أدوات فراسة لمعالجة النص العربي

التحليل اللغوي للنص العربي

مايو 2023

المحتوى

ملخص

شهد مجال معالجة اللغات الطبيعية (NLP) تطورات سريعة مع زيادة القوة الحسابية وتوافر البيانات النصية وبيانات الوسائط المتعددة. بذلت مجموعة تقنيات اللغة العربية في معهد قطر لبحوث الحوسبة جهودًا متضافرة لتعزيز أبحاث معالجة اللغات الطبيعية للغة العربية ولهجاتها وبناء موارد لغوية (بيانات وبرامج) لتسريع الابتكار في هذا المجال.

شكلت صعوبات النص العربي والصرف والنحو للغة العربية الفصيحة ولهجاتها فرصًا لتطوير منهجيات لمعالجة اللغات الطبيعية بشكل عام، وتطوير أساليب وموارد للغة العربية بشكل خاص.

طورت مجموعة تقنيات اللغة العربية مجموعة أدوات فعالة للغاية للمعالجة الآلية الكاملة للنص العربي. يشمل ذلك التحليل الصرفي والنحوي والدلالي.

تم استخدام الموارد والتقنيات اللغوية التي تنتجها المجموعة دوليًا مما أدى إلى زيادة دعم للغة العربية في مجال الأبحاث وأيضا من خلال المنتجات المتاحة تجاريًا. وحفز عمل المجموعة جهودًا مماثلة لدعم اللغات ذات الموارد المحدودة.

يمكن تجربة الأدوات وأيضا واجهة تطبيق البرامج APIs من الموقع: https://farasa.qcri.org

واجهة تطبيق أدوات فراسة

فراسة هي مجموعة من الأدوات لمعالجة وتحليل المحتوى المكتوب باللغة العربية الفصحى الحديثة واللهجات العربية وكذلك اللغة التراثية (الحديث الشريف مثلا). وهي تشمل الوظائف التالية:

- التقطيع الصرفي
- تحديد أقسام الكلام Part of Speech (POS) tagging
- أصول الكلمات Lemmatization
- التشكيل
- التصحيح الإملائي
- التعرف على الأعلام Named Entity Recognition (NER)
- الإعراب

التكامل

تتم مشاركة أدوات فراسة كأكواد مفتوحة المصدر بلغة جافا. تتكامل الأدوات بطريقة بسيطة مع التطبيقات لتتيح مهام أكثر تعقيدًا، على سبيل المثال:

- يعتبر التقطيع الصرفي جزءًا أساسيًا من نظام الترجمة الآلية، مثلا نظام شاهين للترجمة https://mt.qcri.org/
- أصول الكلمات ضرورية لاسترجاع المعلومات
- التشكيل ضروري لتطبيق تحويل النص إلى كلام، مثلا نظام ناطق لنطق النصوص https://tts.qcri.org/
- التعرف على الأعلام والإعراب مهمان لفهم الوثائق ولأنظمة الأسئلة والأجوبة

الدقة

يتم مقارنة دقة أدوات فراسة باستمرار مقابل التقنيات المماثلة، وتحسينها لزيادة الدقة والسرعة للغة العربية الفصحى المعاصرة ومجموعة من اللهجات العربية. تعطي معظم الأدوات أفضل النتائج من منظور الدقة والسرعة.

الخوارزميات

تستخدم الأدوات مجموعة متنوعة من تقنيات التعلم الآلي والعميق. من بين التقنيات:

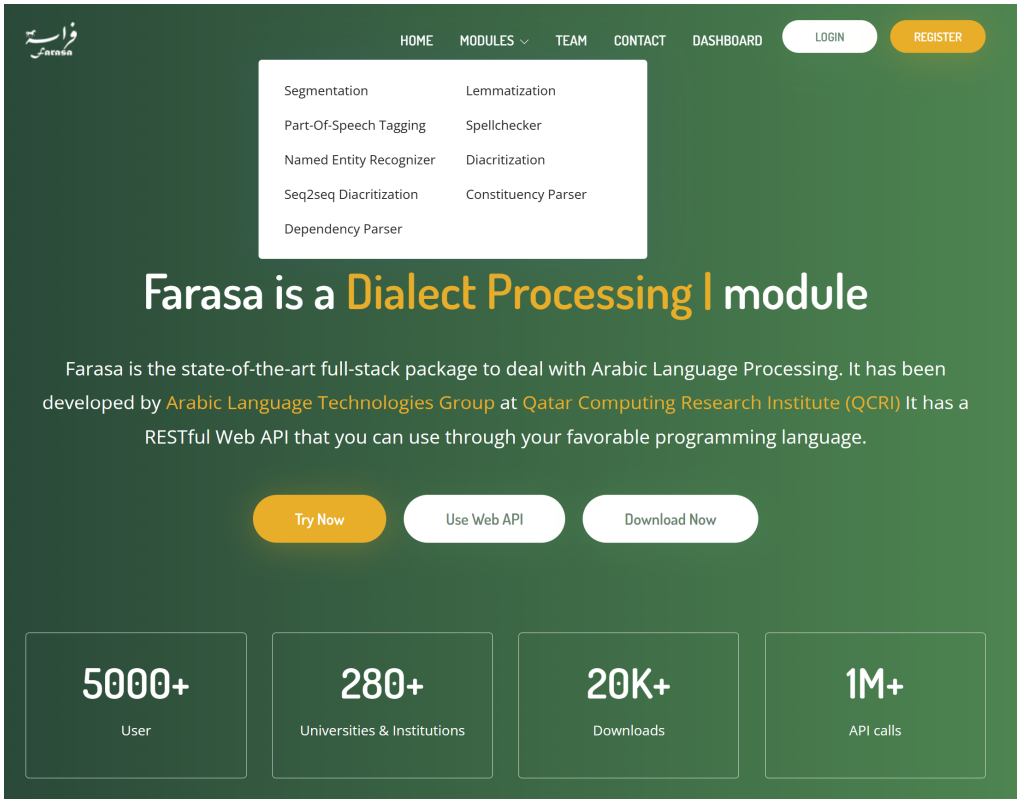
SVM و CRF و DNN و Transformers و Neural MT وغيرها.

تم تدريب البرامج على موارد لغوية تم الحصول عليها من مؤسسات مثل LDC الأمريكية (على سبيل المثال، بنك أشجار الإعراب العربية Arabic TreeBank)، بالإضافة إلى البيانات التي تم إنشاؤها في معهد قطر لبحوث الحوسبة.

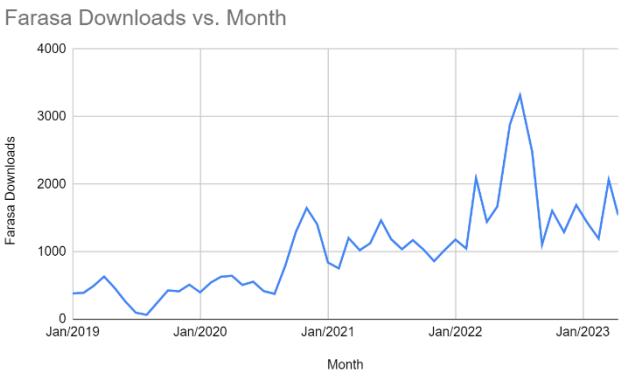
Project*	Accuracy %	State-of-the-art	Algorithm	Conferences
Segmentation	**98.9 (MSA)** **93.4 (DA)**	MADAMIRA (98.8) **MADAMIRA (EGY) (97.5)**	SVM biLSTM	NAACL'16, LREC'16 CoNLL'17, arXiv'17
POS Tagging	**96.3 (MSA)** **88.1 (DA)**	**MADAMIRA (95.3)** **MADAMIRA (EGY) (92.4)**	SVM CFR	WANLP'17 LREC'18, JNLE'20
Spell Checking	78.4 (MSA)	**CMUQ (82.0)**	Noisy Channel	WANLP'14, 15
Lemmatization	**97.3 (MSA)**	MADAMIRA (96.6)	Rule-based	LREC'18
Diacritization	**95.5/94.0 (MSA)** **96.7 (CA)** **98.6 (DA-MA)** **97.5 (DA-TN)**	Microsoft (87.8) (MSA)	Seq2seq/RNN	ACL'19, EMNLP'19, NAACL'19, OSACT-LREC'18, WANLP-EACL'18
Dependency Parsing	**89.38 (MSA)**	Björkelund (88.3)	Randomized Greedy	NAACL'15
Constituency Parsing	79.7 (MSA)	**Berkeley (81.24)**	CRF	-
NER	**84.3 (MSA)** **65.2 (DA)**	Abdul-Hamid (81.0) Darwish (39.9)	CRF CRF	ACL'13 LREC'14
Sentiment Analysis	**80.0 (MSA)** **79.2 (DA)**	- -	Bayesian Classifier	WASSA'13

MSA: Modern Standard Arabic
DA: Dialectal Arabic
CA: Classical Arabic

شكل 1. وظائف فراسة، الخوارزميات، مؤتمرات نشر الأبحاث



أ) عرض أجزاء فراسة لمعالجة النص العربي على الإنترنت

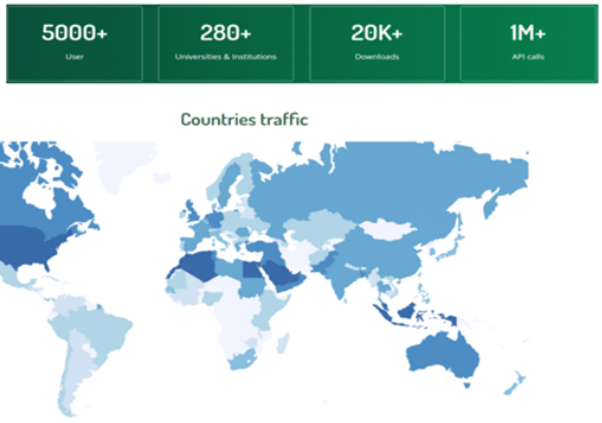


ب) عدد مرات تنزيل حزمة فراسة

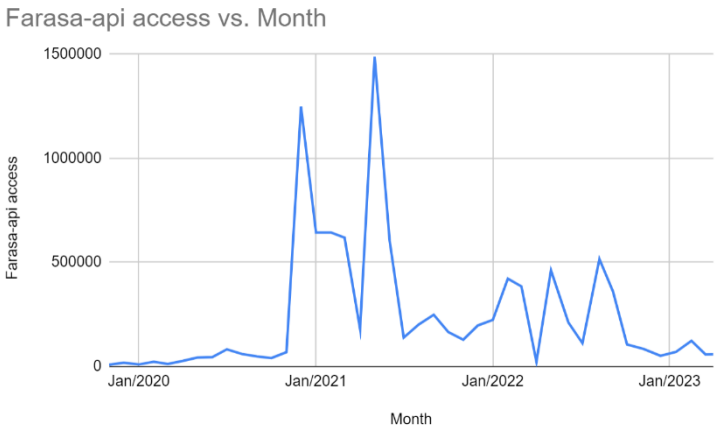
شكل 2. طريقة الوصول إلى منصة فراسة لمعالجة النص العربي عبر الإنترنت. تم تنزيل مجموعة أدوات فراسة أكثر من 20000 مرة.

استخدام واجهة تطبيق البرامج

تتيح واجهة تطبيق فراسة Farasa APIs تكامل وظائف معالجة اللغة العربية مع التطبيقات المختلفة بسهولة كبيرة. الأدوات متاحة للجمهور ويستخدمها الباحثون والمهتمون بهذه الوظائف اللغوية. منذ يناير 2019، تم استدعاء واجهات تطبيق البرامج مليون مرة بواسطة حوالي 300 باحث ومؤسسة حول العالم.



أ) التوزيع الجغرافي لاستخدام واجهة تطبيق البرامج من قبل الباحثين والمؤسسات



ب) عدد مرات استدعاء واجهة تطبيق فراسة

شكل 3. التوزيع الجغرافي لمستخدمي واجهة تطبيق برامج فراسة. تُظهر إحصائيات الاستخدام وجود فترات مكثفة (ارتفاعات) عندما ينخرط مجتمع البحث في تحديات، ويكون هناك احتياج من قبل المنظمات لوظائف معالجة اللغة العربية.

التأثير العلمي

- براءات الاختراع (هناك المزيد تحت الفحص حاليا):

التشكيل الآلي (2020) US 2022/0188515 A1, June 16, 2022

https://patentimages.storage.googleapis.com/15/65/76/71b164b3ee0077/US20220188515A1.pdf

التلخيص (2018) US Patent: 9,990,368 B2, Jun 5, 2018

https://patentimages.storage.googleapis.com/e4/b9/df/9edbe87c5fcb17/US9990368.pdf

- تم نشر 6 أبحاث في مؤتمرات المستوى الأول، و 12 بحثا في مؤتمرات أخرى وورش العمل العربية.
- من المؤتمرات البارزة التي تم نشر الأبحاث بها مؤتمرات: ACL, NAACL, EMNLP, ACM, COLING

- تنظيم ورش عمل اللغة العربية، ومنها على سبيل المثال:

WANLP 2022 (at EMNLP), OSACT 2020 and 2022 (LREC)

- إنشاء ومشاركة عينات اختبار قياسية مع مجتمع البحث العلمي، ومنها على سبيل المثال: مجموعة أخبار ويكي (WikiNews) لاختبار دقة تحديد أقسام الكلام، استخلاص أصول الكلمات، التشكيل الآلي، بالإضافة إلى عينات اختبار اللهجات العربية من تويتر.

- التعاون مع مراكز البحث المختلفة لتطوير الموارد اللغوية للغة العربية والتراث، ومنها المنظمة العالمية للنهوض باللغة العربية، معهد الدوحة التاريخي لإنشاء المعجم التاريخي للغة العربية، وغيرها.

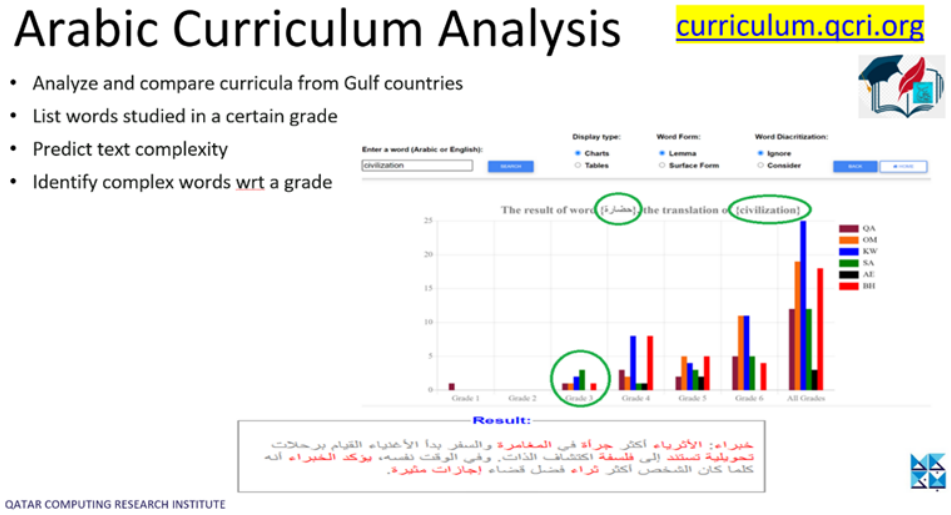
هذه قائمة مختصرة ببعض الأبحاث والمؤتمرات

1. Mubarak, Hamdy, Ahmed Abdelali, Hassan Sajjad, Younes Samih, and Kareem Darwish. "Highly effective Arabic diacritization using sequence to sequence modeling."
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* pp. 2390-2395. 2019. **NAACL-2019**
2. Mubarak, Hamdy, Ahmed Abdelali, Kareem Darwish, Mohamed Eldesouki, Younes Samih, and Hassan Sajjad. "A System for diacritizing four varieties of Arabic."
In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pp. 217-222. 2019. **EMNLP 2019**
3. Darwish, Kareem. "Named entity recognition using cross-lingual resources: Arabic as an example."
In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1558-1567. 2013. **ACL-2013**
4. Abdelali, Ahmed, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. "Farasa: A fast and furious segmenter for arabic." In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations*, pp. 11-16. 2016. **NAACL-2016**
5. Mubarak, Hamdy. "Build fast and accurate lemmatization for Arabic." *arXiv preprint arXiv:1710.06700* (2017). **LREC-2018**
6. Zhang, Yuan, Chengtao Li, Regina Barzilay, and Kareem Darwish. "Randomized greedy inference for joint segmentation, POS tagging and dependency parsing." In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 42-52. 2015. **NAACL-2015**
7. Darwish, K. and Mubarak, H., 2016, May. Farasa: A new fast and accurate Arabic word segmenter.
In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1070-1074). **LREC-2016**

يمكن الاطلاع على القائمة الكاملة للأبحاث على الموقع: https://farasa.qcri.org

الأثر المجتمعي

- دعم التعليم: تحليل المناهج الدراسية في دول الخليج العربي لمساعدة مطوري المناهج على تحسين مهارات الطلاب https://curriculum.qcri.org/



شكل 4. تحليل المناهج باستخدام أدوات فراسة

- تحسين جودة الترجمة الآلية لكسر حاجز اللغة بين المستخدمين.
- زيادة إمكانية الوصول، على سبيل المثال: من خلال قراءة المحتوى لكبار السن والمعاقين بصريًا.
- تعزيز مهارات اللغة العربية لدى الطلاب ومتعلمي اللغة، على سبيل المثال: تصحيح الأخطاء الإملائية، وتبسيط الكلمات المعقدة، إلخ.
- التعاون مع شبكة الجزيرة لتوفير برامج ومصادر لتعلم اللغة العربية (https://learning.aljazeera.net)

نقل التكنولوجيا

- التقطيع الصرفي لتحسين دقة الترجمة (منصة شاهين للترجمة الآلية) واسترجاع المعلومات
  الموقع: https://mt.qcri.org
  البحث: https://aclanthology.org/N16-3003.pdf

- تشكيل النص لاستعادة الحركات المفقودة الضرورية للنطق الصحيح لتطبيق ناطق لتحويل النص إلى كلام
  الموقع: https://tts.qcri.org
  البحث: https://arxiv.org/pdf/2206.07373.pdf

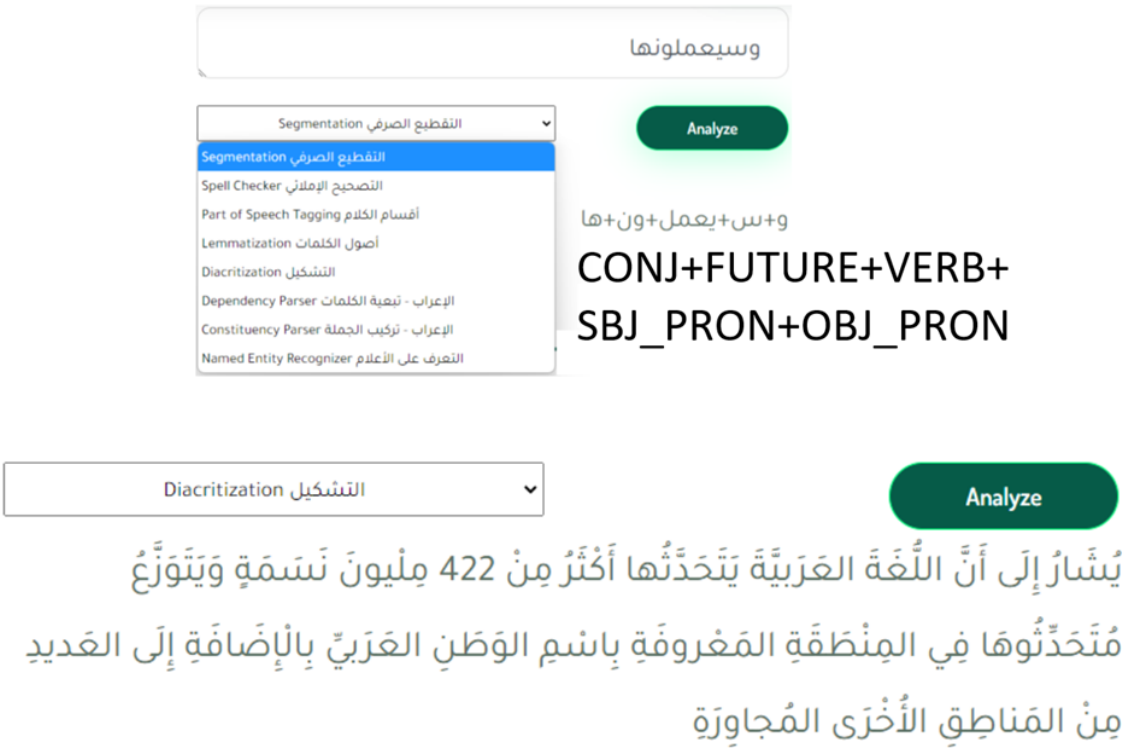
- ترخيص برامج الترجمة الآلية وتحويل النص إلى صوت إلى شركة كناري للذكاء الاصطناعي Kanari AI، الشركة الرائدة في معالجة الكلام للغة العربية الفصحى واللهجات. فازت شركة كناري للذكاء الاصطناعي بجائزة أفضل شركة ناشئة في مجال الذكاء الاصطناعي في معرض جيتكس GITEX 2021 (الإمارات العربية المتحدة)
  موقع الشركة: https://kanari.ai/
  https://www.wamda.com/2021/11/farmin-kanari-ai-win-gitex-supernova-challenge

- ترخيص برنامج التشكيل إلى شركة يابانية رائدة (شركة CJKI) لأغراض تعليمية https://www.cjk.org
- ترخيص برنامج استخلاص أصول الكلمات إلى الشركة البريطانية الرائدة في مجال البحث الصرفي SketchEngine https://www.sketchengine.eu/

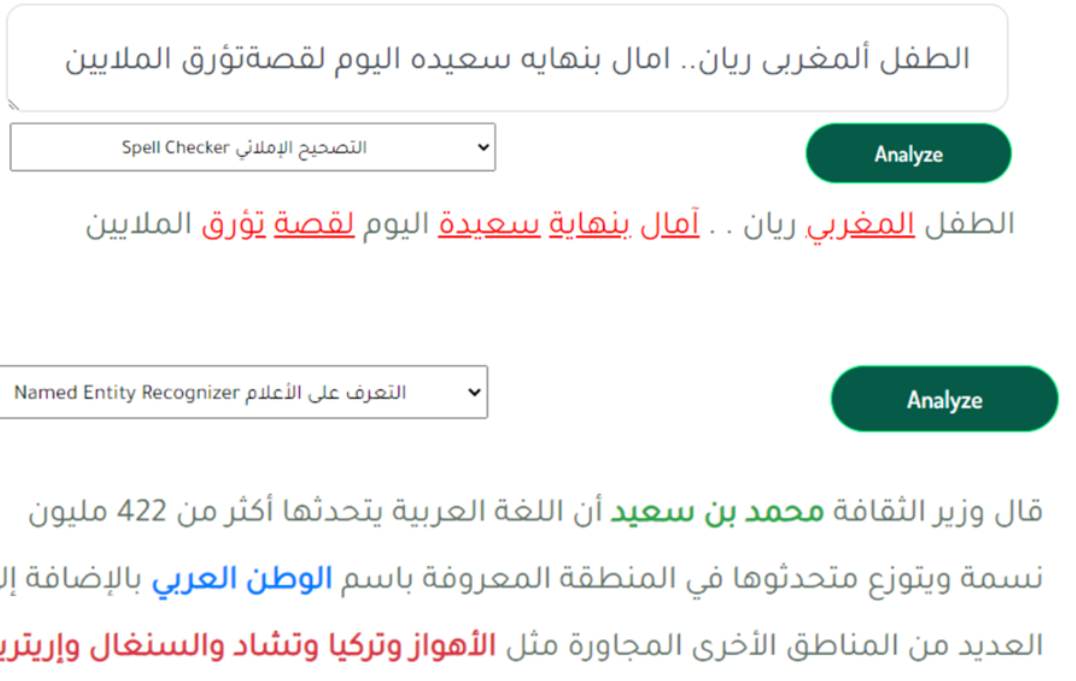
ملحق

واجهة تطبيق فراسة

يمكن تجربة وظائف فراسة المختلفة عن طريق الموقع: https://farasa.qcri.org



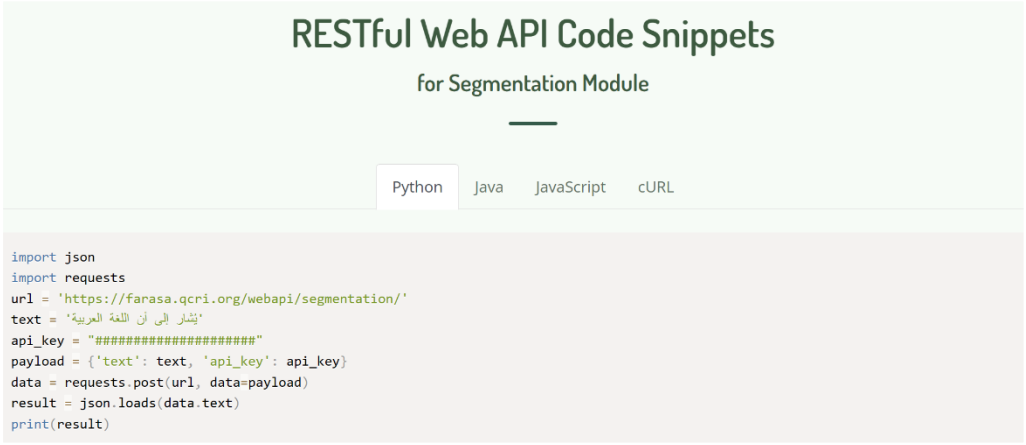
شكل 5. أمثلة لنتائج التقطيع الصرفي، تحديد أقسام الكلام، والتشكيل



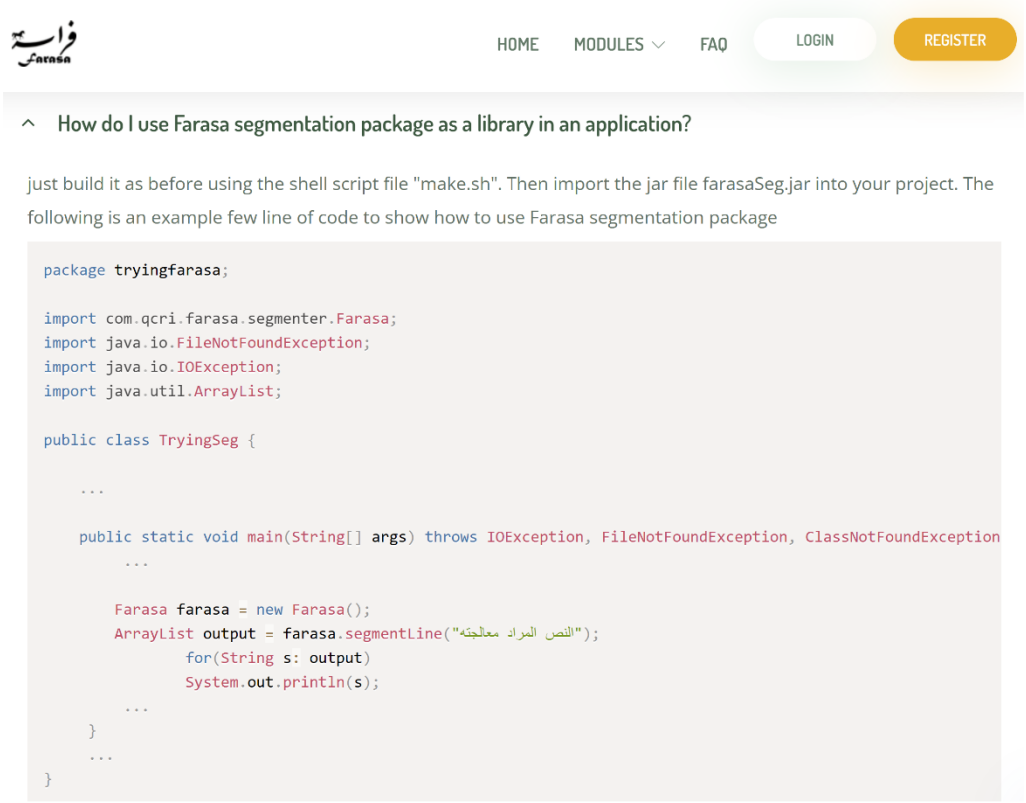
شكل 6. أمثلة لنتائج التصحيح الإملائي والتعرف على الأعلام

شكل 7. أمثلة لنتائج استخلاص أصول الكلمات وإعراب تبعية الكلمات

يمكن استدعاء جميع أدوات فراسة من خلال لغات برمجة مختلفة، مثل بايثون وجافا وغيرها



شكل 8. استدعاء الوظائف عن طريق لغات برمجة مختلفة



شكل 9. خطوات استخدام حزمة برامج فراسة داخل تطبيق مكتوب بلغة جافا