

AMLD Africa Workshop, September 4th, 2021

Moroccan Darija Wikipedia: Basics of Natural Language Processing for a Low-Resource Language

Khalil Mrini, Imane Khaouja, Ihsane gryech, Anass Sedrati and Abdelhak Mahmoudi



Empower AI transformation in Morocco and connect today's and tomorrow's AI leaders.

 Join Us

- Webinars
- Podcasts
- Journal Club
- Study Groups
- ...



Agenda



<https://github.com/MoroccoAI/AMLD>

- 3:00-3:10, Introduction
- 3:10-3:40, Anass Sedrati: **Darija Wikipedia**
- 3:40-3:50, **Explore the Moroccan Darija Wikipedia**
- 3:50-4:10, Ihsane Gryech: **Intro to NLP**
- 4:10-4:40, Imane Khaouja: **Lab1: Wikipedia Darija Cleaning**
- 4:40-5:00, **20 min Break**
- 5:00-5:30, Khalil Mrini: **Lab2: Wikipedia Darija Topic Detection**
- 5:30-6:00, Abdelhak Mahmoudi: **Lab3: NLP Tasks and tools**
- 6:00-6:20, **20 min Break**
- 6:20-8:20, **Explore on your own in breakout rooms**
- 8:20-8:50, Discussion of your results
- 8:50-9:00, Conclusion

Darija Wikipedia

Darija Wikipedia - Background



Anass Sedrati



Agenda

- Wikimedia Overview
- Wikipedia Facts
- Wikimedia in Morocco
 - Wikimedia Morocco User Group
 - Wikipedias in Morocco
- Moroccan Darija - Linguistic overview
- Wikipedia in Darija
- Challenges and way forward

Wikimedia - Overview

- International movement centered around the free digital collaborative encyclopedia: Wikipedia
- “Nobody knows everything, but everyone knows something”.
- Chronology
 - 15 January 2001: Creation of English Wikipedia by Jimmy Wales
 - March 2001: First foreign languages - German, French & Catalan
 - 2003: Arabic Wikipedia
 - 2003: Creation of the Wikimedia Foundation
 - 2008: Darija and Tamazight Wikipedia in the incubator
 - 2015: Creation of Wikimedia Morocco User group
 - July 2020: Official creation of Darija Wikipedia (ary.wikipedia.org)
 - July 2021: Official creation of Tachelhit Wikipedia (shi.wikipedia.org)

Wikimedia - Overview - Numbers

- 15 Wikimedia projects
 - Wikipedia - Wiktionary - Wikibooks - Wikiquote - Wikivoyage - Wikisource - Commons - Wikidata - Wikispecies - Wikinews - etc.
- 450+ staff/contractors at the WMF
- 312 Wikipedia language versions
- 13: Wikipedia's ranking as most visited websites in the world
- 27: Ranking in most visited sites in Morocco
- 55+ million: Total pages in all Wikipedias
- 6.4 million: Number of Wikipedia pages in English
- 1.1 million: Number of Wikipedia pages in Arabic

Wikimedia Projects



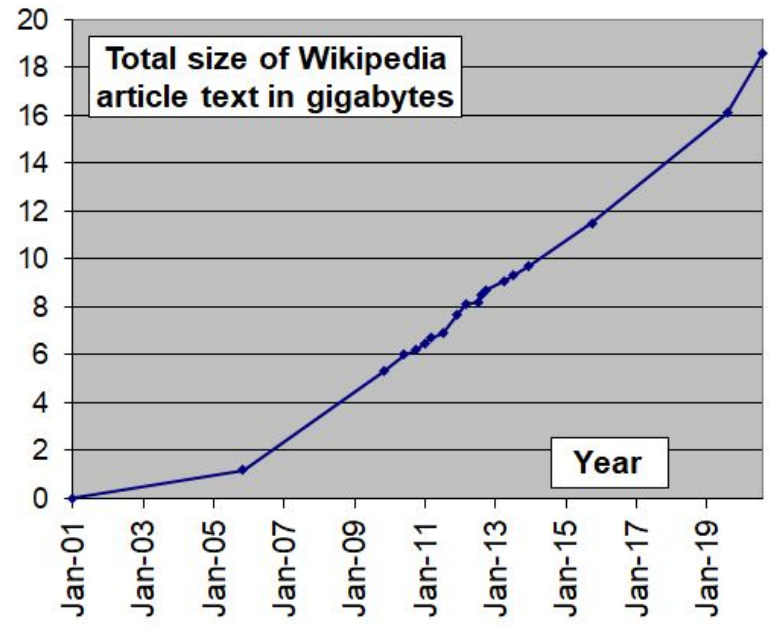
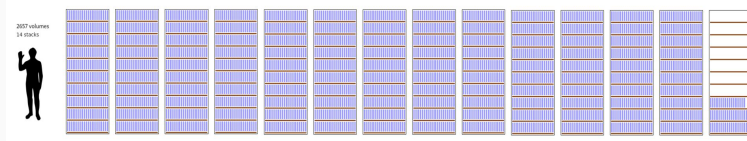
Wikimedia - Overview - Numbers

No ↕	Language ↕	Language (local) ↕	Wiki ↕	Articles ↕
1	English	English↗	en	6,363,265↗
2	Cebuano	Sinugboanong Binisaya↗	ceb	5,897,305↗
3	Swedish	Svenska↗	sv	2,951,702↗
4	German	Deutsch↗	de	2,608,259↗
5	French	Français↗	fr	2,355,160↗
6	Dutch	Nederlands↗	nl	2,064,518↗
7	Russian	Русский↗	ru	1,748,569↗
8	Italian	Italiano↗	it	1,712,973↗
9	Spanish	Español↗	es	1,709,472↗
10	Polish	Polski↗	pl	1,486,974↗
11	Egyptian Arabic	مصرى (Masri)↗	arz	1,322,872↗
12	Japanese	日本語↗	ja	1,286,060↗
13	Vietnamese	Tiếng Việt↗	vi	1,268,439↗
14	Waray-Waray	Winaray↗	war	1,265,477↗
15	Chinese	中文↗	zh	1,222,879↗
16	Arabic	العربية↗	ar	1,132,367↗
17	Ukrainian	Українська↗	uk	1,110,637↗
18	Portuguese	Português↗	pt	1,073,828↗

Wikimedia - Overview - Numbers

If Wikipedia was a book

- 19.52 GB
- 2958 volumes
- 15 stacks



Wikipedia Facts - True

- Wikipedia is an encyclopedia.
- Wikipedia is free.
- Wikipedia is neutral.
- Wikipedia is open source.
- Anyone can write in Wikipedia.
- 99% of Wikipedians are volunteers.
- Wikipedia has rules.

Wikipedia Facts - False

- Wikipedia is limited (in size, article length, etc.)
- Wikipedia is a dictionary
- You can write about everything in Wikipedia
- Wikipedia is a discussion forum
- You can share your opinion and views on Wikipedia
- Wikipedia is a marketing and advertisement place
- People are paid to write in Wikipedia

Wikimedia Morocco User Group

- Created in October 2015
- Supporting and promoting Wikimedia projects in Morocco
- Organizing events and projects related to this purpose



Wikimedia Morocco User Group - Activities

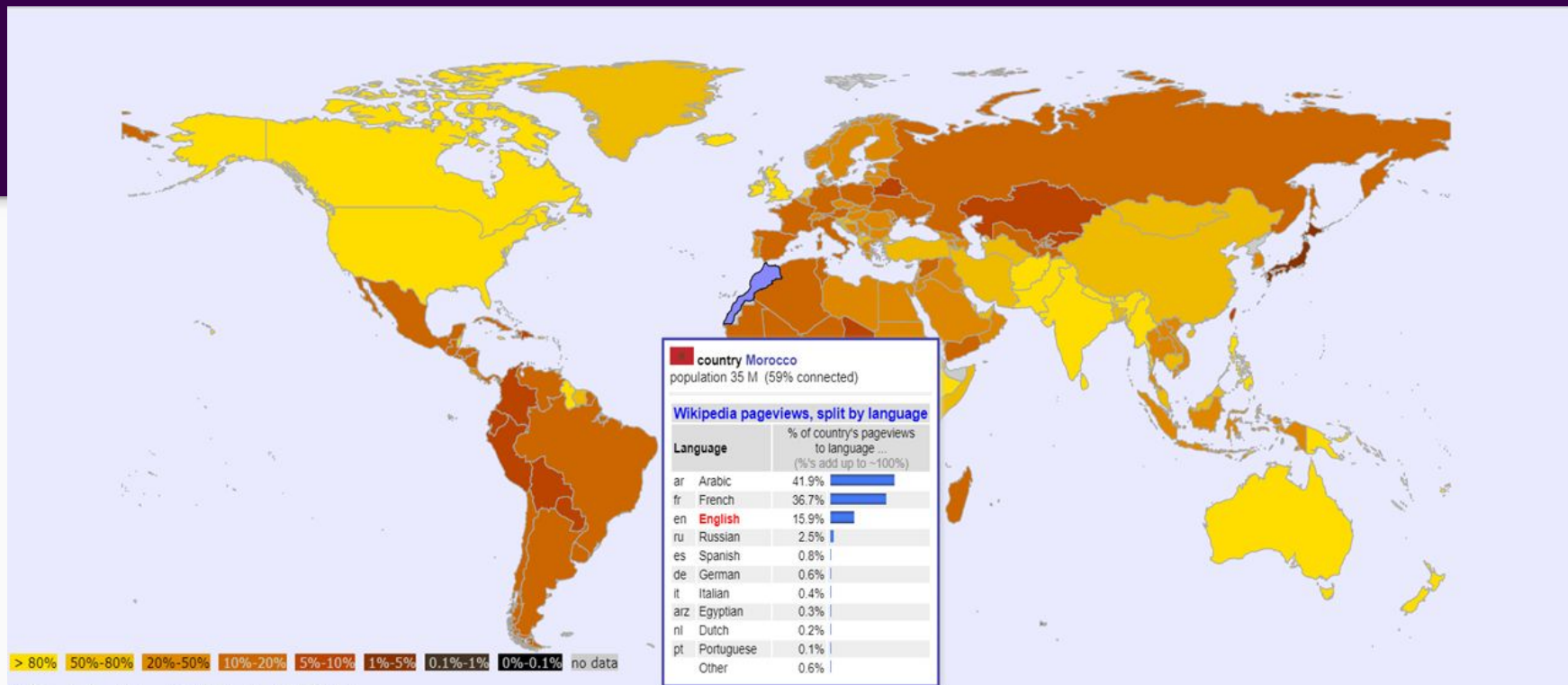
Wikimedia MA User Group holds a number of activities around the country where experienced editors and newcomers can contribute to the Wikimedia projects together. It works also with institutions to improve the quality of content on Wikipedia. Among activities the groups is organizing:

- **Edit-a-thons** - to help new Wikipedians learn editing techniques, to enrich content about Morocco or work on and develop a certain topic.
- **Meetings** - for the Wikimedians in Morocco to exchange ideas and to check what support do they need to continue editing Wikipedia.
- **Editing contests** - to encourage existing users edit Wikipedia and to encourage non-users join.
- **Photography trips and contests** - to encourage users upload photos to Commons.

Wikipedia in Morocco - General Overview

- Languages taught at school:
 - Standard Arabic – Mandatory
 - French – Mandatory
 - English/Spanish – Mandatory
 - Standard Tamazight (Berber) – Optional
- Official languages
 - Arabic
 - Tamazight (Berber)
- Mixed use of all these languages, in addition to some others

Wikipedias in Morocco - General Overview



@Wikimedia Foundation, CC-BY-SA (author Erik Zachte)

Many thanks to Mark DiMarco for DataMaps and to *Our world in Data* team for inspiration. Data supplied by WMF Analytics Team.

See also static version, with Overview and breakdown of pageviews By Country and By Language.

WiViVi version 0.9

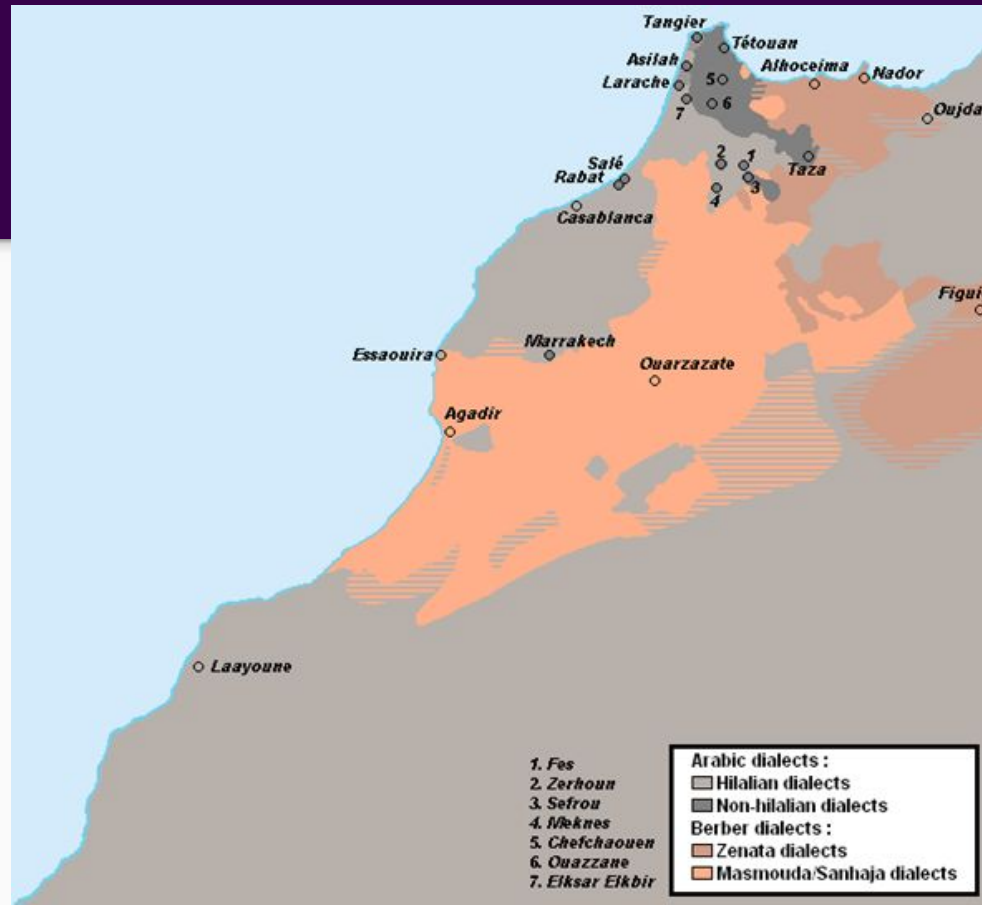
Data read: 223 flags, 10 regions, 223 countries, 96 languages. Canvas: 1536x722

Languages ranked by monthly pageviews. Only languages shown with at least 0.1% of traffic in at least one country.

Moroccan Darija - Linguistic Overview

- Vernacular Arabic spoken in Morocco
- Heavily influenced by Berber, and to a lesser extent by French and Spanish
- Unclear Official Status
- Spoken in daily life
 - ~30 million native speakers
 - Strong presence in media
 - Weak to unexisting presence in written tradition
 - Mostly Oral tradition
- Has its own Wikipedia since July 2020

Moroccan Darija - Linguistic Overview



Wikipedia in Darija - Background

- First Request for a Darija Wikipedia -> January 2008

https://meta.wikimedia.org/wiki/Requests_for_new_languages/Wikipedia_Moroccan

- Very early and persistent Request
- Distinction from the Middle-East
- Creation of a "Moroccan" Wikipedia

Darija Wikipedia - 2011

!Merħḥba bikom fe *Wikipédya* be *I-loġa I-mġerbiyya*


.Hada waħed I-mećroġ bać nšaybo Wikipédya be I-loġa I-mġerbiyya, o merħḥba be I-li bġa yćarek mġana

: Bac tšayeb ci šefħa, kteb I-ġonwan dyalha mbeġd **Wp/ary/ o brek ġla **Šayeb****

O zid ġafak had I-kod fe I-leħher dyal š-šefħa dyaltek bac ttzad mġa š-šefħat I-ħrin : **[[Category:Wp/ary]]**

/Wp/ary

Šayeb



WIKIPÉDYA
L-Mewšoġa L-Ħorra

Gadad š-šefħat: **+1390**

.Brek **hnaya**, bac tcof š-šefħat I-li kaynin

Hada Wikipédya be I-loġa I-mġerbiyya, o mektob be I-ħrof t-romiyya dyal ktbdarija.com, bać ĩkon I-qraya dyal I-mġerbiyya sahla. L-Loġa I-mġerbiyya hiya loġa kaiji le I-ġerbiyya, belħeq ettro ġliha bezzaf ĩ loġat ĩrin I-li homa t-ĩamaziġi, š-špenyoliyya, I-řanřawiyya... L-'Aġlabiyya .dyal I-mġarba kayġerfo yhedro biha

: Kaynin fe dahel I-loġa I-mġerbiyya bezzaf de I-leħjať, I-li homa

: **L-Leħjať L-La Hilaliyya**

L-Leħjať dyal I-modon I-qdam (Ĥađarĩ): L-Leħjať I-qdam dyal R-Rbat, Fas, Ĥitwan, Ĥaza, Sla : Kanjebro fihom ĩ kelmať majyin men š-špenyoliyya ola I-ġerbiyya I-'andalosiyya

L-Leħjať dyal Jbala (qabl-Hilaliyin): Kaytkellmo bih fe I-mentaqa dyal R-Rif I-ġerbi, o fe sawaħil ĩamal-ġerb L-Meġrib mġa Jbala, Senhaja o .Znata

.L-Leħjať I-li 'ettro ġlihom Jbala : L-Leħjať I-qdam dyal Tanja, Wazzan, Āefćawen

.L-Leħjať I-li 'ettro ġlihom I-leħjať dyal I-badiya : L-Leħjať I-qdam dyal Merrakeć o Meknas

.L-Leħjať dyal I-ĩhod I-mġarba

Id	Id	Id	Id	Id	Id
1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36
37	38	39	40	41	42
43	44	45	46	47	48
49	50	51	52	53	54
55	56	57	58	59	60
61	62	63	64	65	66
67	68	69	70	71	72
73	74	75	76	77	78
79	80	81	82	83	84
85	86	87	88	89	90
91	92	93	94	95	96
97	98	99	100	101	102
103	104	105	106	107	108
109	110	111	112	113	114
115	116	117	118	119	120
121	122	123	124	125	126
127	128	129	130	131	132
133	134	135	136	137	138
139	140	141	142	143	144
145	146	147	148	149	150
151	152	153	154	155	156
157	158	159	160	161	162
163	164	165	166	167	168
169	170	171	172	173	174
175	176	177	178	179	180
181	182	183	184	185	186
187	188	189	190	191	192
193	194	195	196	197	198
199	200	201	202	203	204
205	206	207	208	209	210
211	212	213	214	215	216
217	218	219	220	221	222
223	224	225	226	227	228
229	230	231	232	233	234
235	236	237	238	239	240
241	242	243	244	245	246
247	248	249	250	251	252
253	254	255	256	257	258
259	260	261	262	263	264
265	266	267	268	269	270
271	272	273	274	275	276
277	278	279	280	281	282
283	284	285	286	287	288
289	290	291	292	293	294
295	296	297	298	299	300

Darija Wikipedia - 2013

مرحبا بكم ف ويكيبيديا ب داريجة لمغربية

هادا واحد لمشروع باش نصايبو ويكيبيديا ب داريجة لمغربية و مرحبا بللي بغا يشارك معنا

/Wp/ary / باش تصايب شي صفحة، كتب لعنوان ديالها من بعد و برك على صايب

و زيد عفاك هاد لكود ف لقر باش صفحة للي صايبت تزداد معا صفحات لآخرين

[[Category:Wp/ary]]

/Wp/ary

شايب

عدد الصفحات: **1390+**
برك هنايا. باش تشوف الصفحات للي كاينين

Hada Wikipédia be l-loġa l-mġerbiyya, o mektob be l-ħrof ř-řomiyya dyal ktbdarija.com, bač řkon l-qraya dyal l-mġerbiyya sahla. L-Loġa l-mġerbiyya ĥiya loġa katji le l-ġerbiyya, belħeq ettiro ġliha bezzaf ři loġat řrin l-li homa ř-řamaziġt, ř-řpenyoliyya, l-řanřawiyya... L-'Aġlabiyya .dyal l-mġarba kayġerfo yhedro biha

: Kaynin fe daħel l-loġa l-mġerbiyya bezzaf de l-leħjai, l-li homa

: *L-Leħjai L-La Hilaliyya*

Rank	Access	Rank	Access	Rank
1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5
6	6	6	6	6
7	7	7	7	7
8	8	8	8	8
9	9	9	9	9
10	10	10	10	10

Darija Wikipedia - 2015

مرحبا بكم ف ويكيبيديا بالدارجة المغربية

هذا مشروع باش نصايو ويكيبيديا بالدارجة المغربية و مرحبا بللي بغا يشارك معنا

`/Wp/ary` باش تصايب صفحة، كتب عنوانها من بعد
واضبط على صايب

و زيد عفاك هاد الكود ف الجر باش صفحة اللي صايبت تتراد معا الصفحات الأخر

[[Category:Wp/ary]]

عدد الصفحات: **1390+**

اضغط [هنا](#) باش تشوف الصفحات اللي كاينا

`/Wp/ary`

صايب



ويكيبيديا
الموسوعة الحرة

Darija Wikipedia - 2018

ويكيبيديا بالدارجة فيها حتا لداها 93 مقال

مُرحبا بكم فويكيبيديا بالدارجة المغربية

لي كل حد بقدر يقرأ و يكتب فيه بلا ما يخلص !

لماط  سيونس  تكنولوجيا  ثقافة و فنون  مجتمع  سيرة داتية  جغرافيا  تاريخ  المغرب

شنو هي ويكيبيديا



ويكيبيديا هو واحد المشروع فيه براف د اللغات و الهدف دها هو يصاوب موسوعة دقيقة و متكاملة و متتوعة و مفتوحة لكلشي. إلا كنتي كانهضر الدارجة د المغرب أجي شارك معنا فهاد المشروع. بعينا ويكيبيديا تكون تا باللغة ديالنا.

الدعوة موجهة لكأع المهتمين بالمشاركة في تطوير الموسوعة بالمقالات الهادفة. عفاك توجه لصفحة النقاش للمناقشة العامة

ويكيبيديا بالدارجة



هادا مشروع باش نصايبو ويكيبيديا بالدارجة المغربية و مرحبا بللي بها يشارك معنا Hada mchro3 bach nsaybo Wikipedia b darija d lmghrib. Mr7ba belli bgha ichark m3ana

باش تصايب صفحة، كتب عنوانها من بعد Wp/ary و برك على صايب و زيد عفاك هاد الكود ف الجُر باش صفحة اللي صايبتي تتراد معا الصفحات الأخر:

[[Category:Wp/ary]]

مقالة مخيرة



تاريخ دبال المغرب من بعد الثقافات لي بانو قبل التاريخ بحال لي تلقاو الآثار دبالها **فجبل إبعود و تافوغالت** تاريخ لآلاف السنين. وتبيدا من تأسيس دبال **مملكة موريطانية** والممالك الأمازيغية القديمة. حتى لتأسيس دبال الدولة المغربية من طرف سلالة **الأدارسة** ومن بعد جاو سلالات اسلامية أخرى، ومن تما لعهد الحماية والتفترات مابعد الإستقلال. (كمل القراءة...)

مقالة مزينة



ويكيبيديا (بلينگليزيا: Wikipedia) هي واحد الموسوعة حرة **فالانترنت**. الهدف دبالها هي تخلي أي واحد إصاوب و يعدل المقالات. ويكيبيديا هي أكبر واشهر موقع دبال المعلومات فالانترنت. و جا فالرتبة الخامسة من بين المواقع المعروفة. وهو واحد المشروع دبال مؤسسة ويكيبيديا.

(كمل القراءة...)



Wikipedia Darija - User Group involvement

- Discussion with User Group Members
 - Positive feedback towards the Darija projects
- Creation of a Darija Task Force
 - Fully volunteering capacity
 - Sharing different tasks, including:
 - Article creation
 - Article improvement
 - Audio recording
 - Media Wiki Translation
 - Standardization and policy making
 - In collaboration and agreement with the online community
 - Conflict-management and mediation
 - Advocate for the Darija Wikis in media
- Giant online one-year "Edit-a-thon"

Wikipedia Darija - Current Status (2021)

مرحبا بيدا ف ويكيبيديا

لموسوعة لي أي واحد يمكن يساهم فيها. كايبة تال دابا 4,505 مقالة ب دارجة.

لغتون	سبور	سيفيات	طب	سيونص	فلسفة	شيمي	لأديان	لجغرافيا
			تاريخ	لماط	تكنولوجيا			

🌐 ساحة د جماعة 🗨️ لبرلمان 🗳️ ميزان لكلام 🌱 دوزان W على ويكيبيديا 📄 حقوق نشر



اليوم: "لاربيع 25 غشت 2021، موافق 16 محرم 1443هـ"

ويكيبيديا دارجة

مشروع تشاركي باش نتجو إنسكلوبيديا حرة، فابور، ؤ متكاملة. هادا سيت ختباري باش نشوفو واش لإنسكلوبيديا واجدة ؤ لإقبال عليها كاي، دابا وصلنا ل 4,505 أرتيكل. لهدف ديالنا نوصلو ل 5000 أرتيكل باش نزيدو نطورو لموسوعة. لمحتوى ديال ويكيبيديا منشور تحت رخصة كُو ل لوثائق لحرة، شي لي كايغني بلي حور باش يتعاود يتستعمل.

أش طاري ؟



- 15 غشت: طالبان عاودو حكمو أفغانستان مورا وصول ل لعاصمة كابول ؤ رايص أشرف غني هرب.
- 28 يوليوز: الوداد البيضاء داو لبوطولا پرو. والمغرب

لأرتيكل لمزيانة ديال ليوم

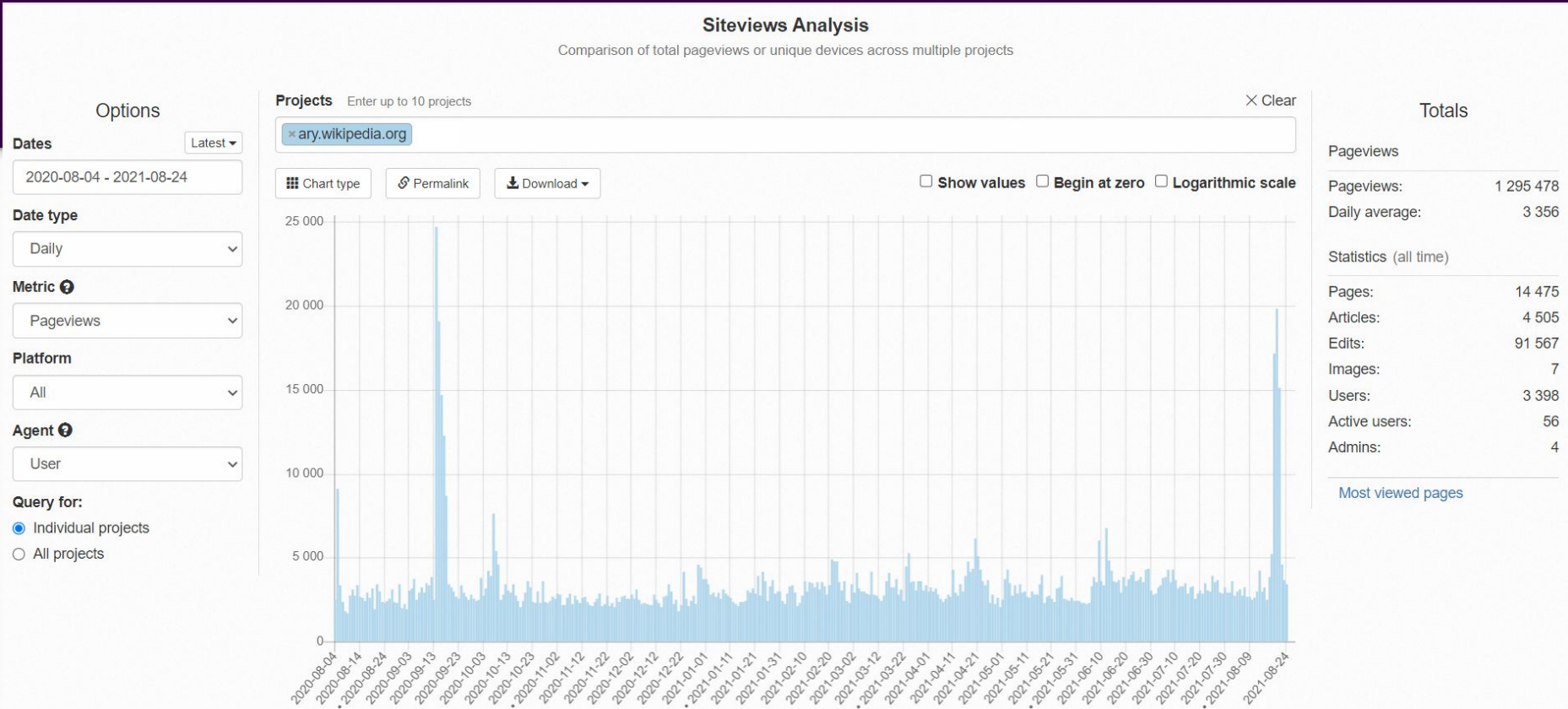


بدا

كاس محمد لغامس هي بوطولة صغيرة ودية كانت كاتتضم فلمغريب (مدون الرباط، الدار البيضاء ومراكش) من 1962 حتال 1989، وكانو كايشاركو فيها ربعة ديال لفراقي، علاقل وحدة من لمغريب ولياقي من لعالم. شاركو فهاد لبوطولة فراقي كبار من لعالم بحال إف سي برصونة وريال مدريد وباريس ميونخ، ومونتخابات بحال ديال رومانيا وتشيكوسلوفاكيا، وفراقي مغربية بحال الوداد البيضاء، الرجاء البيضاء، الجيش الملكي وحسنية أكادير. أكثر فرقة شاركات فلبوطولة هي الجيش الملكي (8 دلمرات) ووصل لفينال جوج مرّات، وأكثر فرقة ذات لكاس... - كمل(ي) القرارية ...

تصويرة لمزيانة ديال ليوم

Wikipedia Darija - Current Status



Wikipedia Darija - Current Status

تعديل ضایة بايکال



ع م      [متقدم](#) < [أحرف خاصة](#) < [مساعدة](#)

{{مقالة ناقصينها عيون لكلام}}

[<p>Baikal.A2001296.0420.250m-NASA.jpg:ملف</p>يسار|تصغير|<P ALIGN=center>ضاية بايكال من السما]

[[ملف:Lake Baikal in winter.jpg|يسار|تصغير|P ALIGN=center>بلاصة محفدة ف الضاية د بايكال</p>]]

'''ضاية بايكال'''، (بالروسية: о́зеро Байка́л) هي ضاية جات ف لجنوب د لمنطقة د سيبيريا، ف [[روسيا]]. هي الضاية اللي عندها أكبر حجم د لما ف لعالم، و فيها تقريبا 22-23% د ببال لما لحلو د لعالم. هي تا أغرق ضاية ف لعالم، و أغرق بلاصة فيها واصله ل 1642 ميטرو د لغرق. بايكال هي وحدا من أقدم الضايات ف لعالم (عندها ما بين 25 و 30 مليون عام)، و هي سابع أكبر ضاية من جهته لمساحة.

الضاية ديال بايكال مسجلة ف اللائحة د التراث العالمي د [[ليونيسكو]] من 1996.

== عيون لكلام ==

{ {عون} }

 $\{\{زراعة\}\}$

{{ضبط مخازن}}

[[تصنيف:ضايقة]]

[[تصنيف: روسيا]]

Another Darija project - Wiktionary

كائن دابا 866 كلمة بداريجة
لفيرسيون لمغربية دبال لبروجي Wiktionary
نسخة دبال تليفون

مرحبا ببيكم ف ويكاموس

لقاموس لحر و لفايور لي أي واحد يمكنلو يحسنو

داريجة — Lingaedjes — Alfabet

أ - ؤ - إ - ي - پ - ت - ج - ح - خ - د - ر - ز - س - ش - ص - ض - ط - ع - غ - ف - ث - ق - ك - گ - ل - م - ن - ه - و - ي

تقديم

لويكاموس هو ديكسيونير مكتوب بداريجة، حر و فابور، لي أي واحد يمكن لو يعاون فيه، ولي كيشرح لكلمات، عبارات، متلات لي كابين بلهجات كاملين دبال داريجة. ويكاموس هو ديكسيونير وصفي كيفس لكلمة بنفس لوعة وكيعطي معاها شرح، مرادفات، مشتقات و حتا نطق.

— تاريخ د ويكاموس

جماعة

ويكاموس محافظة عليه واحد جماعة كاترحب بكلشي، ومستاعدة تعاون أي واحد !

- سول شي سؤال على شي كلمة ولا معنى
 - زيد شي كلمة ناقصة
 - قتارح شي تحسينات ولا أفكار جديدة
 - سول شي سؤال على لمعاونة
 - وبطيعة لحال يمكن ليك حتا نتا تعاون ف لمحتوى بتصحح ولا زيادات.
- ماشي ضروري يكون عندك كونس باش تشوف ولا تعاون، والاكين عادي يعاونك باش تتعلم من ناس لي عندهم خبرة.

كلمة دبال ليوم

[[Template:Wt/ary/كلمة د ليوم/2021/08/25]]

كتاشف لمشروع

كيفاش تشارك :

- بترتيب
- على حساب نوع

كيفاش تشارك :

- طريقة لكتابة
- صفاحي دبال معاونة

مشاريع خرين بداريجة

Another Darija project - Wiktionary

بغل/Wt/ary

ary | Wt >
< بغل > ary

لمضمون [خبي]

1 داريجة

1.1 أصل

1.2 سمية

1.3 نطق

داريجة [بدل | بدل لكود]

أصل [بدل | بدل لكود]

من عربية **بغل**, **بَاغِل**

سمية [بدل | بدل لكود]

بغل \byəl\

1. كائن مزبود فاش كيتزاوج حمار دكر معا عودان نتوة، كيكون عادة عاكر، ؤ عندو قدرة على تحمل كتر من لحمار.
2. سبة كنتال ف حق شي واحد دار فعلة غيبة ؤلا حقيرة.

نطق [بدل | بدل لكود]

• سلا : سمع « بغل/Wt/ary [byəl] »
0:00 ڤكسة

جمع	فرد	
بغال/Wt/ary	بغل/Wt/ary	دكر
بغلات/Wt/ary	بغلة/Wt/ary	نتوة



بغا

بغل

Darija Wikipedia - Remaining challenges

- **Standardization**
 - No official Status
 - Different regions/dialects
- **Keyboards**
 - Arabic Vs Latin Letters
- **Vandalism**
 - Ideological fights
 - Dialect Vs language
- **Scattered online community**
 - Sporadic interest

Further Readings

- Moroccan Arabic (Darija) – Wikipedia Article - https://en.wikipedia.org/wiki/Moroccan_Arabic
- Moroccan Darija Wikipedia - <https://ary.wikipedia.org>
- List of Wikipedias per language - https://meta.wikimedia.org/wiki/List_of_Wikipedias
- Moroccan Darija Wiktionary (incubator) - <https://bit.ly/3zgheic>
- Moroccan Darija in Wikipedia – Paper - <https://revistas.uca.es/index.php/aam/article/view/5795/6028>
- Wikimedia Morocco User Group - https://meta.wikimedia.org/wiki/Wikimedia_MA_User_Group

Explore the Moroccan Darija Wikipedia

3:40-3:50



Introduction to NLP

NLP

- **Definition**
 - Allow Computers to process and analyze large amounts of natural language data.
- **Tasks**
 - SA, QA, NER, Text Generation, Topics Detection, Machine Translation, Language Detection, Targeted Advertising, etc.
- **Datasets**
 - BooksCorpus
 - English Wikipedia
 - GLUE benchmarks
- **Preprocessing steps**
 - Character Encoding, Tokenization, Part-of-Speech Tagging, Chunking, Stemming and Lemmatization, Parsing, Regular Expressions.
- **State of the art**
 - Text representation: Word2vec, GloVe and FastText
 - Self-supervised learning: ELMo, GPT, ULMfit, BERT, XLnet, T5
- **Libraries**
 - NLTK, Spacy, Gensim, HuggingFace, SparkNLP, etc.

Arabic NLP

• Challenges

- Encoding, phonetic, morphology, syntax, semantic
- Classical Arabic, Modern Standard Arabic and Dialect Arabic
- Right to Left
- Diacritic marks (tashkil)
- Letter shape in different positions (example letter f: "ف"/"ف"/"ف")
- Hamza spelling: "أ", "ؤ", "ئ" .. سؤال, سئل, مكتب, رأسمالية
- No capital letters
- Morphology: سيكتبونها, كتب, كتاب, مكتب, رأسمالية
- Transliteration: كتاب -> ktAb, مكتب -> mktb
- ...

• Datasets

- **SA:** HARD (Hotel Arabic Reviews Dataset)
- **NER:** ANERcorp (Arabic NER Corpus)
- **QA:** ARCD (Arabic Reading Comprehension Dataset),
- Training **Language Models:**
 - SANAD (Single-Label Arabic News Articles Dataset for Automatic Text Categorization)
 - OSIAN (Open Source International Arabic News Corpus)
 - Abu El-Khair Corpus (Arabic Corpus)
 - Arabic Wordnet

• Libraries

- Madamira, Farasa, Camel, AraBERT,

Arabic Dialect NLP

- Challenges

- **Dialects:** Maghrebi (North Africa), Hassaniya, Egyptian, Levantine (Syria, Lebanon), Mesopotamian (Iraq), Arabian Peninsula (Gulf), Jordanian, Sudanese, Yemeni, etc.
- **Romanized:**
 - Arabizi, Aransi
- Numbers as letters:
 - maba3rafsh, 9oli, 5oya, 7amdoullah,
- **Inconsistent spelling:** chefti, chetti, chofti,

- Sources

- Social Media, Chats, Forums, Wikis, Emails, SMS, etc.
- User Generated Text (UGT)

- Datasets

- **SA:** ASTD(twitter+egyptian), ArSenTD-Lev (Twitter+Levantine), AJGT (Arabic Jordanian General Tweets)
- **NER:** ?
- **QA:** ?
- MADAR (Multi-Arabic Dialect Applications and Resources)

- Preprocessing steps

- Detect Dialect, remove o..., characters encoding (tfinagh letters)

- Libraries

- TunBERT
- Madamira, Farasa, Camel, AraBERT,

Moroccan Darija NLP

- Datasets

- SA: MSTD [1] (Moroccan Sentiment Twitter Dataset)
- QA: ?
- NER: ?
- Moroccan Darija Wikipedia
- MDED (Moroccan Dialect Electronic Dictionary)
- Moroccan Darija WordNet (MDW) [2]
- DODa (Darija- Open Dataset) [3]

- Challenges

- Backwalters, romanized,
- Resume words (B1, Bn8,)

- Librairies

- ?

- MSTD,



Cleaning & Preprocessing Text Data

Text Cleaning

- **Gathering, sorting, and preparing data** is the most important step in the data analysis process – bad data can have cumulative negative effects downstream if it is not corrected.
- **Data wrangling**, meaning the manipulation of data so that it is most suitable for machine interpretation is therefore critical to accurate analysis.
- The goal of data prep is to produce '**clean text**' that machines can analyze error free.

Text Cleaning

- **Removing extra spaces** - Most of the time the text data that you have may contain extra spaces in between the words, after or before a sentence. So to start with we will remove these extra spaces from each sentence by using regular expressions.

« حمر لڭلالشه جماعة ترابية قروية كاينة في إقليم تارودانت، جهة سوس ماسة »

➡ « حمر لڭلالشه جماعة ترابية قروية كاينة في إقليم تارودانت، جهة سوس ماسة »

- **Removing punctuations** - The punctuations present in the text do not add value to the data. The punctuation, when attached to any word, will create a problem in differentiating with other words.

! السنوات للي خدا فيها اللقب

➡ السنوات للي خدا فيها اللقب

- **Removing latin & special characters** - Sciences de l'Archéologie et du Patrimoine | page=19 | لعصر الحجري لقديم
}} فساحل د لمغريب : رباط، تمارة ؤ لمعمورة

لعصر الحجري لقديم فساحل د لمغريب : رباط، تمارة ؤ لمعمورة

➡

Preprocessing

- **Tokenization** : Splitting a sentence into tokens and creating a list.

Ex:

```
sample = ''
```

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

يُشار، "إلى"، "أن"، "اللغة"، "العربية"، "يتحدثها"، "أكثر"، "من"، "422"، "مليون"، "نسمة"، "و"، "يتوزع"، "متحدثوها"، "في"، "المنطقة"، "المعروفة"، "باسم" ["الوطن"، "العربي"، "بالإضافة"، "إلى"، "العديد"، "من"، "المناطق"، "الأخرى"، "المجاورة"، "مثل"، "الأهواز"، "و"، "تركيا"، "و"، "تشاد"، "و"، "السنغال"، "و"، "إريتريا"، "و"، "غيرها"، "و"، "هي"، "اللغة"، "الرابعة"، "من"، "لغات"، "منظمة"، "الأمم"، "المتحدة"، "الرسمية"، "الست"]

- **Removing Stopwords** - We fetch a list of stopwords in the Arabic dictionary. Then, we remove the tokens that are stopwords. Stop words include : 'إذا', 'إذما', 'إذن', 'أف', 'أقل', 'أكثر', 'ألا', 'إلا', 'التي', 'الذي'

Preprocessing

- **Stemming** : Normalize words into its base root form || done by cutting the end or the beginning of words.

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

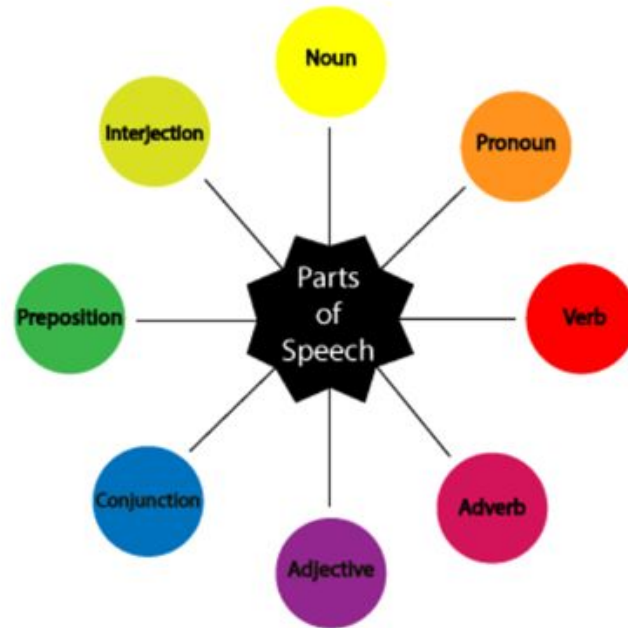
- **Lemmatization** : Takes into consideration the morphological analysis of the word by using a dictionary. Ranges together diff words with the same meaning. Ex: It could map together Going, went, gone into GO.

Very challenging for the Moroccan dialect.

انْقَضَى ، مَضَى ، مُنْتَهَى ، انْتَهَى ، مُسْتَهْلَك ، نَافِد ، تَحَرَّكَ ، تَوَجَّه ، دَهَبَ ، سَافَرَ

Preprocessing

- **POS Tags:** Grammatical type of the word
- **Named Entity Recognition**



30 min Lab 4:10-4:40

Lab1: Wikipedia Darija Cleaning



<https://github.com/MoroccoAI/AMLD>

20 min Break 4:40-5:00



30 min Lab 5:00-5:30

Lab2: Wikipedia Topic Detection



<https://github.com/MoroccoAI/AMLD>

30 min Lab 5:30-6:00

Lab3: NLP Tasks & Tools



<https://github.com/MoroccoAI/AMLD>

20 min Break 6:00-6:20



Explore on your own in breakout rooms 6:20-8:20

<https://github.com/MoroccoAI/AMLD>



Discuss your Results 8:20-8:50



What Next ?

- **More?**

- Train a language model
- Develop new tools

- **Contribute**

- Share your work/data/notebooks on Darija with the community
- Contribute to Darija Wikipedia

- **Datasets**

- Moroccan Darija Wikipedia
- MSTD (Moroccan Sentiment Twitter Dataset)
- MDED (Moroccan Dialect Electronic Dictionary)
- Moroccan Darija WordNet (MDW)
- DODa (Darija- Open Dataset)
- ...

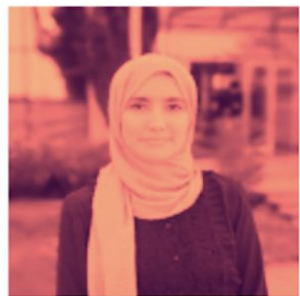


✦ Join Us

Some papers

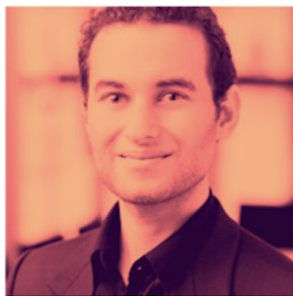
- Shaalan, Khaled & Siddiqui, Sanjeera & Alkhatib, Manar & Monem, Azza. (2018). Challenges in Arabic Natural Language Processing. 10.1142/9789813229396_0003.
- Mihi, Soukaina & AIT, Brahim & EL, Ismail & Arezki, Sarā & Laachfoubi, Nabil. (2020). MSTD: Moroccan Sentiment Twitter Dataset. International Journal of Advanced Computer Science and Applications. 11. 10.14569/IJACSA.2020.0111045.
- Mrini, Khalil, and Francis Bond. "Building the moroccan darija wordnet (mdw) using bilingual resources." *Proceedings of the International Conference on Natural Language, Signal and Speech Processing (ICNLSSP)*. No. CONF. 2017.
- Aissam Outchakoucht and Hamza Es-Samaali. Moroccan Dialect -Darija- Open Dataset, 2021, Arxiv
- Ridouane, Tachicart & Bouzoubaa, Karim. (2021). Moroccan Data-Driven Spelling Normalization Using Character Neural Embedding. Vietnam Journal of Computer Science. 8. 1-19. 10.1142/S2196888821500044.

Thanks!



Imane Khaouja

PhD student,
Mohammed V
University



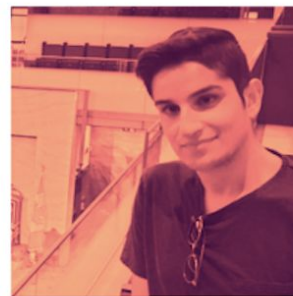
Anass Sedrati

Vice President,
Wikimedia Morocco



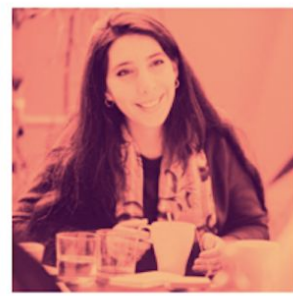
Abdelhak Mahmoudi

Associate Professor,
Mohammed V
University



Khalil Mrini

PhD Candidate in NLP,
University of California
San Diego



Ihsane Gryech

PhD candidate,
International
University of Rabat

