جامعة نيويورك أبوظبي

NYU | ABU DHABI

مختبر كامل

CAMeL Lab

WIDH@NYCDH 2021

Introduction to Arabic Text Processing with Python and CAMeL Tools

Starting soon!

جامعة نيويورك أبوظبي

NYU | ABU DHABI

مختبر كامل

CAMeL Lab

WIDH@NYCDH 2021

Introduction to Arabic Text Processing with Python and CAMeL Tools

Salam Khalifa and Ossama Obeid

Computational Approaches to Modeling Language (CAMeL) Lab

New York University Abu Dhabi

Roadmap

- Arabic NLP Challenges
- CAMeL Lab Solutions:

  CAMeL Tools
- Questions

خارطة الطريق

- تحديات اللغة العربية
- حلول من مختبر كامل:

  أدوات كامل
- أسئلة

The Arabic Language

- Classical Arabic
  - Quranic Arabic
  - Historical texts
- Modern Standard Arabic
  - Official language
  - Language of news & media
  - Standard writing & grammar
  - The National Language
- Dialectal Arabic
  - Predominantly spoken
  - No official standardization
  - The Mother Tongue
  - Increasing use on social media

اللغة العربية

- فصحى التراث
  - لغة القرآن الكريم
  - نصوص تاريخية
- فصحى العصر
  - اللغة الرسمية في البلدان العربية
  - لغة الصحف والإعلام
  - لها إملاء قياسي وقواعد قياسية
  - لغة الأمة
- اللهجات العربية
  - لغة التحدث الرئيسية
  - بدون إملاء أو قواعد قياسية رسمية
  - لغة الأم
  - وجود متزايد على وسائل التواصل الاجتماعية

The Main Challenges for Arabic Language Processing

التحديات الرئيسية للمعالجة الآلية للغة العربية

- Orthographic ambiguity
- Morphological richness
- Dialectal variation
- Orthographic inconsistency
- Resource poverty (data & tools)
- Limited research

- الإبهام الإملائي
- الغنى الصرفي
- تعدد اللهجات
- الأخطاء الإملائية
- فقر موارد البيانات والأدوات
- البحث العلمي المحدود

Arabic Script

الخط العربي

العربية

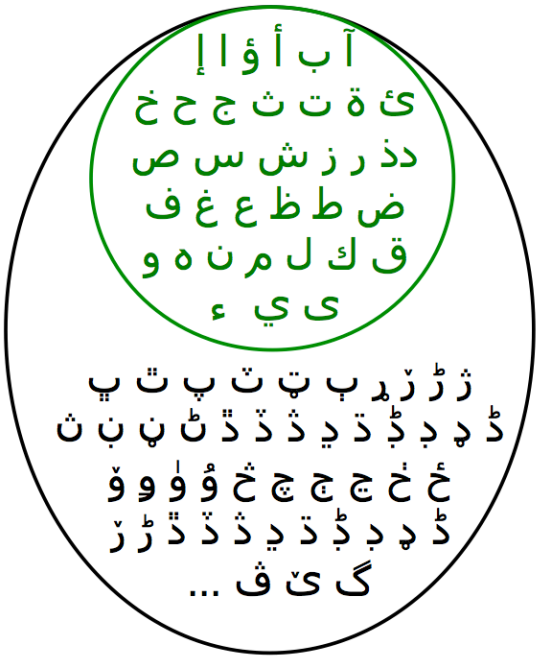
ال ع ر ب ي ة

- Written right-to-left
- Letters have contextual variants
- Used to write many languages besides Arabic: Persian, Kurdish, Urdu, Pashto, etc.

- مكتوب من اليمين إلى اليسار
- حروف أشكالها متغيرة سياقيا
- يستخدم لكتابة العديد من اللغات إلى جانب العربية: الفارسية والكردية والأردية والباشتو ، إلخ.

Arabic Script

الخط العربي



العربية

ال ع ر ب ي ة

- Written right-to-left
- Letters have contextual variants
- Used to write many languages besides Arabic: Persian, Kurdish, Urdu, Pashto, etc.

- مكتوب من اليمين إلى اليسار
- حروف أشكالها متغيرة سياقيا
- يستخدم لكتابة العديد من اللغات إلى جانب العربية: الفارسية والكردية والأردية والباشتو ، إلخ.

Orthographic Ambiguity

الإبهام الإملائي

- Arabic script uses optional diacritical marks
  - 1.5% of newspaper words have some diacritical marks
  - Standard Arabic has 6.8 diacritizations and 2.7 lemmas per word

- الإملاء بالخط العربي يستخدم التشكيل الاختياري
  - ١،٥ ٪ من كلمات نصوص الصحف تحتوي على علامات تشكيل
  - الكلمات العربية الفصحى ٦،٨ تشكيلات و٢،٧ مداخل معجمية بالمتوسط

وبعقدنا

وَبِعُقَدِنَا وَبِعِقْدِنَا وَبِعَقْدِنَا وْبِعَقِّدْنَا

and with our (contract, necklace, psychoses) | and he stresses us out

Morphological Richness

الغنى الصرفي

- Arabic has a very rich inflectional system
  - For example, Arabic verbs have 5,400 inflected forms
  - Whereas English verbs have 6 and Chinese verbs have 1!

- للغة العربية نظام صرفي غني جدًا ينتج عنه تراكيب كثيرة
  - فمثلاً، للفعل العربي حوالي ٥،٤٠٠ تصريف
  - بينما للفعل الإنجليزي ٦ تصاريف، وللفعل الصيني تصريف واحد!

قال، قالت، قالا، قالوا، قلتَ، قلتِ، قلتما، قلتم، قلتن، يقولُ، يقولَ، يقل، تقولُ، تقولَ، تقل، تقولين، تقولي، ...فقال، فقالت، فقالا... وسأقولها، **وسنقولها** ... ،

وسنقولها
/wasanaqūluhā/
و+ س+ ن+ قول + ها
wa+sa+na+qūl+u+hā
and+will+we+say+it
And we will say it

Dialectal Variation

تعدد اللهجات

- Coarse Classification
  - Gulf Arabic
  - Levantine Arabic
  - Egyptian Arabic
  - Maghrebi Arabic
- Classification Problems
  - Countries with multiple varieties
  - Sub-dialectal varieties:
    e.g., Urban, Rural, and Bedouin
- Diglossia and Multilinguality

- التصنيف العام
  - اللهجة الخليجية
  - اللهجة الشامية
  - اللهجة المصرية
  - اللهجة المغاربية
- مشاكل تصنيفية
  - تعدد اللهجات في نفس البلد
  - تعدد اللهجات الفرعية:
    المدنية، والفلاحية والبدوية، مثلا
- الازدواجية والتعددية اللغوية

Phonological Variation

الاختلافات الصوتية

- First noticed differences
  - Not the most important

- الاختلافات الصوتية تُلاحظ فورياً
  - ليست الأكثر أهمية

MSA الفصحى		Dialects اللهجات	
ق	/q/	ق، ك ، ء، گ، ج	/q/, /k/, /ʔ/, /g/, /ʤ/
ث	/θ/	ث، ت، س	/θ/, /t/, /s/
ذ	/ð/	ذ، د، ز	/ð/, /d/, /z/
ج	/ʤ/	ج، گ، ج	/ʤ/, /g/, /ʒ/

قلب → /qalb/, /galb/, /ʔalb/ *(heart)*

/fagr/ → فجر ، فقر *(poverty, dawn)*

Morphological Variation

الاختلافات الصرفية

- Compared to Standard Arabic, dialectal morphology can be simpler or more complex
  - No case
  - New complex clitics

- مقارنة بصرف الفصحى، فإن صرف اللهجات أحياناً أبسط وأحياناً أكثر تعقيداً
  - اختفاء الحركات الإعرابية
  - تعدد التراكيب اللواصقية

كتابْ → كتابُ، كتابٌ، كتابَ، كتاباً، كتابِ، كتابٍ

(book)

وماكتبتوهالوش → ولم تكتبوها له

(you all did not write it for him)

Morphological Variation

الاختلافات الصرفية

- Dialects vary systematically among themselves

- هناك العديد من الاختلافات المنهجية الصرفية بين اللهجات

	Past الماضي	Future المستقبل
فصحى Standard	كتب /kataba/	سيكتب /sajaktubu/
شامي Levantine	كتب /katab/	حيكتب /ħajiktob/
مصري Egyptian	كتب /katab/	هيكتب /hajiktib/
مغربي Moroccan	كتب /kteb/	غيكتب /ʁajekteb/

Lexical Variation

الاختلافات المعجمية

طوماطيش

طماطيس

مطيشة

قوطة

طماطم

بندورة

طماطة

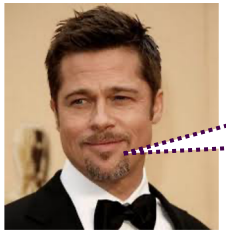
طماط

*You say **to-MAY-to**,*
*I say **to-MAH-to**!*

براد؟

/b a r r aa d/

/b a r r aa d/



أنا أشهر براد بيت!
I'm the most famous "*brad*"

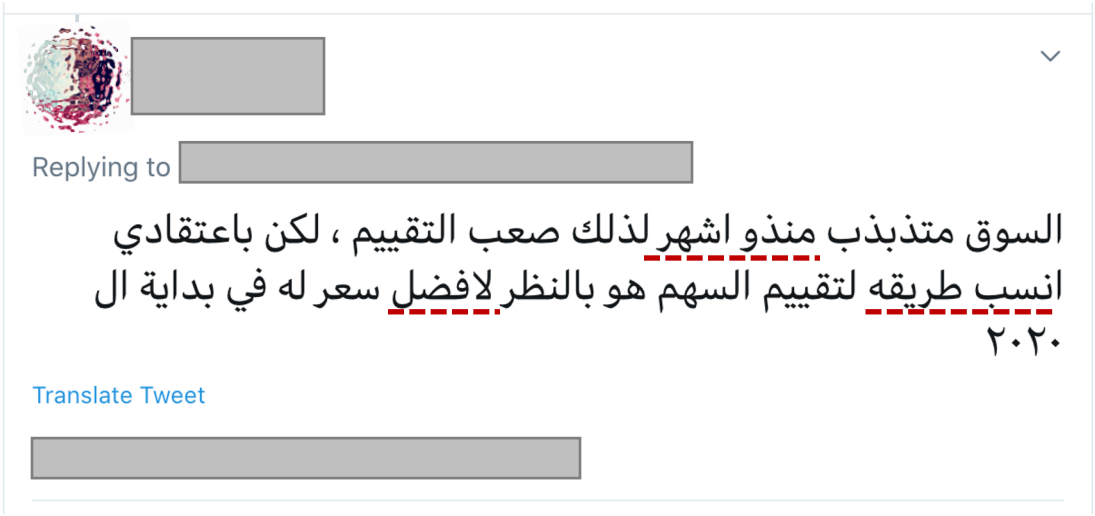
/b r aa d/

Common Spelling Errors and Inconsistency

- Unedited standard Arabic text on news commentaries has an error rate of ~30% (QALB project)

كثرة الأخطاء الإملائية والتضارب الإملائي

- نسبة الخطأ في التعليقات الغير محررة المكتوبة بالعربية الفصحى على صفحات بعض مواقع الأخبار تصل إلى ٣٠٪ (مشروع «قلب»)



Common Spelling Errors and Inconsistency

كثرة الأخطاء الإملائية والتضارب الإملائي

- But in the dialects, there is no official orthography! So, there are many free forms including Arabizi.

- ولكن في اللهجات، لا يوجد معيار إملائي قياسي! لذا ، هناك تنوع في التهجئات بما في ذلك الإملاء المرومن (العربيزي).

وجدنا على الإنترنت ٢٧ تهجئة لهذه الكلمة باللهجة المصرية

We found 27 spellings online of the Egyptian Arabic word

ما بيقولهاش

/mabiʔulhāʃ/

“he does not say it”

(Habash et al., 2018)

Arabic Orthography	Arabic Transliteration	Frequency
مبيقولهاش	*mbyqwlhAš*	≈ 26,000
ما بيقولهاش	*mA byqwlhAš*	≈ 13,000
مابقلهاش، مبقولهاش، مبقلهاش، ما بقلهاش، مابيقولهاش	*mAbqlhAš, mbqwlhAš, mbqlhAš, mA bqlhAš, mAbyqwlhAš*	≤ 10,000
مابقولهاش، ما بقولهاش، مبيقلهاش، ما بيقلهاش مابيقلهاش	*mAbqwlhAš, mA bqwlhAš, mbyqlhAš, mA byqlhAš mAbyqlhAš*	≤ 1,000
مبئلهاش، مابيئولهاش، ما بيئولهاش، مابيؤلهاش	*mbŷlhAš, mAbyŷwlhAš, mA byŷwlhAš, mAbyŵlhAš*	≤ 100
ما بيؤلهاش، مابئلهاش، مبيئولهاش، ما بيئلهاش، مابئولهاش، ما بئلهاش، ما بؤلهاش، مبئولهاش، مبيؤلهاش، مابؤلهاش، مبؤلهاش	*mA byŵlhAš, mAbŷlhAš, mbyŷwlhAš, mA byŷlhAš, mAbŷwlhAš, mA bŷlhAš, mA bŵlhAš, mbŷwlhAš, mbyŵlhAš, mAbŵlhAš, mbŵlhAš*	≤ 10

Arabizi variants: mabi**2**ulhash, mabi**q**ulhash, mabi’ulhash,… :بضعة تهجئات **عربيزية**

The Main Challenges for Arabic Language Processing

التحديات الرئيسية للمعالجة الآلية للغة العربية

- Orthographic ambiguity
- Morphological richness
- Dialectal variation
- Orthographic inconsistency
- **Resource poverty (data & tools)**
- Limited research

- الإبهام الإملائي
- الغنى الصرفي
- تعدد اللهجات
- الأخطاء الإملائية
- **فقر موارد البيانات والأدوات**
- البحث العلمي المحدود

Roadmap

- Arabic NLP Challenges
- **CAMeL Lab Solutions:**

  CAMeL Tools
- Questions

خارطة الطريق

- تحديات اللغة العربية
- **حلول من مختبر كامل:**

  أدوات كامل
- أسئلة

Computational Approaches to Modeling Language Lab

- Researchers
- Research
- Teaching
- Publications
- Resources
- News
- Work With Us

Computational Approaches to Modeling Language Lab

The Computational Approaches to Modeling Language (CAMeL) Lab is a research lab at New York University Abu Dhabi established in September 2014. CAMeL's mission is research and education in artificial intelligence, specifically focusing on natural language processing, computational linguistics, and data science. The main lab research areas are Arabic natural language processing, machine translation, text analytics and dialogue systems.

مختبر كامل
CAMeL Lab

www.camel-lab.com

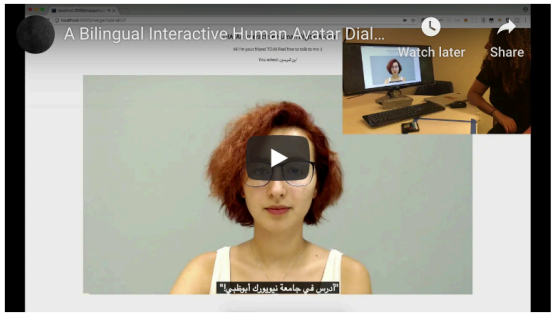
موارد كامل
CAMeL Resources

أدوات كامل
CAMeL Tools

Featured Projects

TOIA

TOIA (time-offset interaction application) is a bilingual (Arabic-English) conversational agent, similar to a chat bot, except that the avatar is based on pre-recorded videos of an actual human being.



مدار
The MADAR Project

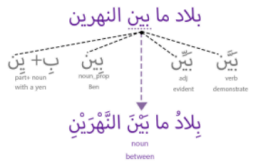
MADAR

Parallel Dialectal Corpus and Lexicon for 25+ cities

مكنز قُمر
Gumar Corpus

Gumar

A browsable 100-million word data set of Gulf Arabic



MADAMIRA

Arabic Artificial Intelligence

Corpora

[+] Expand all

- ANERcorp - CAMeL Lab Train/Test Splits
- Arab-Acquis: A parallel corpus of Arabic with all the languages of the European Parliament
- The Arabic Parallel Gender Corpus
- Gumar: The Gulf Arabic Corpus
- Margarita Dialogue Corpus
- Multi-Arabic Dialect Corpus (MADAR Corpus)
- NYUAD Universal Dependency of Arabic (NUDAR) Treebank
- Palm Treebank: A Multi-genre Arabic Dependency Treebank
- Qatar Arabic Language Bank (QALB Corpus)

Lexicons

[+] Expand all

- Arabic Multidialectal Word Embeddings
- ArabScribe
- Multi-Arabic Dialect Lexicon (MADAR Lexicon)
- SAMER Readability Lexicon

Tools

[+] Expand all

- ADIDA: Arabic Dialect Identification
- BOTTA: An Arabic Dialect Chatbot
- CALIMA-STAR
- CAMeLParser
- CAMeL Tools
- DALILA: The Dialectal Arabic Linguistic Learning Assistant
- Gulf Arabic Morphological Analyzer
- MADAMIRA: Morphological Analyzer and Disambiguator for Arabic
- MADARi: Web-base Morphological Annotation Tool
- Palmyra
- SIMMR
- YAMAMA: Arabic Morphological Analyzer and Disambiguator

Guidelines

[+] Expand all

- CAMeL Arabic Phonetic Inventory (CAPHI)
- CAMeL Part-of-speech tagset and guidelines (CAMEL POS)
- CODA: Conventional Orthography for Dialectal Arabic

أدوات كامل
CAMeL Tools

The Camel Tools Team

فريق أدوات كامل

أسامة عبيد
Ossama Obeid

ناصر زلموط
Nasser Zalmout

سلام خليفة
Salam Khalifa

ديما التاجي
Dima Taji

مي عودة
Mai Oudah

بشار الحفني
Bashar Alhafni

جو إينوي
Go Inoue

فضل الإرياني
Fadhl Eryani

اليكساندر إردمان
Alex Erdmann

نزار حبش
Nizar Habash

- **An Open-source Python Toolkit for Arabic Natural Language Processing**
- Arabic Specific
- Flexibility in use and reuse
- Modularity
- High Performance
- Ease of use for beginner

- **مجموعة أدوات مفتوحة المصدر بلغة بايثون لمعالجة اللغة العربية**
- التركيز على اللغة العربية
- مرونة الاستخدام وإعادة الاستخدام
- وحدات مستقلة سهلة التجميع
- الأداء عالي الجودة
- سهولة الاستخدام للمبتدئين

أدوات كامل
CAMeL Tools

Comparison with other toolkits المقارنة مع رزم أخرى

	CAMeL Tools	MADAMIRA	Stanford CoreNLP	Farasa	
Target Language	Arabic خاص بالعربية	Arabic خاص بالعربية	Multilingual متعدد اللغات	Arabic خاص بالعربية	نوع الحزمة
Programming Language	Python بايثون	Java جافا	Java/Python جافا/بايثون	Java جافا	لغة البرمجة
CLI	✓	✓	✓	✓	واجهة سطر الأوامر
API	✓	✓	✓	✓	واجهة برمجة التطبيقات
Exposed Pre-processing	✓				المعالجة المسبقة للنصوص
Morphological Modeling	✓	✓			نمذجة التصريف
Disambiguation	✓	✓	✓	✓	حل الالتباس
POS Tagging	✓	✓	✓	✓	توسيم قسم الكلام
Diacritization	✓	✓		✓	التشكيل
Tokenization	✓	✓	✓	✓	التقطيع
Lemmatization	✓	✓		✓	المدخل المعجمي
Named Entity Recognition	✓	✓	✓	✓	التعرف على الكيانات المسماة
Sentiment Analysis	✓				تحليل المشاعر
Dialect ID	✓				تحديد اللهجة

- Other new toolkits include Adwat (Zerrouki, 2020) and ASAD (Mubarak et al., 2020)
- من بين الرزم الأخرى الحديثة، رزمة «أدوات» (زروقي، ٢٠٢٠) ورزمة «أسد» (مبارك وآخرون، ٢٠٢٠)

Roadmap

- Arabic NLP Challenges
- CAMeL Lab Solutions:

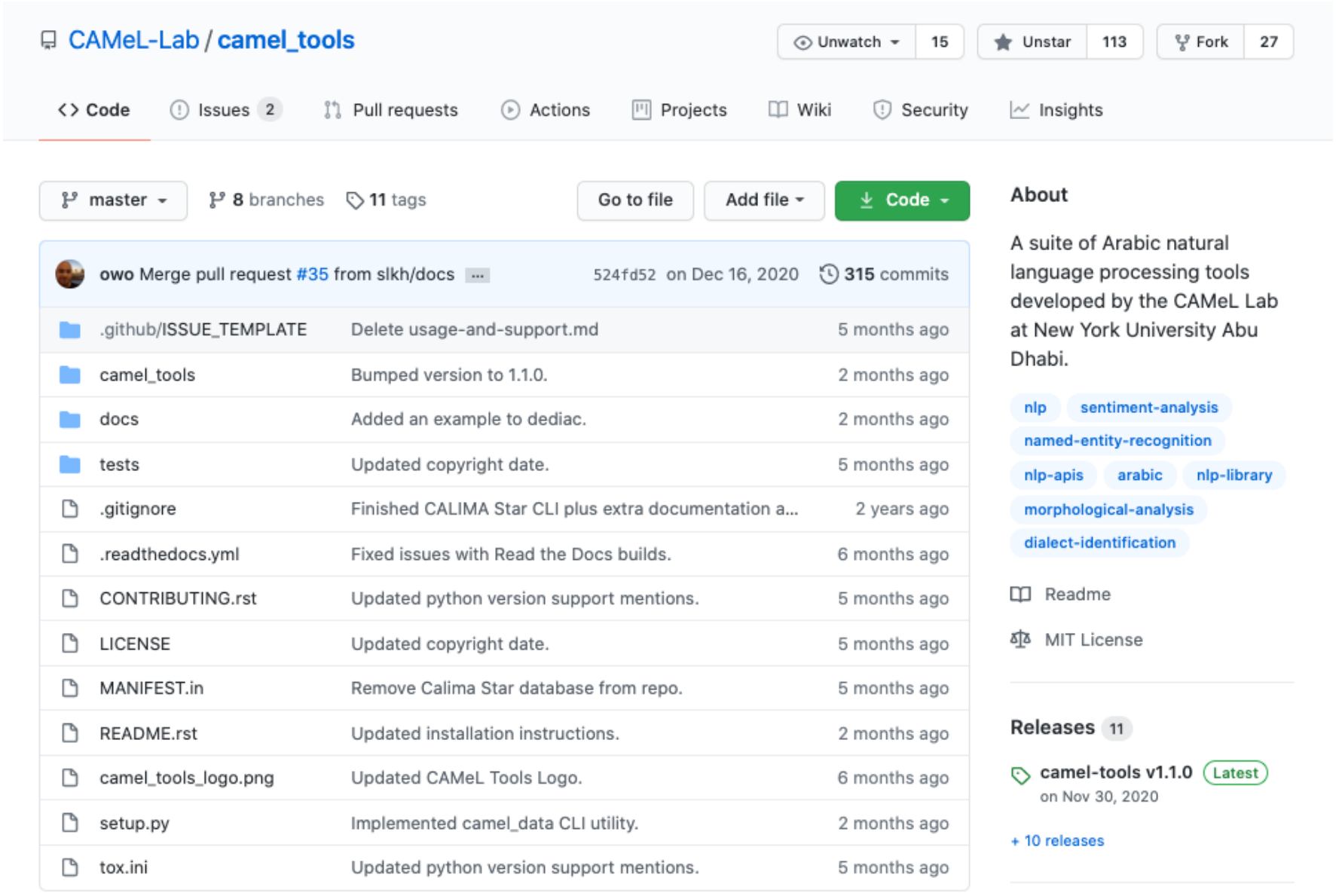
  CAMeL Tools
- **Questions**

خارطة الطريق

- تحديات اللغة العربية
- حلول من مختبر كامل:

  أدوات كامل
- **أسئلة**

https://github.com/CAMeL-Lab/camel_tools



CAMeL-Lab / camel_tools

Unwatch 15 | Unstar 113 | Fork 27

Code | Issues 2 | Pull requests | Actions | Projects | Wiki | Security | Insights

master | 8 branches | 11 tags | Go to file | Add file | Code

owo Merge pull request #35 from slkh/docs ... 524fd52 on Dec 16, 2020 315 commits

.github/ISSUE_TEMPLATE	Delete usage-and-support.md	5 months ago
camel_tools	Bumped version to 1.1.0.	2 months ago
docs	Added an example to dediac.	2 months ago
tests	Updated copyright date.	5 months ago
.gitignore	Finished CALIMA Star CLI plus extra documentation a...	2 years ago
.readthedocs.yml	Fixed issues with Read the Docs builds.	6 months ago
CONTRIBUTING.rst	Updated python version support mentions.	5 months ago
LICENSE	Updated copyright date.	5 months ago
MANIFEST.in	Remove Calima Star database from repo.	5 months ago
README.rst	Updated installation instructions.	2 months ago
camel_tools_logo.png	Updated CAMeL Tools Logo.	6 months ago
setup.py	Implemented camel_data CLI utility.	2 months ago
tox.ini	Updated python version support mentions.	5 months ago

About

A suite of Arabic natural language processing tools developed by the CAMeL Lab at New York University Abu Dhabi.

nlp | sentiment-analysis | named-entity-recognition | nlp-apis | arabic | nlp-library | morphological-analysis | dialect-identification

Readme

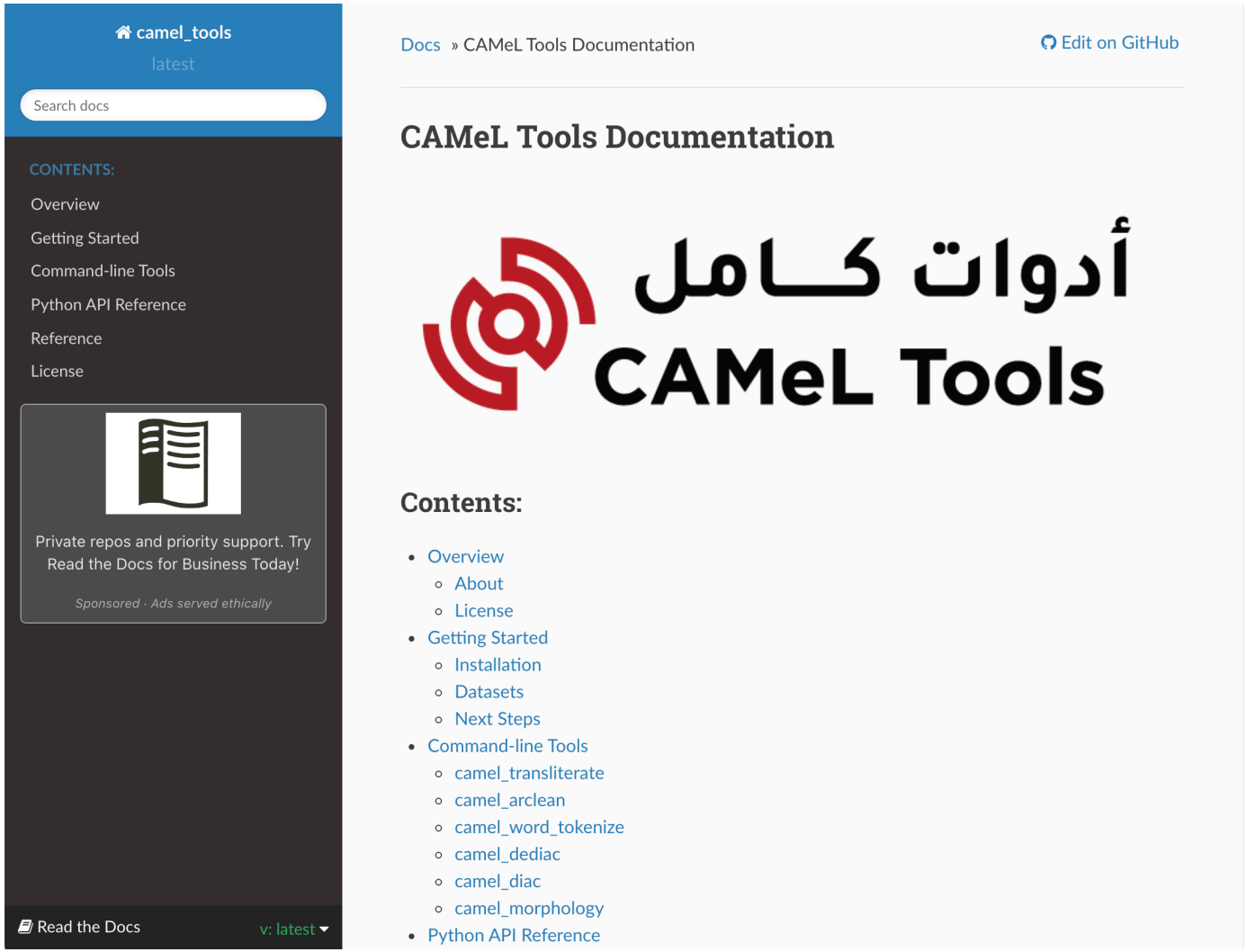
MIT License

Releases 11

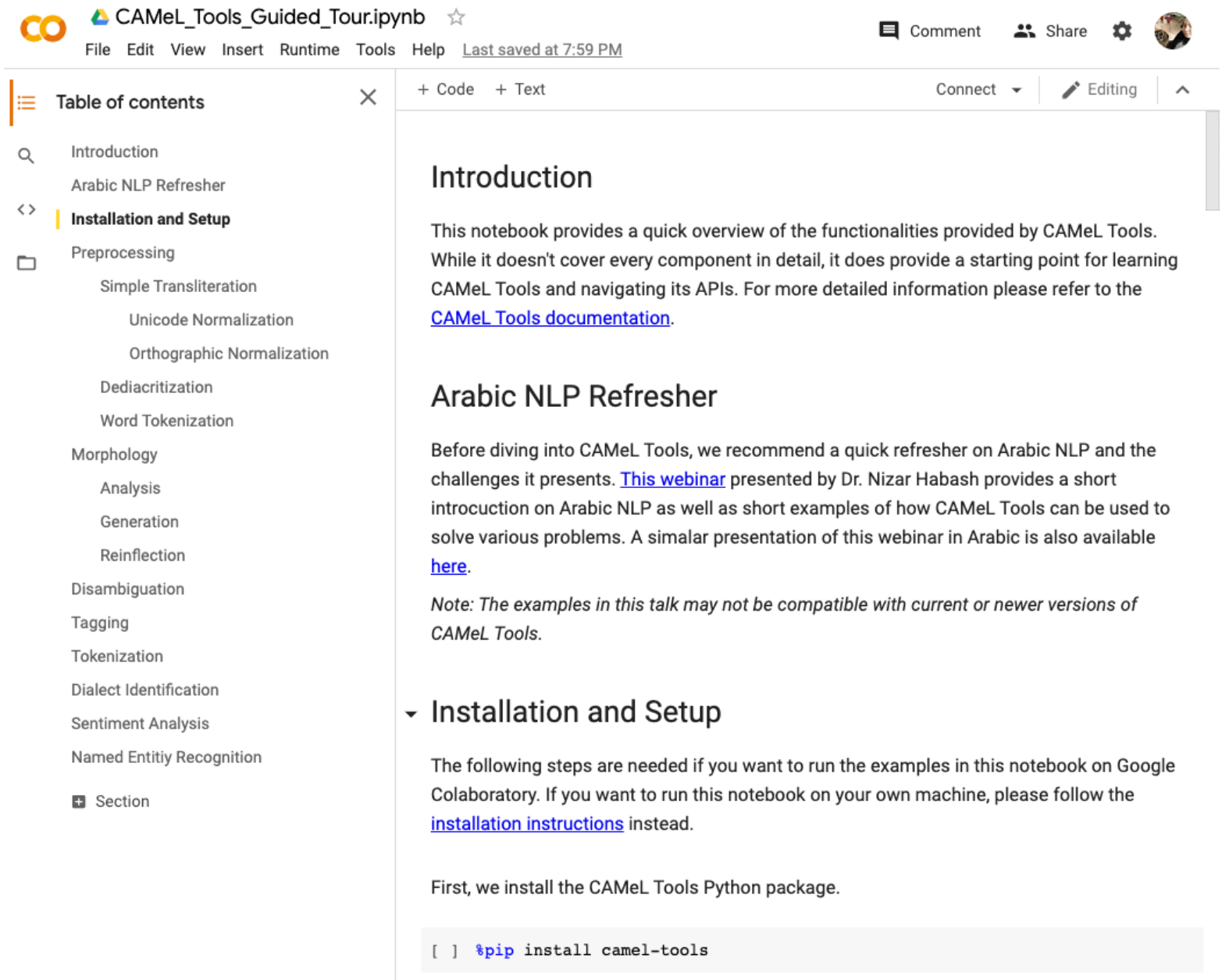
camel-tools v1.1.0 Latest
on Nov 30, 2020

+ 10 releases

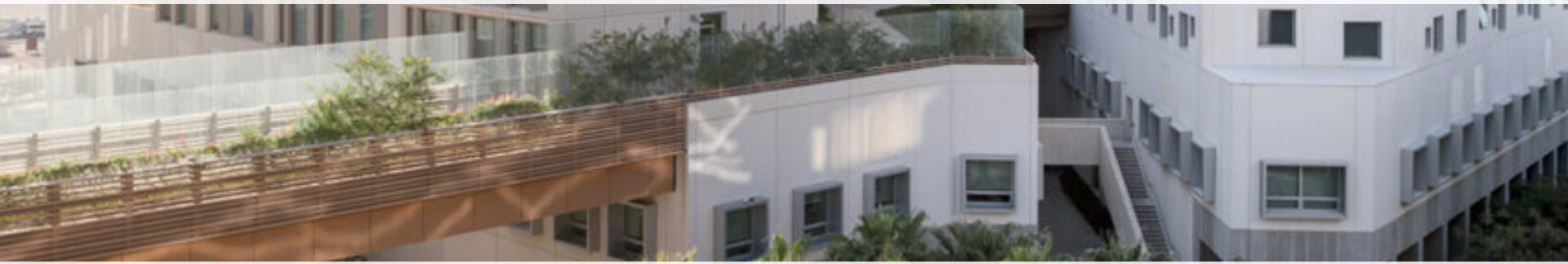
https://camel-tools.readthedocs.io/



https://bit.ly/3a40q3O



Thank You !شكراً على المتابعة



www.camel-lab.com
{oobeid,salamkhalifa}@nyu.edu

diac	وَسَيَكْتُبُونَها	التشكيل
lex	كَتَب-u_1	المدخل المعجمي
root	ك.ت.ب	الجذر
pattern	وَسَيُ123ونَها	الوزن
caphi	w_a_s_a_y_a_k_t_u_b_uu_n_a_h_aa	اللفظ
gloss	and_+_will_+_they_(people)+write+it;them;her	الترجمة
pos	verb	قسم الكلام
catib6	PRT+PRT+VRB+NOM	قسم الكلام/كاتب
ud	CONJ+AUX+VERB+PRON	قسم الكلام/يو دي
prc2	wa_conj	سابقة٢
prc1	sa_fut	سابقة١
per	3	الشخص
gen	m	الجنس
num	p	العدد
asp	i	الزمن
vox	a	البناء
mod	i	الصيغة
enc0	3fs_dobj	لاحقة١
d1tok	وَ+_سَيَكْتُبُونَها	تقطيع د١
d2tok	وَ+_سَ+_يَكْتُبُونَها	تقطيع د٢
atbtok	وَ+_سَ+_يَكْتُبُونَ_+ها	تقطيع بنك بن الشجري
d3tok	وَ+_سَ+_يَكْتُبُونَ_+ها	تقطيع د٣
bwtok	وَ+_سَ+_يَ+_كْتُبُ_+ونَ_+ها	تقطيع باكو التر
bw	IVSUFF_DO:3FS/هاIVSUFF_SUBJ:MP_MOOD:I+/ونIV+/كْتُبُIV3MP+/يَFUT_PART+/سَCONJ+/وَ	تحليل باكو التر
pos_logprob	-1.023208	احتمالية قسم الكلام
lex_logprob	-3.648503	احتمالية المدخل المعجمي