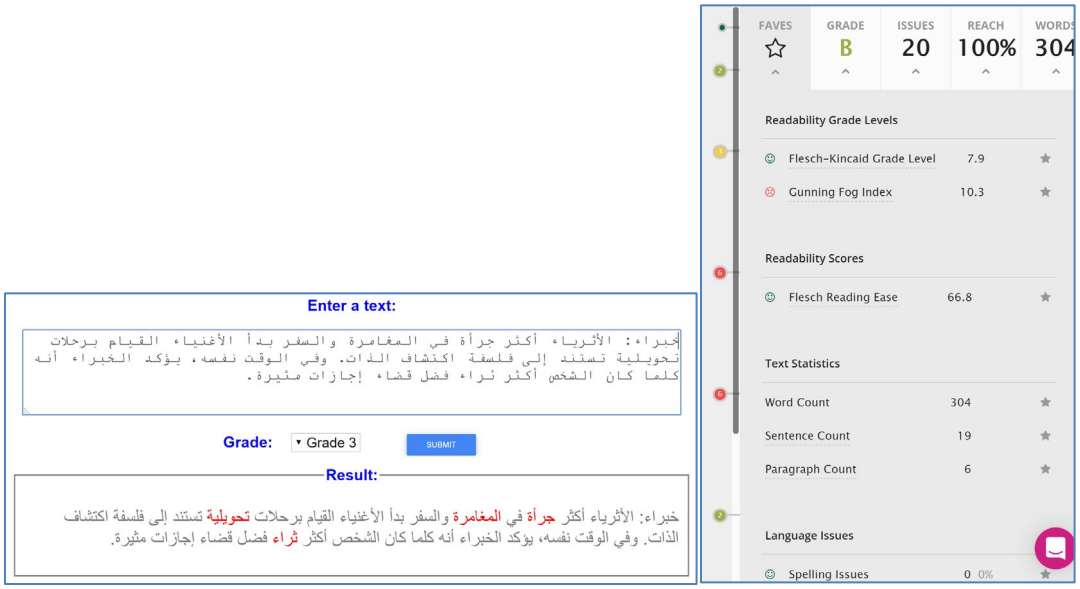
Arabic Text Readability Analysis

Project Description:

Motivation: Readability scores are a way to measure whether a written text is likely to be understood by the intended reader. Text that scores a readability grade level of 8 or below will be readable by around 85% of the general public. This can help in improving language skills for language learners, proficiency tests, text simplification, designing of educational materials, etc.

Available Resources: Arabic curricula for Gulf countries (Grades 1 to 6), Automatic processing (Farasa NLP tools) for diacritization, stemming, parsing, etc. huge Arabic corpora from Wikipedia, Aljazeera and other sources.

We already built an initial interface for spotting complex words that didn't appear in a selected grade.



Duties/Activities:

- We will study different readability formulas and apply to Arabic.
- We will build a website to provide readability analysis for an input text and give tips to improve its readability.

Required Skills:

- Java/Python
- Web development
- Understanding Arabic is a plus

Preferred Intern Academic Level:

- Junior/Senior levels

Learning Opportunities:

- This internship will help interns to immerse in coding and using various NLP libraries. It will help as well to get familiar with web development and build full stack profile. Also, you about recent research in reading analysis and apply them in real scenarios.

Expected Team Size: *it is preferable to have team projects*

- 2 to 4

Mentors

Name: Hamdy Mubarak email: hmubarak@hbku.edu.qa
Name: Ahmed Abdelali email: aabdelali@hbku.edu.qa
Name: Kareem Darwish email: kdarwish@hbku.edu.qa

Arabic Text Readability Analysis

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

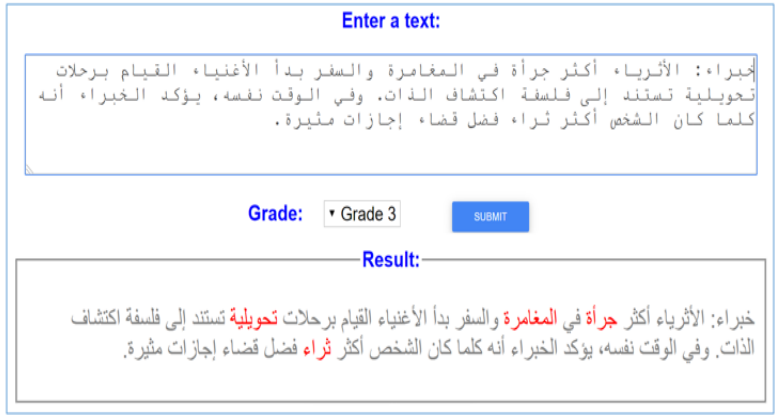
Mentors: Hamdy Mubarak, Ahmed Abdelali, Kareem Darwish {hmubarak, aabdelali, kdarwish}@hbku.edu.qa

Motivation: Readability scores are a way to measure whether a written text is likely to be understood by the intended reader. Text that scores a readability grade level of 8 or below will be readable by around 85% of the general public. This can help in improving language skills for language learners, proficiency tests, text simplification, designing of educational materials, etc.

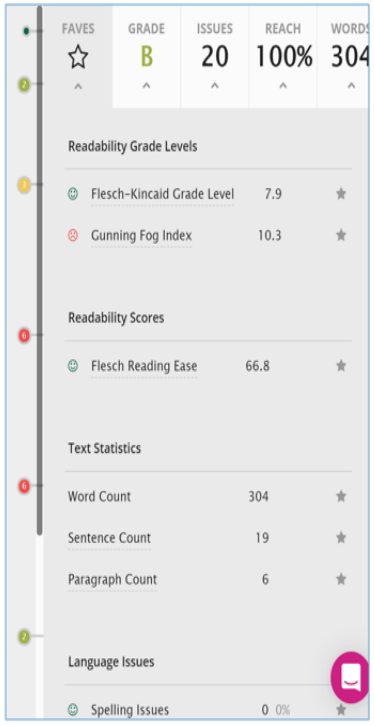
Available Resources: Arabic curricula for Gulf countries (Grades 1 to 6), Automatic processing (Farasa NLP tools) for diacritization, stemming, parsing, etc. huge Arabic corpora from Wikipedia, Aljazeera and other sources.

We already built an initial interface for spotting complex words that didn't appear in a selected grade.

- ❑ We will study different readability formulas and apply to Arabic.
- ❑ We will build a website to provide readability analysis for an input text and give tips to improve its readability.



Readability Analysis



❖ Skills

- ✓ Experience in Java/Python for data processing, calling APIs…
- ✓ Experience in web development: HTML, CSS
- ✓ Passionate about language analysis and education
- ✓ Understanding Arabic is a plus!

* Show possible word forms of a given lemma (stem) with pronunciation (TTS) and examples of usage:

Ex: term "reading" in Qatar (Grade 6)

قِراءة	قراءة	130	الْقِرَاءَةِ	القراءة	43	الْقِرَاءَةُ	القراءة	23	قِرَاءَةٌ	قراءة	16	قِراءةٌ	قراءة	12	القراءةِ	القراءة	11
			القراءةُ	القراءة	6	قِرَاءَةً	قراءة	6	لِقِرَاءَةِ	لقراءة	3	قراءةً	قراءة	2	القراءةِ	القراءة	1
			الْقِراءةِ	القراءة	1	الْقِرَاءةِ	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.

❖ 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!