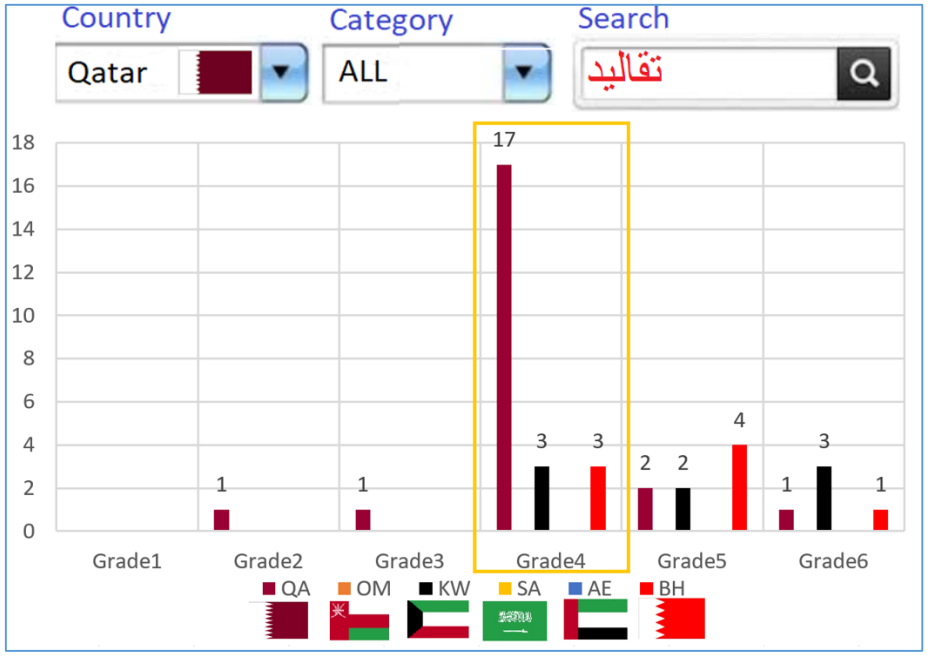
Available Resource: Arabic curriculums for **Gulf** countries (Grades 1 to 6):

Automatic processing (Farasa NLP tools): diacritized and lemmatized (stemmed), statistics, and human subjects categorization

❖ Solution

Build a website to do the following tasks:

* Browse all terms and term usage at the country and grade levels
(term "traditions": Qatar is the top country, and its grade is Grade4+)



* Browse usage of broad concepts (synonyms, related words), ex: good morals, family and society, etc.

كرم، شهامة، شجاعة..، الأسرة والعائلة

* Browse categories at the country and grade levels
(in "Animal" category, term "horses" is #1 in Qatar)

* Learn 20 words only covers 80% of all animals used in all countries (Grades 1 to 6):

(animal, camel, fox, lion, rabbit, dog, cat, wolf, sheep, … monkey, snake)

* Show possible word forms of a given lemma (stem) with pronunciation (TTS) and examples of usage:
Ex: term "reading" in Qatar (Grade 6)

قِراءَة	قراءة	**130**	الْقِرَاءَةِ	القراءة	43	الْقِرَاءَةُ	القراءة	23	قِرَاءَةً	قراءة	16	قِراءةً	قراءة	12	القراءةِ	القراءة	11
			القراءةُ	القراءة	6	قِرَاءَةَ	قراءة	6	لِقِرَاءَةِ	لقراءة	3	قراءةَ	قراءة	2	القراءةِ	القراءة	1
			الْقراءةِ	القراءة	1	الْقِرَاءَةِ	القراءة	1	الْقِرَاءَةَ	القراءة	1	بِقِرَاءَةٍ	بقراءة	1	قراءةِ	قراءة	1
			قِرَاءَتِكَ	قراءتك	1	لِلْقِرَاءَةِ	للقراءة	1									

جَلَسَا لِقِرَاءَةِ الْقُرْآنِ بِتَدَبُّرٍ

* Show statistics about each grade in each country: #words, #lemmas, development with time…
* Show top terms that are common across all countries (for a certain grade), and terms appear in one country but not in the others.
* Augment text data with images (top results from Google image search)

* Determine grade level for any input text, highlight complex words, and suggest simpler ones.

اللغة السريانية.. تمتد لآلاف السنين

وتأبى الاندثار وترفض الانقراض

السريانية هي عمق التاريخ الذي لا يمكن لشخص أو فئة -صغرت أو كبرت- التجرد منها، فهي نسب من نوع آخر، فموسيقى أبجدية حروفها عزف حضارات تمتد لآلاف السنين.

شارك

Grade 6+

❖ Skills

✓ Experience in Java/Python for data processing, calling APIs…
✓ Experience in web development: HTML, CSS
✓ Passionate about language and culture studies, and preparing resources for language learners (native and non-native speakers)
✓ Understanding Arabic is a plus!