بناء وتطوير ذخيرة لُغويَّة لمُحرِّك بحث عربيّ مفتوح المصدر

دكتور المُعتزّ بالله السَّعيد

ورقة عمل مُقدَّمة إلى

المُنظَّمة العربيَّة للتَّربية والثَّقافة والعُلُوم
ALECSO

اجتماع خُبراء مُحرِّكات البحث العربيَّة
تونُس، 26-27 أبريل 2017

1. توصيف الذَّخيرة المُنجَزة:

- استمدَّت الذَّخيرةُ *Linguistic Corpus* مادَّتَها من مورِدَين رئيسَين؛ على النَّحو الآتي:

م	المورد	عدد الوثائق	عدد الكلمات	الكلمات الفريدة
1	موسوعة ويكيبيديا	306366	42894090	931790
2	صحيفة الحياة	26919	20072264	408681

- تمَّت تهيئة الذَّخيرة في صُورَتَين؛ إحداهُما: الذَّخيرة بصُورتها الخام *Raw Corpus*، والأخرى في قاعدة بيانات، بصيغة *XML*.
- تمَّ استخلاص قوائم الكلمات الفريدة *Unique Words* لكلّ وثيقة على حدة؛ وتمَّ التَّعامُل معها باعتبارها كلمات مِفتاحيَّة *Key Words*.
- تمَّ استخلاص إحصاءات الذَّخيرة. واشتمَلَت هذه الإحصاءات على:
  1. إحصاءات الكلمات.
  2. إحصاءات الكلمات الفريدة.
  3. إحصاءات تردُّدات الكلمات الفريدة.
  4. إحصاءات وثائق الذَّخيرة، مع تصنيف المقالات.

2. نموذج من الذَّخيرة المُنجَزة:

عُنوان المقالة	أيون الهيدروجين	
المورد	موسوعة ويكيبيديا	
نموذج النُّصُوص	أيون الهيدروجين. أوصى الاتحاد الدولي للكيمياء البحتة والتطبيقية *IUPAC* بمصطلح عام لأيونات الهيدروجين ونظائره بالاعتماد على شحنة الأيون. 1- أيون موجب الشحنة. 2- أيون سالب الشحنة. وتحت ظروف مائية معينة كما توجد في الكيمياء الحيوية توجد أيونات الهيدروجين في صورة هيدرات مكونة جزيئات الهيدرونيوم *H3O+*, ويسمى غالبا أيون الهيدروجين أو بروتون في الكيمياء الحيوية. - أيون الهيدروجين :له شحنة أولية موجبة واحدة (*H+*) - البروتون : *H1+* هو أيون الهيدروجين، ويحمل شحنة أولية واحدة موجبة، يستخدم هذا التعبير مرادفا لأيون الهيدروجين. - ديترون : *H+, D2*، وهو أيون الهيدروجين الثقيل (نظير) حيث تتكون نواته ليس من بروتونا واحدا فقط وإنما يتكون من 1 بروتون و1 نيوترون لهاذا يسمى الهيدروجين الثقيل، ويرمز إليه بالرمز *D*، وهو يحمل شحنة أولية واحدة موجبة مثل أيون الهيدروجين. ...	
إحصاء الكلمات	179	
إحصاء الكلمات الفريدة	98	
الكلمات الفريدة المفتاحيَّة	**الكلمة**	**التَّردُّد**
	الهيدروجين	16
	أيون	14
	شحنة	5
	بروتون	3
	موجبة	3
	الحيوية	2
	الكيمياء	2
	الهيدرونيوم	2

3. موارد مُقتَرَحة لتطوير الذَّخيرة:

- تقترحُ الورقة إثراءَ مادَّة الذَّخيرة بنُصُوص العربيَّة المُعاصِرة المُستمدَّة من الموارد الآتية:

1. وثائق مُؤتمر (تيد *TED*):

- تأتي هذه المادَّة تمثيلًا للُّغة العربيَّة العلميَّة.
- يُتوقَّع ألَّا يقلَّ عددُ كلمات هذه المادَّة عن 2 مليون كلمة.
- تضُمُّ التَّرجمة العربيَّة لمُحاضرات المُؤتمر الأكاديميّ (تيد *TED*).
- يمتدُّ الإطارُ الزَّمنيُّ لهذه النُّصُوص من عام (2012م) حتَّى عام (2017م).

2. وثائق مُنظَّمة الأُمم المُتَّحدة (*UN*):

- تأتي هذه المادَّة تمثيلًا للُّغة العربيَّة الرَّسميَّة.
- يُتوقَّع ألَّا يقلَّ عددُ كلمات هذه المادَّة عن 5 ملايين كلمة.
- تضُمُّ النُّسخة العربيَّة لقرارات مُنظَّمة الأمم المُتَّحدة (*UN*).
- يمتدُّ الإطارُ الزَّمنيُّ لهذه النُّصُوص من عام (2000م) حتَّى عام (2010م).

3. وثائق الصَّحافة العربيَّة المُعاصرة:

- تأتي هذه المادَّة تمثيلًا للُّغة العربيَّة العامَّة.
- يُتوقَّع ألَّا يقلَّ عددُ كلمات هذه المادَّة عن 10 ملايين كلمة.
- تضُمُّ نُصُوص المقالات الصَّحفيَّة المُستمدَّة من الصُّحف العربيَّة.
- يمتدُّ الإطارُ الزَّمنيُّ لهذه النُّصُوص من عام (2000م) حتَّى عام (2017م).

4. وثائق الموقع المفتوح (ويكي هاو *Wikihow*):

- تأتي هذه المادَّة تمثيلًا للُّغة العربيَّة العلميَّة.
- يُتوقَّع ألَّا يقلَّ عددُ كلمات هذه المادَّة عن 100 ألف كلمة.
- تضُمُّ النُّسخة العربيَّة لمقالات موقع الويكي (*Wikihow*) المعرفيّ.
- يمتدُّ الإطارُ الزَّمنيُّ لهذه النُّصُوص من عام (2008م) حتَّى عام (2017م).

5. وثائق الموسوعة الحُرَّة ويكيبيديا "مُستدرَكة":

- تأتي هذه المادَّة تمثيلًا للُّغة العربيَّة الموسوعيَّة.
- يُتوقَّع ألَّا يقلَّ عددُ كلمات هذه المادَّة عن 5 ملايين كلمة.
- تضُمُّ النُّسخة العربيَّة من الموسوعة الحُرَّة "ويكيبيديا *Wikipedia*"
- يمتدُّ الإطارُ الزَّمنيُّ لهذه النُّصُوص من عام (2014م) حتَّى عام (2017م).

4. المهامّ والمُخرَجات والإطار الزَّمنيّ لبناء وتطوير الذَّخيرة اللُّغويَّة:

المرحلة	المهامّ	المُخرَجات		الإطار الزَّمنيّ
1	استيراد وتهيئة وتنقية نُصُوص الذَّخيرة	الذَّخيرة اللُّغويَّة الخام [*XML*].		4 أشهر
		وثائق الذَّخيرة المُنقَّاة من الكشائد والزَّوائد.		
2	العمليَّات الإحصائيَّة	تحليل *N-Gram*	الكلمات الأحاديَّة (1)	4 أشهر
			المُتلازمات الثُّنائيَّة (2)	
			المُتلازمات الثُّلاثيَّة (3)	
		إحصاءات الكلمات والمُتلازمات	قائمة الكلمات الفريدة *Wordlist – 01 Gram*	
			قائمة الثُّنائيَّات الفريدة *Wordlist – 02 Gram*	
			قائمة الثُّلاثيَّات الفريدة *Wordlist – 03 Gram*	
3	العمليَّات الحاسُوبيَّة	الفهرسة الآليَّة	قوائم كلمات الذَّخيرة وتردُّداتها – لكلّ وثيقة	4 أشهر
		التَّعرُّف الآليّ على عناوين الوثائق	قائمة عناوين وثائق الذَّخيرة	
4	العمليَّات اللُّغويَّة	الوسم التَّركيبيّ	نماذج من الذَّخيرة اللُّغويَّة مُوسَّمة تركيبيًّا، توسيمًا آليًّا جُزئيًّا.	4 أشهر
		التَّجذيع *Stemming*	كلمات الذَّخيرة بعد القطع الجُزئيّ للزَّوائد	

		التَّقييم	نتائج تقييم عمل مُحرِّك البحث، استنادًا إلى مُخرجات الذَّخيرة	

5. فريق العمل المُقتَرَح لبناء وتطوير الذَّخيرة:

م	الباحث	التَّوصيف
1	المُعتزّ بالله السَّعيد	باحث رئيس
2	ريم الأعطر	باحث مُساعد

6. توفيق المهامّ والمُخرَجات مع مُتطلَّبات المُعالجة اللُّغويَّة وتقييم أداء مُحرِّك البحث:

المُعالجة	م	إجراءات المُعالَجة	المُخرَجات
المُعالَجة اللُّغويَّة	1	توسيع مُعجم النِّظام	قائمة الكلمات الفريدة *Wordlist – 01 Gram*
	2	مُطابقة مُفردات الذَّخيرة بمُفردات المُعجم	قوائم كلمات الذَّخيرة وتردُّداتها – لكلّ وثيقة
	3	تعيين أسماء الأعلام غير الموجودة في مُعجم النِّظام	قائمة عناوين وثائق الذَّخيرة
	4	تعيين المُصطلحات الدَّائرة غير الموجودة في مُعجم النِّظام	الكلمـات الأحاديَّة (1)
			المُتلازمات الثُّنائيَّة (2)
			المُتلازمات الثُّلاثيَّة (3)
	5	إحصاءات المُفردات والمُصطلحات	قائمة الكلمات الفريدة *Wordlist – 01 Gram*
			قائمة الثُّنائيَّات الفريدة *Wordlist – 02 Gram*
			قائمة الثُّلاثيَّات الفريدة *Wordlist – 03 Gram*
	6	إحصاء مُفرَدات الذَّخيرة بعدَ القطع السَّطحيّ للسَّوابق واللَّواحق	كلمات الذَّخيرة بعد القطع الجُزئيّ للزَّوائد
	7	تحديد نسبة نجاح القطع السَّطحيّ	كلمات الذَّخيرة بعد القطع الجُزئيّ للزَّوائد
تقييم أداء مُحرِّك	8	بناء قاعدة بيانات خاصَّة بالنُّصُوص العلميَّة المُعاصرة	الذَّخيرة اللُّغويَّة الخام [*XML*]
			وثائق الذَّخيرة المُنقَّاة من الكشائد والزَّوائد

	9	بناء استعلامات الذَّخيرة	الذَّخيرة اللُّغويَّة الخام [*XML*]
			وثائق الذَّخيرة المُنقَّاة من الكشائد والزَّوائد
	10	المُقارنة بينَ نتائج البحث الآليَّة والنَّتائج المُعَدَّة يدويًّا	الذَّخيرة اللُّغويَّة الخام [*XML*]
			قوائم كلمات الذَّخيرة وتردُّداتها – لكلّ وثيقة

7. نموذج مُخرَجات الذَّخيرة اللُّغويَّة لمُحرِّك البحث:

- يُمثِّلُ النَّموذجُ الآتي صُورة مُصَغَّرة من الذَّخيرة اللُّغويَّة المنشودة.

أوَّلًا: إحصاءات الوثائق:

م	عُنوان الوثيقة	عدد الكلمات	م	عُنوان الوثيقة	عدد الكلمات
1	آشور	*3940*	**26**	المحقق كونان	*2319*
2	آن بولين	*3760*	**27**	المشتري	*3948*
3	أغالبة	*3320*	**28**	الوطن العربي	*4473*
4	ألماس	*5990*	**29**	بابوية كاثوليكية	*2809*
5	ألومنيوم	*4607*	**30**	برج إيفل	*3717*
6	أورانوس	*4517*	**31**	بلاد الشام	*2750*
7	أور	*2881*	**32**	بنين	*2276*
8	إدوارد سعيد	*3240*	**33**	بيل غيتس	*3242*
9	إسحاق شامير	*2941*	**34**	تاريخ الإسكندرية	*5489*
10	إسحاق نيوتن	*3299*	**35**	تخلخل العظم	*2738*
11	إسرائيل	*3544*	**36**	تدخين	*3716*
12	إعصار قمعي	*3956*	**37**	تويوتا	*5498*
13	إلكترون	*4658*	**38**	جامعة الدول العربية	*3505*
14	الأرض	*6421*	**39**	جوجل	*4960*
15	الأندلس	*4073*	**40**	ج. ك. رولينج	*5140*
16	الإمبراطورية البيزنطية	*4246*	**41**	حديد	*4436*

17	الدولة المرينية	*4956*	**42**	حرب بونيقية ثانية	*4651*
18	الزهرة	*4156*	**43**	حشرة	*4194*
19	السكري	*3099*	**44**	ديمقراطية	*2600*
20	الشمس	*5249*	**45**	ذكاء اصطناعي	*2087*
21	الظاهر بيبرس	*3485*	**46**	ربو	*2791*
22	القاهرة	*4323*	**47**	روبوت	*3581*
23	القضية الفلسطينية	*5276*	**48**	رومانيا	*4240*
24	القمر	*4037*	**49**	زحل	*3942*
25	المجموعة الشمسية	*4597*	**50**	زينون	*4722*

م	عُنوان الوثيقة	عدد الكلمات	م	عُنوان الوثيقة	عدد الكلمات
51	ستيف جوبز	*6116*	**76**	كيمونو	*4214*
52	شركة فورد	*3592*	**77**	لاوس	*2754*
53	صحافة	*4221*	**78**	لغة سريانية	*3480*
54	صهيونية	*2638*	**79**	لوحة أم	*2578*
55	ضفدع	*2992*	**80**	ليثيوم	*3501*
56	طاقة شمسية	*3953*	**81**	ماري أنطوانيت	*2339*
57	عبد الحميد بن باديس	*2761*	**82**	ماري كوري	*3573*
58	عصر النهضة	*5592*	**83**	مالكوم إكس	*4745*
59	عكا	*3243*	**84**	مايكروسوفت ويندوز	*3175*
60	عيد الحب	*3987*	**85**	محرك بحث	*2731*
61	غابرييل غارثيا ماركيث	*4177*	**86**	مدغشقر	*3806*
62	غابرييل فوري	*2240*	**87**	مسيحية	*3638*
63	غيشا	*4861*	**88**	مندائية	*3067*
64	فتوحات إسلامية	*3266*	**89**	منظمة التحرير الفلسطينية	*2646*
65	فرانكلين روزفلت	*1827*	**90**	مونستر	*12488*
66	فلاديمير لينين	*2270*	**91**	ميلانو	*4258*
67	فولتير	*2949*	**92**	ميونخ	*3852*

68	فيروس	*3642*	**93**	نفط	*2663*
69	فيلكا	*2679*	**94**	نيكولاي ريمسكي كورساكوف	*3933*
70	فينسنت فان غوخ	*3380*	**95**	هيدروجين	*2978*
71	كارلوس الثعلب	*3819*	**96**	هيليوم	*2813*
72	كازاخستان	*4128*	**97**	ولاية برج بوعريريج	*5024*
73	كهرباء	*2691*	**98**	يهودية إصلاحية	*4608*
74	كوكب	*3263*	**99**	يوفنتوس	*2835*
75	كولاجين	*3309*	**100**	يوهانس برامس	*5367*

ثانيًا: إحصاءات الذَّخيرة:

Raw Corpus: Texts: 100 | Arabic Types: 59498 | Arabic Tokens:383057

ثالثًا: من قائمة الكلمات وتردُّداتها:

م	الكلمة	التَّردُّد
1	التخطيط	*8*
2	التخفيضات	*1*
3	التخفيف	*2*
4	التخفيفات	*1*
5	التخلص	*10*
6	التخلق	*1*
7	التخلل	*1*
8	التخلي	*14*
9	التخليق	*7*
10	التخمر	*1*
11	التخمين	*1*
12	التخوف	*2*
13	التخوم	*1*
14	التدابير	*4*
15	التداخل	*2*
16	التداخلات	*1*

17	التدخل	*11*
18	التدخلات	*1*
19	التدخين	*77*
20	التدرب	*1*
21	التدرج	*3*
22	التدرن	*1*
23	التدريب	*9*
24	التدريبية	*1*
25	التدريجي	*9*
26	التدريس	*6*
27	التدفق	*13*
28	التدفئة	*6*

رابعًا: قائمة المُتلازمات الثُّنائيَّة وتردُّداتها:

م	الكلمة	التَّردُّد
1	الشريط البحري	*2*
2	الشريط الذي	*1*
3	الشريط الساحلي	*2*
4	الشريعة الإسلامية	*2*
5	الشريعة الموسوية	*1*
6	الشريعة لشمس	*1*
7	الشريف التلمساني	*1*
8	الشريف الخليلي	*1*
9	الشريف حسين	*3*
10	الشريك التجاري	*1*
11	الشريك الرئيسي	*1*
12	الشريكان الرئيسيان	*1*
13	الشط الغربي	*1*
14	الشط المالح	*1*
15	الشطر الثاني	*1*
16	الشطر الغربي	*1*

17	الشعار الحالي	*1*
18	الشعار باللون	*1*
19	الشعار ذو	*1*
20	الشعار هو	*1*
21	الشعارات المعادية	*1*
22	الشعاع الكهربائي	*1*
23	الشعاع تصرف	*1*
24	الشعائر الإصلاحية	*1*
25	الشعائر المرتبطة	*1*
26	الشعائر اليهودية	*3*
27	الشعب الإسرائيلي	*1*
28	الشعب الثوري	*2*

خامسًا: قائمة المُتلازمات الثُّلاثيَّة وتردُّداتها:

م	الكلمة	التَّردُّد
1	الاتحاد الصهيوني البريطاني	*1*
2	الاتحاد العالمي لليهودية	*2*
3	الاتحاد العبري المعهد	*1*
4	الاتحاد العبري عام	*1*
5	الاتحاد العبري فرعا	*1*
6	الاتحاد العربي لكرة	*1*
7	الاتحاد الفلكي الدولي	*13*
8	الاتحاد المسيحي الديمقراطي	*1*
9	الاتحاد اليهودي للتربية	*1*
10	الاتحاد اليهودي للموسيقى	*1*
11	الاتحاد اليهودي مركز	*1*
12	الاتحاد أو الجمهورية	*1*
13	الاتحاد بدون مشغل	*1*
14	الاتحاد بين الشركتين	*1*
15	الاتحاد بين اليمن	*2*
16	الاتحاد في سنة	*1*

17	الاتحاد للإدارة العامة	*1*
18	الاتحاد مؤتمرا عالميا	*1*
19	الاتحاد مؤسسات في	*1*
20	الاتحاد وقالت بالطبع	*1*
21	الاتحادات القارية لكرة	*1*
22	الاتحادات القطرية العربية	*1*
23	الاتحادي الثالث وذلك	*1*
24	الاتحادي لتمثيل ألمانيا	*1*
25	الاتصال الأوروبي بدأ	*1*
26	الاتصال السريع مثل	*1*
27	الاتصال الشبكي الشهير	*1*
28	الاتصال الفكري بين	*1*

سادسًا: نموذج التَّوسيم التَّركيبيّ الآليّ لنُصُوص الذَّخيرة:

[PO]إن [CN]النتائج [RP]التي [VI]تمكن [PO]أن [VI]يحصل [PRE]عليها [CN]الملك [CN]الآشوري، [PO]لا [CN]سيما [PRE]في [CN]التخلص [PRE]من [CN]النفوذ [CN]الميتاني [CO]و [CN]التمكن [PRE]من اقتسام بلادهم، [PO]أن جعله [VI]يتوجه [AD]نحو توطيد أواصر علاقاته [CN]السياسية [AD]مع [CN]القوى [CN]السياسية [CN]الفاعلة، [AD]حيث [VI]أقدم [PRE]على [CN]الزواج [PRE]من ابنة [CN]الملك [CN]الكاشي [RP]الذي [VP]كان [VI]يفرض نفوذه [PRE]على بابل. [CO]و [PO]قد [VI]حظيت مملكة آشور بملوك خلفوا آشور أوبلط [CO]و [VP]كانوا [PRE]على مستوى [CN]المسؤولية [CO]و [VP]انتهجوا ذات [CN]الأسلوب [RP]الذي [VP]سار [PRE]عليه [PO]ل [VI]يثمر [PRE]عن [DE]ذلك [AD]خلال [CN]القرن [CNU]التاسع [AD]ق.م بلوغ مستوى [CN]الإمبراطورية [CN]الآشورية [PRE]ب [CN]كل قوتها ونفوذها [CN]السياسي. [PRE]من [CN]الملوك [CN]الآشوريين [CN]البارزين شلمنصر [CNU]الأول [RP]الذي [VP]دام حكمه [CNU]1266 - [CNU]1243 [AD]ق.م، [CO]و [VI]تطلع [PRE]إلى توجيه [CN]العديد [PRE]من [CN]الحملات [CN]العسكرية وعمل [PRE]على [CN]استبدال [CN]العاصمة آشور [PRE]ب [CN]مدينة نمرود. [PO]أما [CN]العمل [CN]الأبرز [CO]ف [VP]كان [PRE]على يد [CN]الملك توكلتي نورتا [CNU]1243 - [CNU]1221 [AD]ق.م، [RP]الذي [VP]تمكن [PRE]من [CN]السيطرة [PRE]على بلاد بابل، وتوسيع سلطانه [PRE]في [CN]الجهات [CN]الشرقية [CO]و [CN]الغربية. [PO]لكن [AD]بعد وفاة [DE]هذا [CN]الملك [VP]دخلت آشور [PRE]في مرحلة [CN]الضعف [CN]السياسي، نتيجة لوصول ملوك ضعاف [CN]الشخصية، [EX]غير قادرين [PRE]على إدارة مقاليد [CN]الحكم، [CO]و [VP]استمرت [DE]هذه [CN]الفترة [AD]حوالي [CNU]مائة عام، [CO]حتى بلوغ [CN]الملك تجلات بلاسر [CNU]الأول [CNU]1116 - [CNU]1090 [AD]ق.م [PRE]إلى سدة [CN]الحكم، [PO]ل [VI]تكون [DE]هذه [CN]الفترة مليئة [PRE]ب [CN]الإنجازات [CN]العسكرية [CN]الكبيرة، [AD]حيث [VP]تمكن [PRE]من تحقيق [CN]الانتصارات [CN]المتوالية [PRE]في [CN]الأصقاع [CN]البعيدة، [PRE]في [CN]البحر [CN]الأسود وسواحل آسيا [CN]الصغرى [CO]و [CN]المدن [CN]الفينيقية [PRE]على [CN]الساحل [CN]السوري.