Automatic Classification of Arabic ETDs

Eman Abdelrahman and Fatimah Alotaibi
Supervised by: Dr. Edward Fox
CS6604 Course Project
December 10th, 2019
Virginia Tech, Blacksburg, 24061

This document has our additional work about the “**Otrouha: Automatic Classification of Arabic ETDs**” research project.

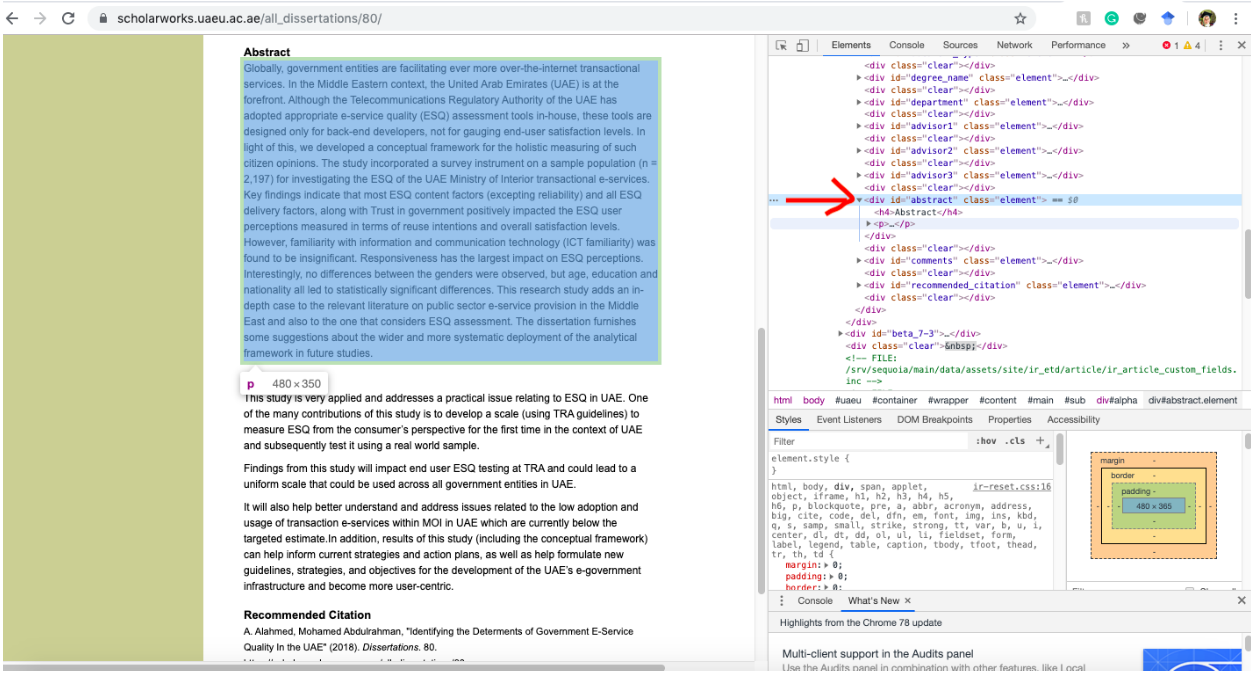
Data gathering:

● ProQuest and NDLTD.org

Initially, we looked for Arabic ETDs in the popular ETD digital libraries such as ndltd.org and ProQuest. We found that the ones tagged with the Arabic language were not really Arabic; some of them were Farsi or other different languages.

● UAEU

Then we chose United Arab Emirates University “Scholarworks @ UAEU” to collect Arabic ETDs. We inspected their website and found that it is highly crawlable, as shown in the figure below. Accordingly, we designed a crawler that can download the required data. Their digital library contains 75 dissertations and 687 theses. The ETDs that contained Arabic translation of their abstracts, keywords, and title are about 300 ETDs including different disciplines. Once we downloaded the ETDs, we harvested the metadata corresponding to each ETD and saved them as JSON files. The following is an example of the metadata harvested.



scholarworks.uaeu.ac.ae/all_dissertations/80/

Abstract

Globally, government entities are facilitating ever more over-the-internet transactional services. In the Middle Eastern context, the United Arab Emirates (UAE) is at the forefront. Although the Telecommunications Regulatory Authority of the UAE has adopted appropriate e-service quality (ESQ) assessment tools in-house, these tools are designed only for back-end developers, not for gauging end-user satisfaction levels. In light of this, we developed a conceptual framework for the holistic measuring of such citizen opinions. The study incorporated a survey instrument on a sample population (n = 2,197) for investigating the ESQ of the UAE Ministry of Interior transactional e-services. Key findings indicate that most ESQ content factors (excepting reliability) and all ESQ delivery factors, along with Trust in government positively impacted the ESQ user perceptions measured in terms of reuse intentions and overall satisfaction levels. However, familiarity with information and communication technology (ICT familiarity) was found to be insignificant. Responsiveness has the largest impact on ESQ perceptions. Interestingly, no differences between the genders were observed, but age, education and nationality all led to statistically significant differences. This research study adds an in-depth case to the relevant literature on public sector e-service provision in the Middle East and also to the one that considers ESQ assessment. The dissertation furnishes some suggestions about the wider and more systematic deployment of the analytical framework in future studies.

p 480 × 350

This study is very applied and addresses a practical issue relating to ESQ in UAE. One of the many contributions of this study is to develop a scale (using TRA guidelines) to measure ESQ from the consumer's perspective for the first time in the context of UAE and subsequently test it using a real world sample.

Findings from this study will impact end user ESQ testing at TRA and could lead to a uniform scale that could be used across all government entities in UAE.

It will also help better understand and address issues related to the low adoption and usage of transaction e-services within MOI in UAE which are currently below the targeted estimate.In addition, results of this study (including the conceptual framework) can help inform current strategies and action plans, as well as help formulate new guidelines, strategies, and objectives for the development of the UAE's e-government infrastructure and become more user-centric.

Recommended Citation

A. Alahmed, Mohamed Abdulrahman, "Identifying the Determents of Government E-Service Quality In the UAE" (2018). *Dissertations*. 80.

```
        <div class="clear"></div>
      ▶<div id="degree_name" class="element">…</div>
        <div class="clear"></div>
      ▶<div id="department" class="element">…</div>
        <div class="clear"></div>
      ▶<div id="advisor1" class="element">…</div>
        <div class="clear"></div>
      ▶<div id="advisor2" class="element">…</div>
        <div class="clear"></div>
      ▶<div id="advisor3" class="element">…</div>
        <div class="clear"></div>
...   ▼<div id="abstract" class="element"> == $0
          <h4>Abstract</h4>
        ▶<p>…</p>
        </div>
        <div class="clear"></div>
      ▶<div id="comments" class="element">…</div>
        <div class="clear"></div>
      ▶<div id="recommended_citation" class="element">…</div>
        <div class="clear"></div>
      </div>
    </div>
  ▶<div id="beta_7-3">…</div>
    <div class="clear">&nbsp;</div>
    <!-- FILE:
    /srv/sequoia/main/data/assets/site/ir_etd/article/ir_article_custom_fields.
    inc -->
```

html body #uaeu #container #wrapper #content #main #sub div#abstract.element

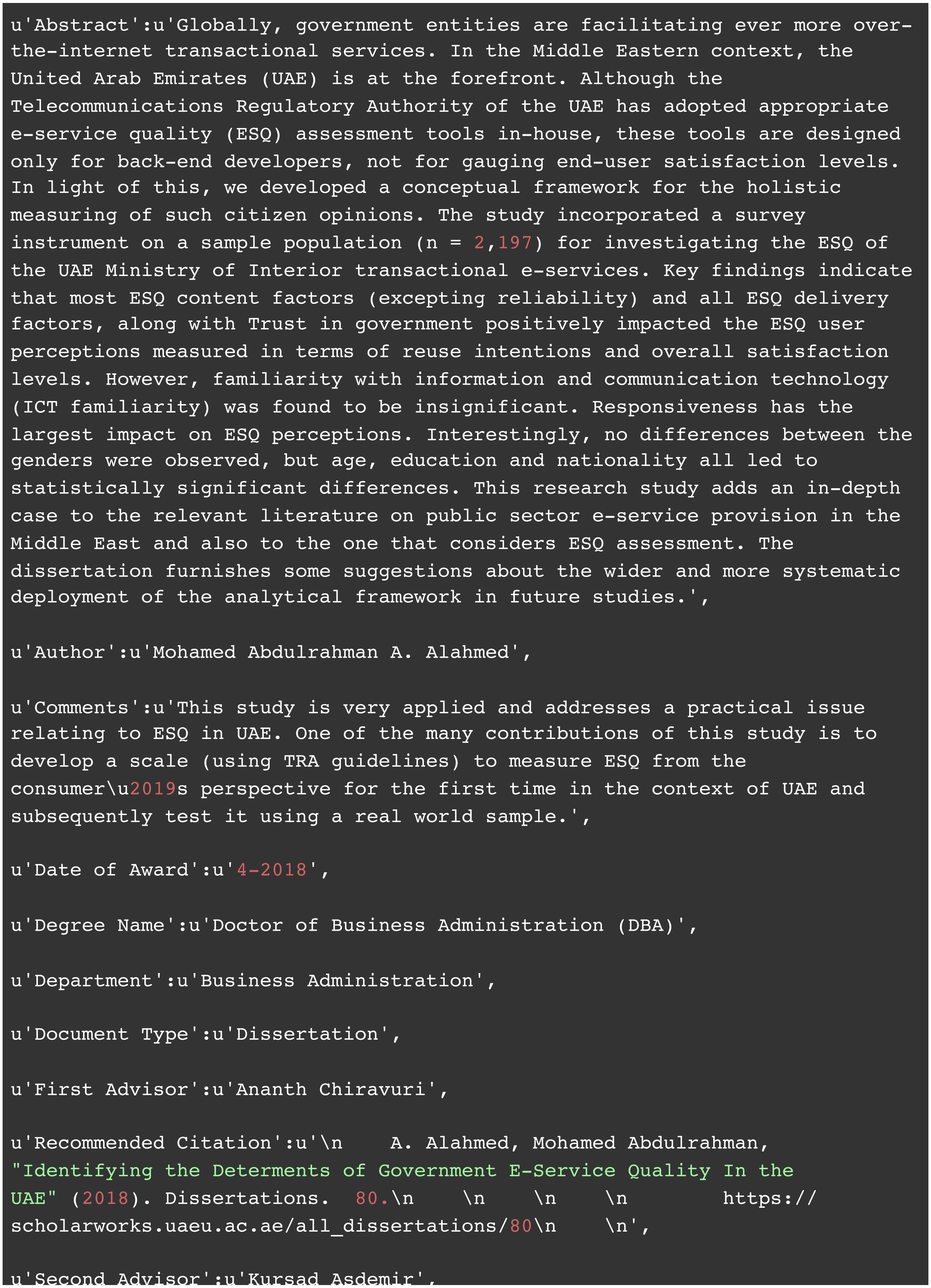
```
element.style {
}
html, body, div, span, applet,                          ir-reset.css:16
object, iframe, h1, h2, h3, h4, h5,
h6, p, blockquote, pre, a, abbr, acronym, address,
big, cite, code, del, dfn, em, font, img, ins, kbd,
q, s, samp, small, strike, strong, tt, var, b, u, i,
center, dl, dt, dd, ol, ul, li, fieldset, form,
label, legend, table, caption, tbody, tfoot, thead,
tr, th, td {
    margin: ▶ 0;
    padding: ▶ 0;
    border: ▶ 0;
```

Highlights from the Chrome 78 update

Multi-client support in the Audits panel

Use the Audits panel in combination with other features, like Local

Figure 1. Inspection and HTML tags of UAEU website



```
u'Abstract':u'Globally, government entities are facilitating ever more over-
the-internet transactional services. In the Middle Eastern context, the
United Arab Emirates (UAE) is at the forefront. Although the
Telecommunications Regulatory Authority of the UAE has adopted appropriate
e-service quality (ESQ) assessment tools in-house, these tools are designed
only for back-end developers, not for gauging end-user satisfaction levels.
In light of this, we developed a conceptual framework for the holistic
measuring of such citizen opinions. The study incorporated a survey
instrument on a sample population (n = 2,197) for investigating the ESQ of
the UAE Ministry of Interior transactional e-services. Key findings indicate
that most ESQ content factors (excepting reliability) and all ESQ delivery
factors, along with Trust in government positively impacted the ESQ user
perceptions measured in terms of reuse intentions and overall satisfaction
levels. However, familiarity with information and communication technology
(ICT familiarity) was found to be insignificant. Responsiveness has the
largest impact on ESQ perceptions. Interestingly, no differences between the
genders were observed, but age, education and nationality all led to
statistically significant differences. This research study adds an in-depth
case to the relevant literature on public sector e-service provision in the
Middle East and also to the one that considers ESQ assessment. The
dissertation furnishes some suggestions about the wider and more systematic
deployment of the analytical framework in future studies.',

u'Author':u'Mohamed Abdulrahman A. Alahmed',

u'Comments':u'This study is very applied and addresses a practical issue
relating to ESQ in UAE. One of the many contributions of this study is to
develop a scale (using TRA guidelines) to measure ESQ from the
consumer\u2019s perspective for the first time in the context of UAE and
subsequently test it using a real world sample.',

u'Date of Award':u'4-2018',

u'Degree Name':u'Doctor of Business Administration (DBA)',

u'Department':u'Business Administration',

u'Document Type':u'Dissertation',

u'First Advisor':u'Ananth Chiravuri',

u'Recommended Citation':u'\n    A. Alahmed, Mohamed Abdulrahman,
"Identifying the Determents of Government E-Service Quality In the
UAE" (2018). Dissertations.  80.\n    \n    \n    \n        https://
scholarworks.uaeu.ac.ae/all_dissertations/80\n    \n',

u'Second Advisor':u'Kursad Asdemir',
```

Table 1. Metadata of UAEU

The metadata contains the field of “Department” which is the ETD’s class. Since we were planning to be using the ProQuest categorization system, we mapped the Scholarworks@UAEU categories to ProQuest categories. For the categories that don’t exactly match, we chose the nearest discipline. The following figures 2 and 3 show examples of the UAEU and ProQuest Architecture category and its subcategories.

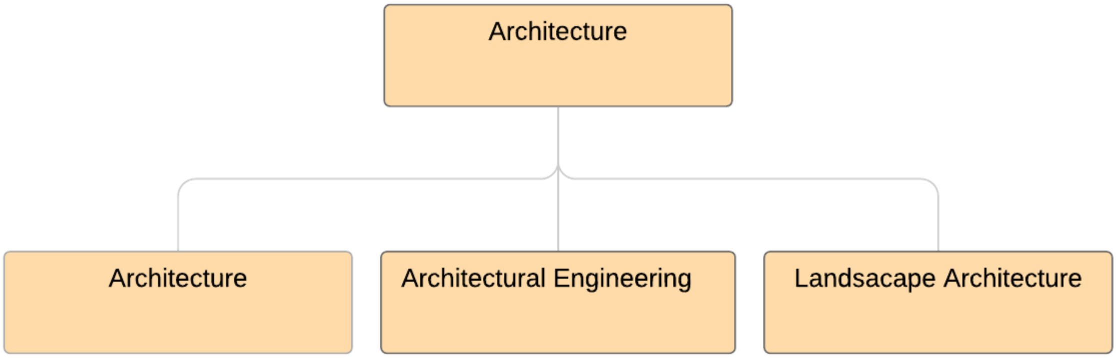


Figure 2. ProQuest Category of Architecture

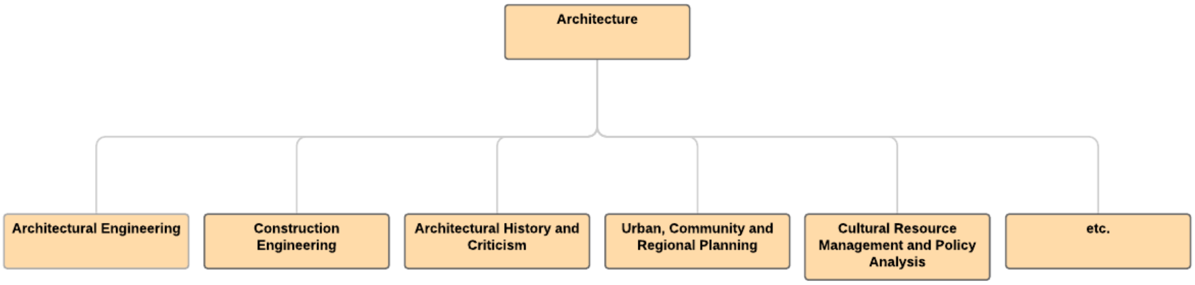
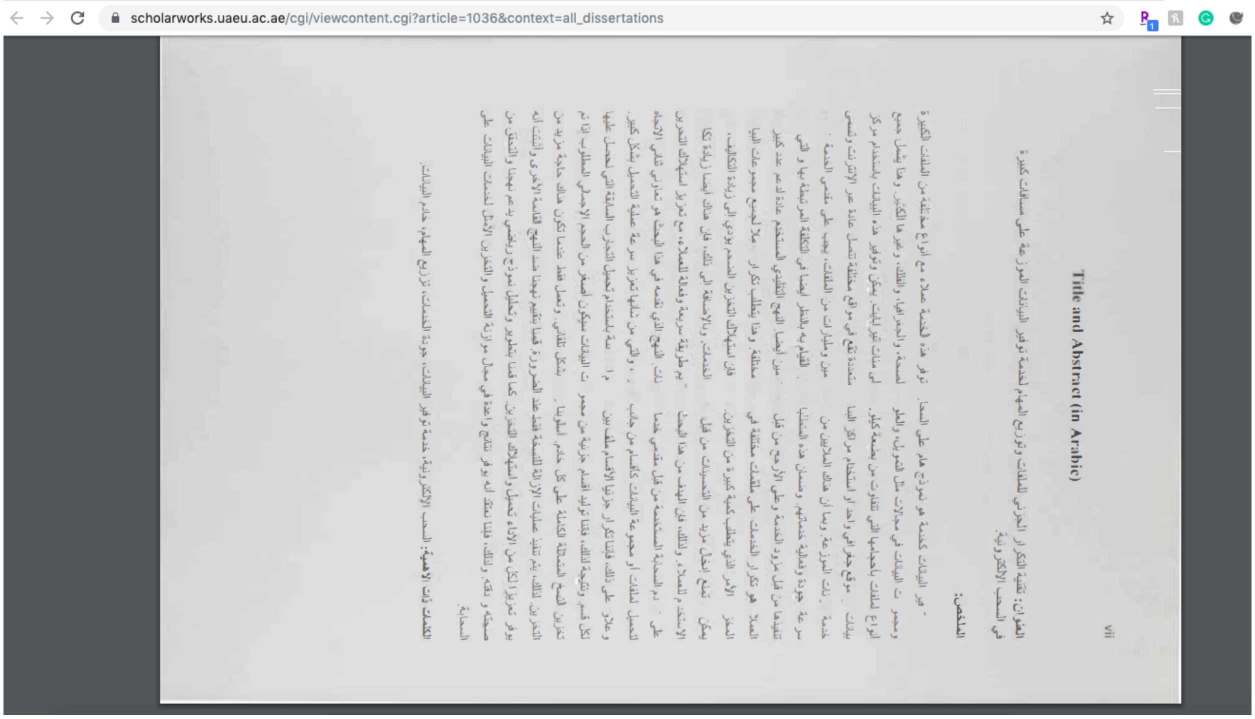

Architecture
Architectural Engineering
Construction Engineering
Architectural History and Criticism
Urban, Community and Regional Planning
Cultural Resource Management and Policy Analysis
etc.

Figure 3. UAEU Category of Architecture

Once we downloaded a number of ETDs to work on, we found that most of them are not native PDFs, but rather scanned PDFs as shown in Figure 4, which made extracting the text not straightforward.



scholarworks.uaeu.ac.ae/cgi/viewcontent.cgi?article=1036&context=all_dissertations

vii

Title and Abstract (in Arabic)

العنوان: تقنية التكرار الجزئي للملفات وتوزيع المهام لخدمة توفير البيانات الموزعة على مسافات كبيرة
في السحب الالكترونية.

الملخص:

فير البيانات كخدمة هو نموذج هام على السحا توفر هذه الخدمة عملاء مع أنواع مختلفة من الملفات الكبيرة
ومجمو ت البيانات في مجالات مثل التمويل، والعلو لصحة، والجغرافيا، والفلك، وغيرها الكثير. وهذا يشمل جميع
أنواع لملفات بأحجامها التي تتفاوت من بضعة كيلو. لى منات تيرابايت. يمكن وتوفير هذه البيانات باستخدام مركز
بيانات موقع جغرافي واحد أو استخدام مراكز البيا متعددة تقع في مواقع مختلفة تتصل عادة عبر الإنترنت وتسمى
خدمة نات الموزعة. وبما أن هناك الملايين من مين ومليارات من الملفات، يجب على مقدمي الخدمة
سرعة جودة وفعالية خدماتهم. وضمان هذه المتطلبا القيام به بالنظر أيضا في التكلفة المرتبطة بها و التي
تنفيذها من قبل مزود الخدمة وعلى الأرجح من قبل مين أيضا. النهج التقليدي المستخدم عادة لدعم عدد كبير
العملا هو تكرار الخدمات على ملقمات مختلفة في مختلفة. وهذا يتطلب تكرار ملا لجميع مجموعات البيا
المخز الأمر الذي يتطلب كمية كبيرة من التخزين. فإن استهلاك التخزين الضخم يودي إلى زيادة التكاليف،
يمكن تمنع إدخال مزيد من التحسينات من قبل الخدمات. وبالإضافة إلى ذلك، فإن هناك أيضا زيادة تكا
الاستخد م للعملاء. ولذلك، فإن الهدف من هذا البحث يم طريقة سريعة وفعالة للعملاء، مع تعزيز استهلاك التحرين
على دم السحابة المستخدمة من قبل مقدمي خدما نات. النهج الذي نقدمه في هذا البحث هو تعاوني ثنائي الاتجاه
لتحميل لملفات أو مجموعة البيانات كأقسام من جانب ، والتي من شأنها تعزيز سرعة عملية التحميل بشكل كبير.
وعلاو على ذلك، فإننا تكرار جزئيا الأقسام ملف بين م ا سة باستخدام تحميل التجارب السابقة التي نحصل عليها
لكل قسم. ونتيجة لذلك، فإننا توليد أقسام جزئية من مجمو ت البيانات سيكون أصغر من الحجم الاجمالي المطلوب اذا تم
تخزين النسخ المتماثلة الكاملة على كل خادم. أسلوبنا بشكل تلقائي. وتعمل فقط عندما تكون هناك حاجة مزيد من
التخزين. لذلك، يتم تنفيذ عمليات الإزالة للنسخة فقط عند الضرورة. قمنا بتقييم نهجنا ضد النهج القائمة الأخرى وأثبتت أنه
يوفر تعزيزا لكل من الاداء تحميل واستهلاك التخزين. كما قمنا بتطوير وتحليل نموذج رياضي يدعم نهجنا والتحقق من
صحته و دقته. ولذلك، فإننا نعتقد أنه يوفر نتائج واعدة في مجال موازنة التحميل والتخزين الامثل لخدمات البيانات على
السحابة.

الكلمات ذات الاهمية: السحب الإلكترونية، خدمة توفير البيانات، جودة الخدمات، تززيع المهام، خادم البيانات.

Figure 4. Example of scanned PDF from UAE library

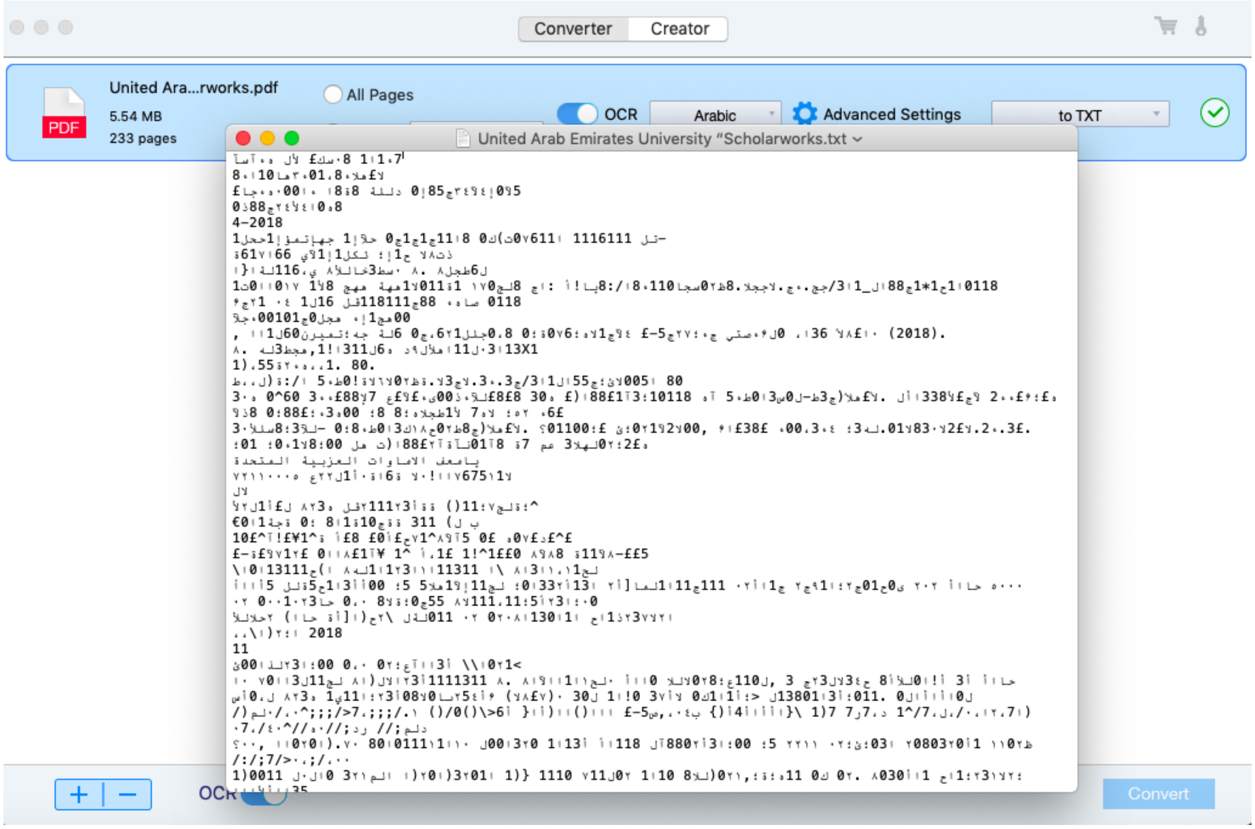
Arabic OCRs tools

To solve the aforementioned issue and extract the text, we tried different Arabic OCR tools such as ABBYY FineReader, Enolsoft, and CISDEM. Unfortunately, none of these tools has worked. They convert characters incorrectly and give unreadable output as shown in Figure 4. So we decided to use another source for Arabic ETDs. There are some other Arabic digital libraries that contain Arabic ETDs in terms of the amount and labeling. However, these digital libraries are geo-restricted so can not be accessed from US.

We found AskZad that is part of the Saudi Digital Library because it has a large number of Arabic ETDs and most importantly, it provides the Abstract as metadata in their webpage.

Create student account at Saudi Digital Library

In order to have access to the Saudi Digital Library, you need to have a student account. Creating an account there has many restrictions. The user has to be a Saudi student who is enrolled in one of the Saudi universities, or a Saudi student who studies abroad. Fatimah asked for an "Mubtath" account through the "Saudi Arabia Cultural Mission". This process took a long time which held us back from making progress in our project.

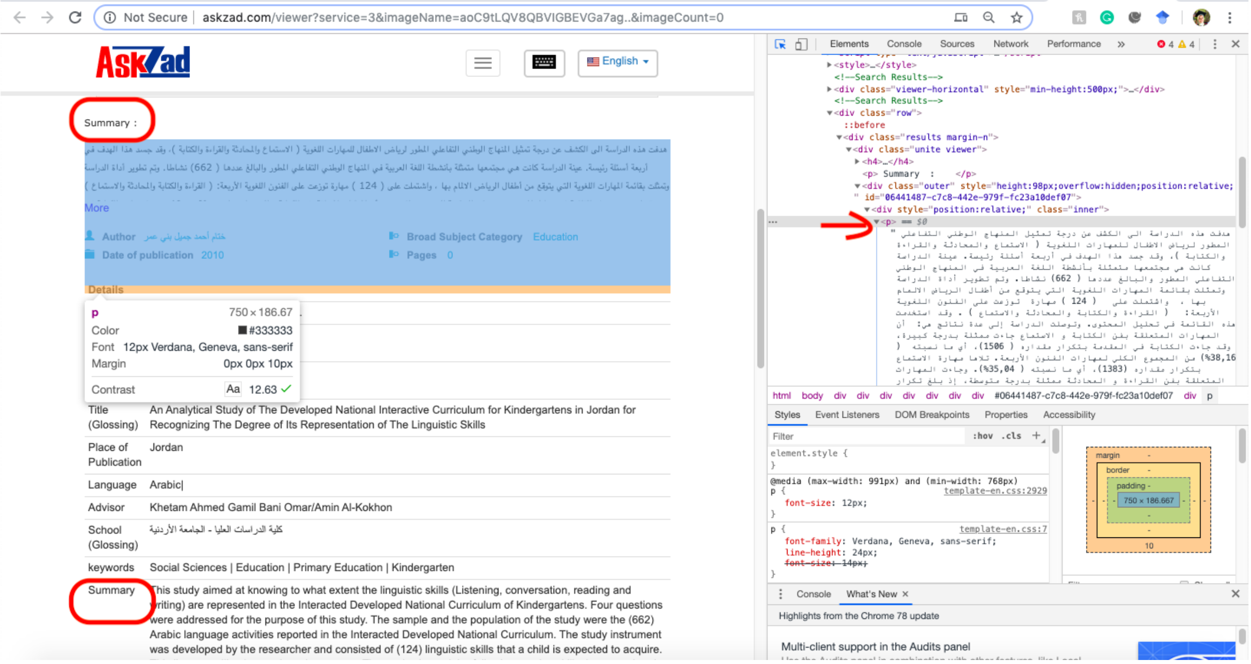


```
آسآء• لأل £سك•8 1؛1،7
8•؛10هـا٢•01.8•ملا£لا
£جا•••00؛• ؛8ة8 دللة 0؛85ع٣٤٩٤؛0٩5
0ذ88ع٢٤لا٤؛0،8
4-2018
1حل1إ‌جهإتمؤ 1إ‌حلا 0ع1ع1ع11؛8 0ك(ت0٧611؛ 1116111 تل-
ة61٧؛66 لاي1إ1تكل ؛إ1ع ذتلا٨
؛)}؛لة116،ي ٨لالاخ3سط• ٨. ل6طجل٨
1ت0؛؛0؛٧ 1لا8 مهج مهة1لا011ة1 ؛٧0لج8 ؛ج : أ!؛يا8:/؛8،110سجا0٢ط8.لاجحلا.ع••جج/3؛1_ل؛88ع1*1ع1؛0118
ع٢1 •٤ 1ل16 قل118111ع88 ••صاد 0118
جلا•00101ع0مجل ؛• إ1مج00
, ؛؛1ل60تميرن؛جه لة6 0ع،6٢1ل0.8جلل 0؛ة0٧6؛هلا1ع٩لا ٤ £-5ع٢٧؛•ع صتي•ل0 ،؛36 لا٨£؛• (2018).
٨. مجطلة3,1!؛311ل6• دلالهق؛11ل•3؛13X1
1).55ة٢•••1. 80.
ط••ل)ة:/؛ 5•ط0!ةلا٦٦0٢ط.ةلا3ع.3••3ع/3؛1ل55ع؛لاى005؛ 80
3•، 0^60 3••£88لا7 ع£٩£•ى00ذ•لا٥£8£8 30، £)؛88£1آ3:10118 ،آ 5•ط0؛3س0ل-ط3ع)ملا£لا. أل؛338لا£ع٩ 2••£٠:£،
٩ذ8 0:88£:•3،00 :8 8:طجلا،1لا 7،لا :٢• •6£
3•لالاس8:3لا٥- 0:8•ط0؛3ك٨ا0ج٢ط8ع)ملا£لا. ؟01100:£ ى:0٢1٩2لا00, ٠؛£38£ •00.3•٤ :3لـه.01لا83•لا2£لا.2••3£.
:01 :0•1لا8:00 ل مل ت)؛88£٢آة آنة01آ8 ة7 مع 3للهلا0٢:2£،
بامعف الاماوات العزبية المتحدة
٧٢١١•••ه ع٢٢ل1أ•ة؛6ة لا•!؛؛٧675؛1لا
لال
لا٢ل1أ£ل ٨٢3، قل٢111٢3أةة ()11:٧قلج؛:^
€0؛1ةجة 0: 8؛1ة10ججة 311 (ل ب
10£^آ!£¥1^ة أ£8 £0أ£ج٧1^٨٩آ5 0£ ،0٧£ذ£^£
£-ة£٩٧1٢£ 0؛؛٨£1آ¥ 1^ أ.1£ 1!^1££0 ٨٩٨8 ة11٩٨-££5
\؛0؛13111ج)؛ ٨لده1؛1٢3؛؛؛11311 ؛\ ٨؛3؛؛،؛1لج
أ؛؛ة5 قلل5ج1؛3أ؛00 :5 5ملا1٩إ11لج :0؛33٢أ13؛ ٢أ]لما1؛11ع111 •٢أ؛1ع ٢ع٩1؛:٢ع01ع0ى ٢•٢ أ؛اح ه•••
•٢ 0••1•٢3ح 0.• 8لاة:0ع55 ٨لا111.11:5أ٢3؛:•0
لالاحلا٢ (؛اح ةأ]؛)ع٢\ لق011 •٢ 0٢•٨؛130؛1؛ ع1ذ٢3٧لا٢؛
••\؛)٢:؛ 2018
11
ق00؛لل٢3؛:00 0.• 0٢:عآ؛؛3أ \\؛0٢1<
؛• ٧0؛؛3ل11لج ٨؛)لال٢؛3أ1111311 ٨. ٨؛٩؛؛؛1؛لج• أ؛؛0 لالـ0٢8:ع110ل, 3 ع٢3لال3ع٤ 8أ٧لـ0؛!أ 3أ أ؛اح
س0أ.ل ٨٢3، 1ي11؛:٢3أ08لا0ل٢5٤أ٠ (لا٨£٧)• 30ل 1؛!0 3٧أ٧ 0ك1؛1أ:> ل13801؛3أ:011. 0ل؛أ؛أ؛0ل
/)مل•/.•^;;;/>7.;;;/.؛ ()/0()\>6أ }؛أ)؛؛()؛؛؛ £-5ص,••٤ب {)أ4؛أ؛أ؛}\ 1)7 7ر7.د 1^/7.ل./•.؛٢.7؛)
•7./٤•^//ه•//;رد //;ملد
؟••, ؛؛0٢0؛).٧• 80؛0111؛1؛؛• ل00؛3٢0 1؛13أ أ؛118 لآ880٢أ3؛:00 :5 ٢٢؛؛ •٢:ق:03؛ ٢0803٢0أ1 ؛؛0٢ط
/:/;7/>•.;/.••
1)0011 ل•ل؛0 3٢؛مل؛ ؛)٢0؛)3٢01؛ 1)} 1110 ٧11ل0٢ 1؛10 8لا)0٢؛,:ة:ه11 0ك 0٢. ٨030أ؛1 ع1:٢3؛لا٢:
؛؛؛لا؛؛35
```

Figure 5. The output of CISDEM OCR tool

- AskZad

There were two challenges when crawling the AskZad Digital Library website. First, the website's crawlability is very low. Therefore, we had to use the full XPath of each element we are looking for. While revising the data, we found that some of the data contains a lot of noise due to the change of the website structure. This leads to getting elements that are not what we are looking for. For example, getting the Advisor Name instead of getting the Abstract. The second challenge was that the Crawl Rate Limit is 45, so among the ETDs downloaded within that limit, some of them weren't what we are looking for. These two challenges significantly reduced the data size.


Not Secure | askzad.com/viewer?service=3&imageName=aoC9tLQV8QBVIGBEVGa7ag..&imageCount=0
AskZad
English
Elements
Console
Sources
Network
Performance
»
4
4

Summary :

هدفت هذه الدراسة الى الكشف عن درجة تمثيل المنهاج الوطني التفاعلي المطور لرياض الاطفال للمهارات اللغوية (الاستماع والمحادثة والقراءة والكتابة)، وقد جسد هذا الهدف في
أربعة أسئلة رئيسة. عينة الدراسة كانت هي مجتمعها متمثلة بأنشطة اللغة العربية في المنهاج الوطني التفاعلي المطور والبالغ عددها (662) نشاطا. وتم تطوير أداة الدراسة
وتمثلت بقائمة المهارات اللغوية التي يتوقع من أطفال الرياض الالمام بها ، واشتملت على (124) مهارة توزعت على الفنون اللغوية الأربعة: (القراءة والكتابة والمحادثة والاستماع)

More

Author ختام أحمد جميل بني عمر
Date of publication 2010
Broad Subject Category Education
Pages 0

Details

p 750 × 186.67
Color ■ #333333
Font 12px Verdana, Geneva, sans-serif
Margin 0px 0px 10px
Contrast Aa 12.63 ✓

Title (Glossing)	An Analytical Study of The Developed National Interactive Curriculum for Kindergartens in Jordan for Recognizing The Degree of Its Representation of The Linguistic Skills
Place of Publication	Jordan
Language	Arabic|
Advisor	Khetam Ahmed Gamil Bani Omar/Amin Al-Kokhon
School (Glossing)	كلية الدراسات العليا - الجامعة الأردنية
keywords	Social Sciences | Education | Primary Education | Kindergarten
Summary	This study aimed at knowing to what extent the linguistic skills (Listening, conversation, reading and writing) are represented in the Interacted Developed National Curriculum of Kindergartens. Four questions were addressed for the purpose of this study. The sample and the population of the study were the (662) Arabic language activities reported in the Interacted Developed National Curriculum. The study instrument was developed by the researcher and consisted of (124) linguistic skills that a child is expected to acquire.

```
▶<style>…</style>
 <!--Search Results-->
▶<div class="viewer-horizontal" style="min-height:500px;">…</div>
 <!--Search Results-->
▼<div class="row">
   ::before
  ▼<div class="results margin-n">
    ▼<div class="unite viewer">
      ▶<h4>…</h4>
       <p> Summary  :    </p>
      ▼<div class="outer" style="height:98px;overflow:hidden;position:relative;
      " id="06441487-c7c8-442e-979f-fc23a10def07">
        ▼<div style="position:relative;" class="inner">
...        ▼<p> == $0
            " هدفت هذه الدراسة الى الكشف عن درجة تمثيل المنهاج الوطني التفاعلي
            المطور لرياض الاطفال للمهارات اللغوية ( الاستماع والمحادثة والقراءة
            والكتابة )، وقد جسد هذا الهدف في أربعة أسئلة رئيسة. عينة الدراسة
            كانت هي مجتمعها متمثلة بأنشطة اللغة العربية في المنهاج الوطني
            التفاعلي المطور والبالغ عددها ( 662) نشاطا. وتم تطوير أداة الدراسة
            وتمثلت بقائمة المهارات اللغوية التي يتوقع من أطفال الرياض الالمام
            بها ، واشتملت على ( 124 ) مهارة توزعت على الفنون اللغوية
            الأربعة: ( القراءة والكتابة والمحادثة والاستماع ) . وقد استخدمت
            هذه القائمة في تحليل المحتوى. وتوصلت الدراسة إلى عدة نتائج هي: أن
            المهارات المتعلقة بفن الكتابة و الاستماع جاءت ممثلة بدرجة كبيرة،
            وقد جاءت الكتابة في المقدمة بتكرار مقداره ( 1506)، أي ما نسبته (
            %38,16) من المجموع الكلي لمهارات الفنون الأربعة. تلاها مهارة الاستماع
            بتكرار مقداره (1383)، أي ما نسبته ( %35,04). وجاءت المهارات
            المتعلقة بفن القراءة و المحادثة ممثلة بدرجة متوسطة، إذ بلغ تكرار
html body div div div div div div div #06441487-c7c8-442e-979f-fc23a10def07 div p
```

```
element.style {
}
@media (max-width: 991px) and (min-width: 768px)
p {                                   template-en.css:2929
  font-size: 12px;
}
p {                                   template-en.css:7
  font-family: Verdana, Geneva, sans-serif;
  line-height: 24px;
  ~~font-size: 14px;~~
}
```

Highlights from the Chrome 78 update

Multi-client support in the Audits panel

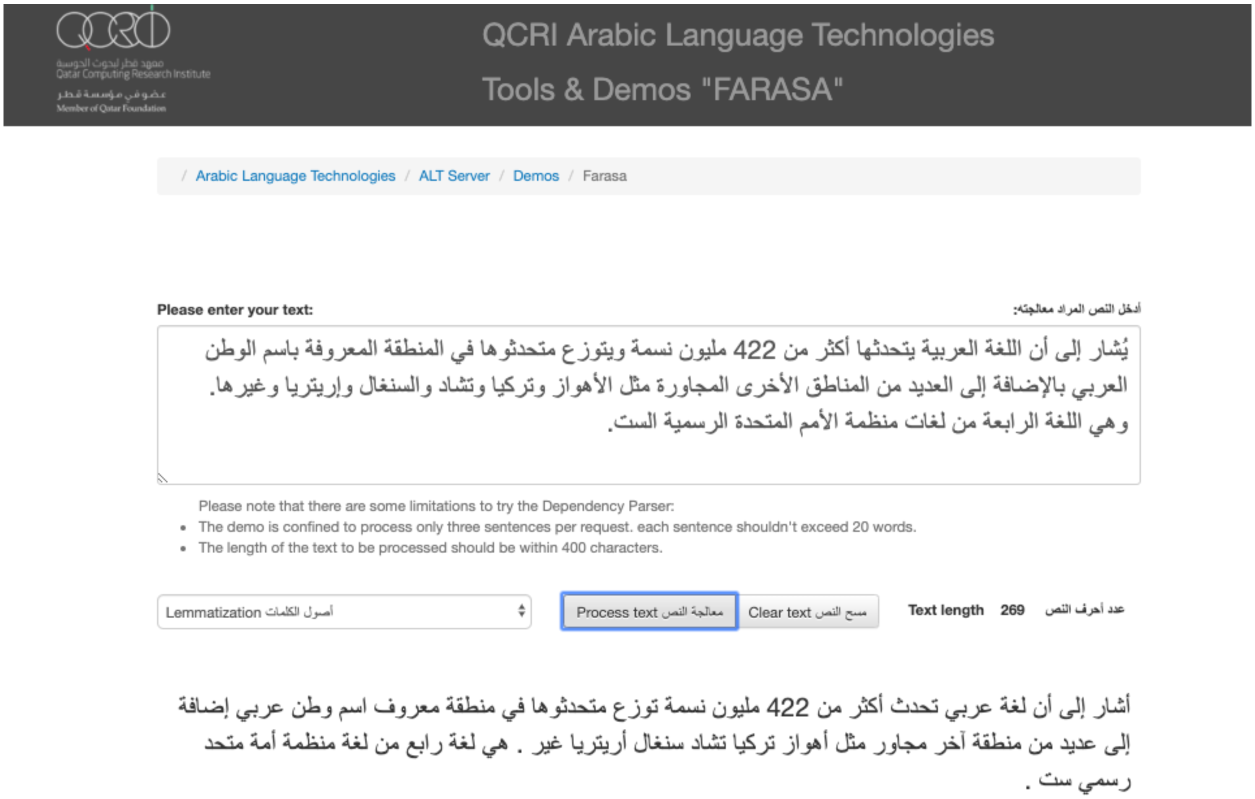
Figure 6. Inspection of Askzad page

Data Preprocessing

Stemming vs. lemmatization

First, we used different stemmers to preprocess our data. We gave the same input text for these stemmers and each of them provided a different output based on the way each of them stems words. Also, these stemmers give the stem of the word which most of the time changes the meaning of the context. Throughout our research we found that lemmatization shows more efficiency, especially in the Arabic language, since it is a highly inflectional and derivational language. As a result, we chose the FARASA lemmatizer.

Figures 7, 8, and 9 are illustrative examples of the difference between stemming and lemmatization in the Arabic language.


معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
عضو في مؤسسة قطر
Member of Qatar Foundation

QCRI Arabic Language Technologies

Tools & Demos "FARASA"

/ Arabic Language Technologies / ALT Server / Demos / Farasa

Please enter your text: أدخل النص المراد معالجته:

يُشار إلى أن اللغة العربية يتحدثها أكثر من 422 مليون نسمة ويتوزع متحدثوها في المنطقة المعروفة باسم الوطن
العربي بالإضافة إلى العديد من المناطق الأخرى المجاورة مثل الأهواز وتركيا وتشاد والسنغال وإريتريا وغيرها.
وهي اللغة الرابعة من لغات منظمة الأمم المتحدة الرسمية الست.

Please note that there are some limitations to try the Dependency Parser:

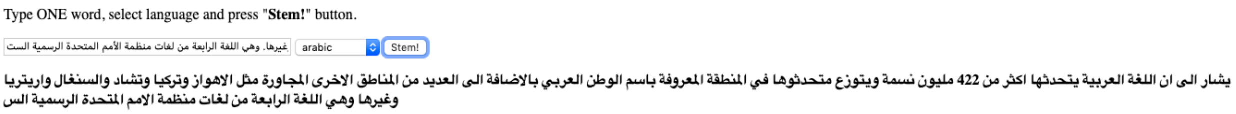
- The demo is confined to process only three sentences per request. each sentence shouldn't exceed 20 words.
- The length of the text to be processed should be within 400 characters.

Lemmatization أصول الكلمات

Text length 269 عدد أحرف النص

أشار إلى أن لغة عربي تحدث أكثر من 422 مليون نسمة توزع متحدثوها في منطقة معروف اسم وطن عربي إضافة
إلى عديد من منطقة آخر مجاور مثل أهواز تركيا تشاد سنغال أريتريا غير . هي لغة رابع من لغة منظمة أمة متحد
رسمي ست .

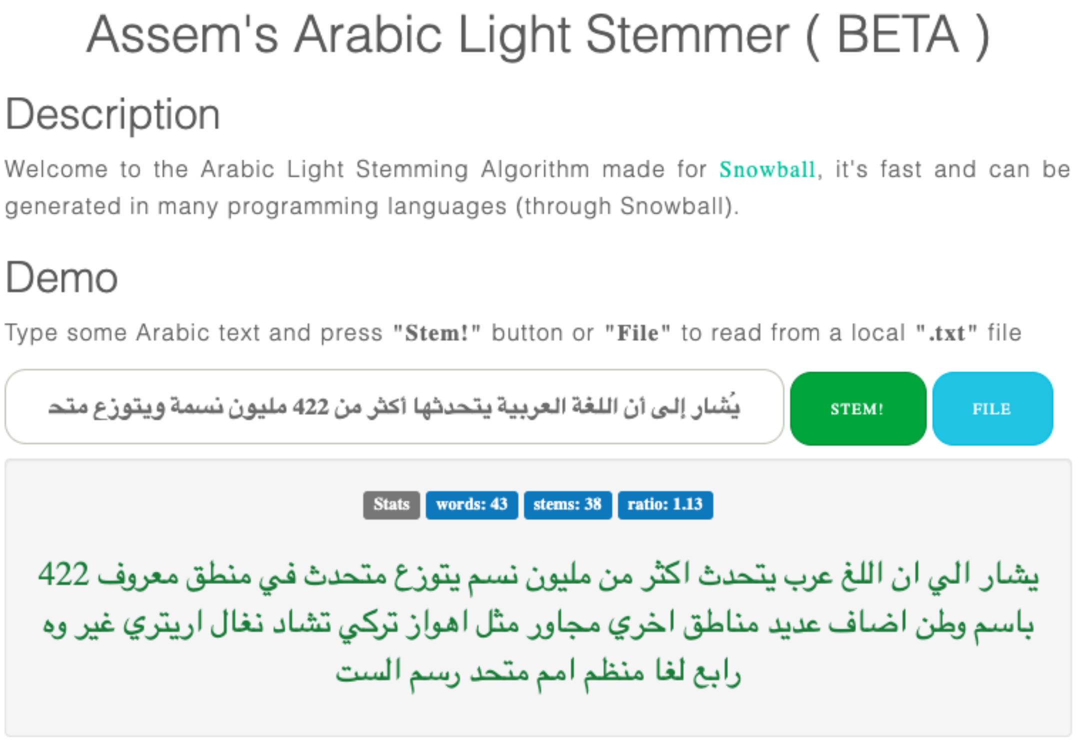
Figure 7. Result of lemmatization by Farasa


Type ONE word, select language and press "**Stem!**" button.

غيرها. وهي اللغة الرابعة من لغات منظمة الأمم المتحدة الرسمية الست arabic

يشار الى ان اللغة العربية يتحدثها اكثر من 422 مليون نسمة ويتوزع متحدثوها في المنطقة المعروفة باسم الوطن العربي بالاضافة الى العديد من المناطق الاخرى المجاورة مثل الاهواز وتركيا وتشاد والسنغال واريتريا
وغيرها وهي اللغة الرابعة من لغات منظمة الامم المتحدة الرسمية الس

Figure 8. Results of Snowball Arabic Stemmer



Assem's Arabic Light Stemmer (BETA)

Description

Welcome to the Arabic Light Stemming Algorithm made for Snowball, it's fast and can be generated in many programming languages (through Snowball).

Demo

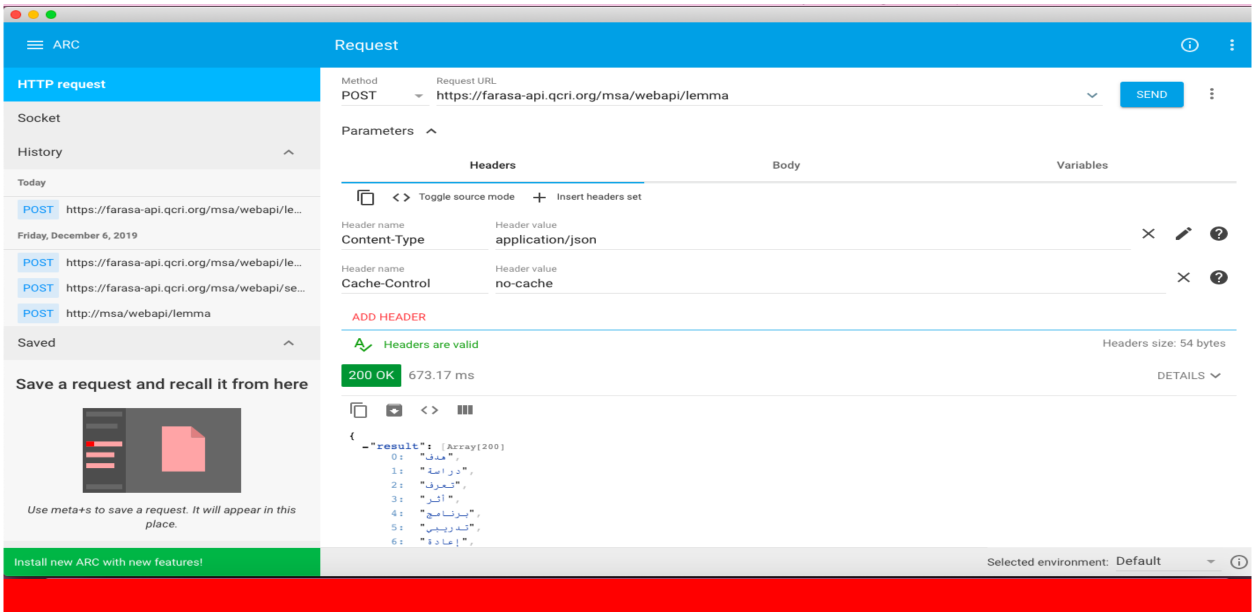
Type some Arabic text and press "**Stem!**" button or "**File**" to read from a local "**.txt**" file

يُشار إلى أن اللغة العربية يتحدثها أكثر من 422 مليون نسمة ويتوزع متح

يشار الي ان اللغ عرب يتحدث اكثر من مليون نسم يتوزع متحدث في منطق معروف 422
باسم وطن اضاف عديد مناطق اخري مجاور مثل اهواز تركي تشاد نغال اريتري غير وه
رابع لغا منظم امم متحد رسم الست

Figure 9. Results of stemming by Assem

During our work with the FARASA API, we needed to send a post request and make sure that it sends valid data back every time. To do that, we used the Advanced REST client tool. Figure 10 shows how this tool works.



POST https://farasa-api.qcri.org/msa/webapi/le...

Friday, December 6, 2019

POST https://farasa-api.qcri.org/msa/webapi/le...

POST https://farasa-api.qcri.org/msa/webapi/se...

POST http://msa/webapi/lemma

Save a request and recall it from here

Use meta+s to save a request. It will appear in this place.

Method: POST

Request URL: https://farasa-api.qcri.org/msa/webapi/lemma

Header name: Content-Type — Header value: application/json

Header name: Cache-Control — Header value: no-cache

Headers are valid

Headers size: 54 bytes

200 OK 673.17 ms

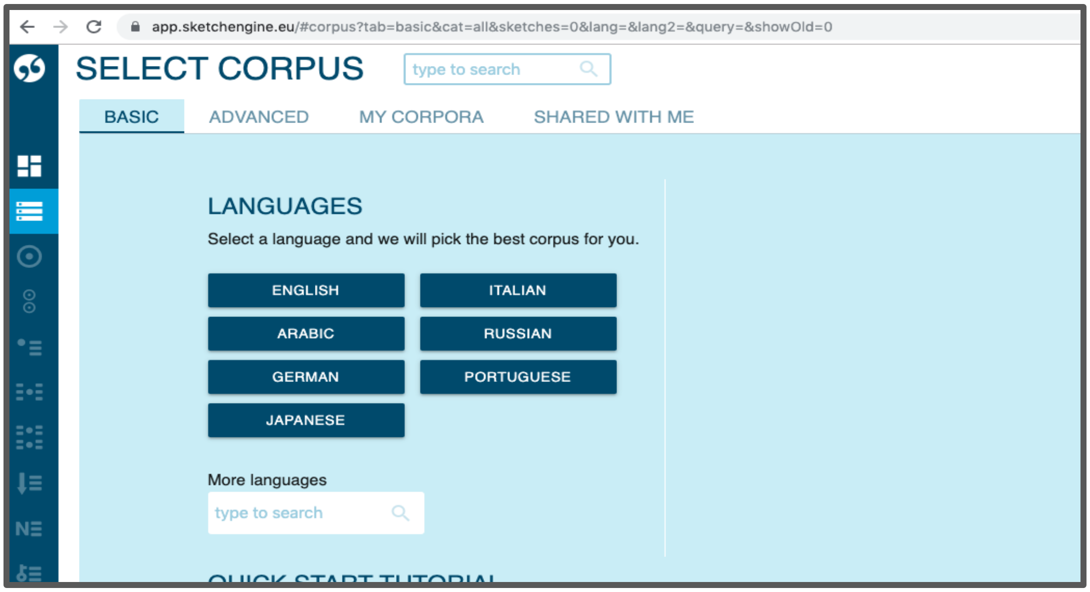
```
{
  -"result": [Array[200]
     0:  "هدف",
     1:  "دراسة",
     2:  "تعرف",
     3:  "أثر ",
     4:  "برنامج",
     5:  "تدريبي",
     6:  "إعادة",
```

Selected environment: Default

Figure 10.Rest client tool connected with FARSA

Classification Process

After we prepared clean data, we ran the SVM model for training. However, the performance was very low.
We tried the tree decision classifier and there was no significant difference.
In this first experiment which is multi-class classification, none of the used models gave a good accuracy. We finally tried Random Forest, which gave us a slight higher difference, so we decided to choose it for a second experiment which is binary classification.
In the binary classification, Random Forest gave a satisfactory results based on the recall, precision, and F-1 measures, and the accuracy. Finally, we think that working on larger dataset will make the accuracy better and the results more credible. Therefore, we want to increase our corpus for further investigation and benchmarking. Also, we aim to make this corpus available for researchers who want to pursue the work on Arabic ETDs. To do that, we will use Sketch Engine. This engine helps building corpora in different languages.



app.sketchengine.eu/#corpus?tab=basic&cat=all&sketches=0&lang=&lang2=&query=&showOld=0

SELECT CORPUS

LANGUAGES

Select a language and we will pick the best corpus for you.

More languages

Figure 11. Sketch Engine