

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

تحليل إحصائي لدعم التعرف الآلي على الكتابة العربية

A Statistical Analysis for Arabic Text Recognition Support

حسني المحتسب صبري محمود

جامعة الملك فهد للبترول والمعادن، الظهران

رامي قهوجي

جامعة برادفورد، برادفورد

مسار

- مقدمة
- طبيعة الدراسة
- من فوائد الدراسة
- تعريفات
- التحليل الإحصائي
- بعض الجداول
- بعض النتائج
- خاتمة

مقدمة

• زيادة اهتمام الباحثين بالحواسبة العربية وتطبيقاتها المختلفة ومنها

• التعرف الآلي (التعرف الضوئي) على الكتابة العربية
• كتابة مطبوعة أو بخط اليد

• فهرسة واسترجاع المعلومات من المخطوطات بناء على المحتوى
• التدقيق الإملائي

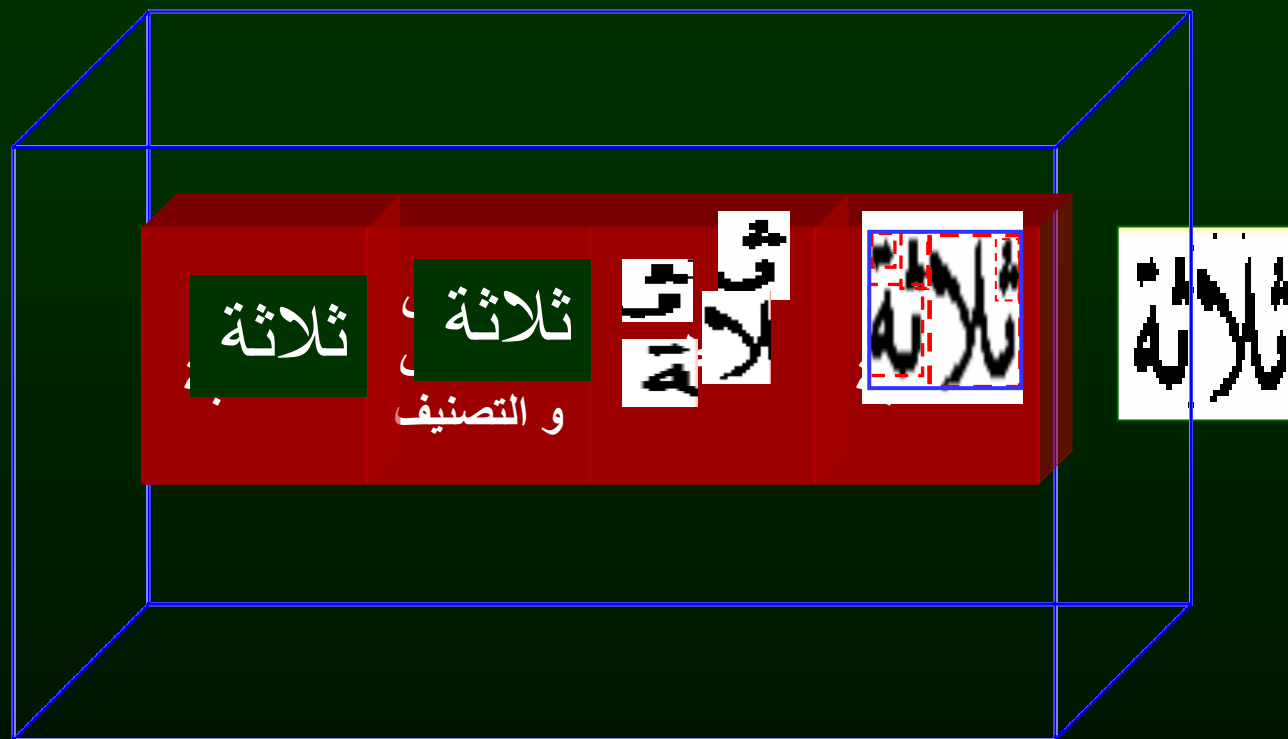
التعرف الضوئي على الكتابة



التعرف الضوئي على الكتابة



التعرف الضوئي على الكتابة



طبيعة الدراسة

- دراسة إحصائية لأعداد ظهور حروف ومقاطع الكلمات في اللغة العربية
 - تكرار كل حرف من الحروف العربية في كل مقطع من المقاطع
 - تكرار الحرف والحرف الذي يليه في المقاطع المختلفة لكل الحروف
 - إحصائيات استخدام الحروف والمقاطع ونسبة استخدام كل منها في حالات الاستخدام المختلفة في اللغة العربية
- الدراسة على كتابي صحيح البخاري ومسلم

من فوائد الدراسة

- المساعدة في عملية التعرف الآلي على الكتابة العربية
 - احتمالات تتابع الحروف العربية
 - دون الحاجة إلى تحليل عينات التدريب في كل مرة
- عملية تصحيح الأخطاء بعد عملية التعرف
- التعرف على صورة الكلمات في فهرسة واسترجاع المعلومات من المخطوطات بناء على المحتوى

تعريفات

• مقطع

– مجموعة الحروف المتصلة ببعضها في كلمة واحدة

• "مقطع": مقطع واحد بطول أربعة حروف

• "مجموعة": مقطعين

– "مجمو" بطول أربعة حروف

– "عة" بطول حرفين

• "الاستحسان": أربعة مقاطع

– "ا" (١)

– "لا" (٢)

– "ستحسا" (٥)

– "ن" (١)

تعريفات

• منفصل

- غير مرتبط بما بعده، وغير مرتبط بما قبله
- مقطعا لوحده
- مثال: "أن" مقطعين
 - كل مقطع منها حرفا منفصلا
 - "أ" حرف منفصل
 - "ن" حرف منفصل أيضا
- مثال: "لأنه" مقطعين
 - "أ" يمثل حرفا متصلا
 - "ن" في المقطع الثاني مثلا على الحرف المتصل

تعريفات

• متصل

- مرتبطا بما بعده، أو بما قبله، أو بكليهما
- مثال: "مرتبطا" كل حرف من حروفها مثالا على الحروف المتصلة

• أول حرف

- الحرف الأول في المقطع
 - قد يكون غير متصل بما بعده فيكون مقطعا لوحدته بطول ١
 - مثال: "أول"
 - ثلاثة مقاطع
 - كل حرف منها هو أول حرف في المقطع (وهو آخر حرف في المقطع
 - مثال: "متحابين"
 - مقطعين
 - أول حرف في المقطع الأول "م"
 - أول حرف في المقطع الثاني "ب"

تعريفات

• آخر حرف

- الحرف الأخير في المقطع
- غير متصل بما بعده
- إذا لم يكن متصلا بما قبله فيكون مقطعا لو حده بطول ١
- مثال ذلك كلمة "متحابين"
- آخر حرف في المقطع الأول هو "ا" وهو متصل
- آخر حرف في المقطع الثاني هو "ن" وهو أيضا متصل

• طول المقطع

• عدد الحروف التي تكون المقطع

- مثال: "فرحتنا" من مقطعين
- "فر" بطول ٢
- "حتنا" بطول ٤

التحليل الإحصائي

- نص كتابي صحيح البخاري وصحيح مسلم
- مثال على النصوص العربية المتعلقة بالحديث الشريف
- تعداد المقاطع في الكتابين ٢٢١٧١٧٨ مع التكرار
- ١٨١٧٠ مقطعا بعد حذف المكرر
- الحروف والرموز في الكتابين ٤٤٠٥٣١٨
- عدد الكلمات ١٠٩٥٢٧٤ مع التكرار
- ٥٠٣٦٧ كلمة بعد حذف المكرر.

الأكثر تكرارا من مجموع ١٨١٧٠ مقطعا

(جزء من الجدول)

مقطع	نسبة %	مقطع	نسبة %	مقطع	نسبة %	مقطع	نسبة %	مقطع	نسبة %	مقطع	نسبة %
ا	9.62	و	5.04	أ	4.69	ل	3.07	الله	2.64	ر	2.53
بن	2.4	حد	1.96	ن	1.88	قا	1.88	عن	1.79	ثنا	1.43
(1.34)	1.34	إ	1.34	عليه	1.05	صلى	0.99	سلم	0.98
د	0.85	ة	0.79	من	0.78	لا	0.73	سو	0.71	ب	0.7
بي	0.67	م	0.67	[0.66]	0.66	ما	0.64	با	0.62
ذ	0.61	عبد	0.59	في	0.56	ء	0.54	نا	0.51	ه	0.49
لى	0.47	فقا	0.45	ي	0.44	بو	0.43	لأ	0.42	لنبي	0.4
ز	0.39	1	0.38	خبر	0.36	2	0.34	كا	0.33	على	0.33
عمر	0.32	لر	0.31	لك	0.29	ثم	0.28	ني	0.28	محمد	0.28
ت	0.28	فأ	0.27	لو	0.27	له	0.26	س	0.26	ير	0.26
يا	0.26	مر	0.26	بو	0.26	يد	0.25	ثني	0.25	5	0.24

تكرار المقاطع حسب طولها ونسبة ظهورها

طول المقطع	عدد	تكرار	نسبة %
١	٥٧	٩٤٢٠٩٧	42.49079
٢	٥٣٧	٦٣٨٣٥٩	28.79149
٣	٣١٩٢	٤١٨٨٨٥	18.89270
٤	٦٧٧٨	١٦٩٩٣٨	7.664604
٥	٤٨٩٦	٣٩٠٢٦	1.760164
٦	٢١١٨	٧٣٦١	0.331998
٧	٥٢٢	١٣٢٢	0.059625
٨	٦٢	١٦٢	0.007306
٩	٨	٢٨	0.001262
المجموع	18170	2217178	100

كل المقاطع بطول ٩ ونسبتها

نسبة %	مقطع	نسبة %	مقطع	نسبة %	مقطع
٠,٠٠٠٠٩	قسطنطينية	٠,٠٠٠٠٩	فليقطعهما	٠,٠٠٠٨٦	لمستضعفين
٠,٠٠٠٠٥	يستبينهما	٠,٠٠٠٠٥	فاتقتانهم	٠,٠٠٠٠٥	ليثينهما
		٠,٠٠٠٠٥	فلايتمسها	٠,٠٠٠٠٥	مستقبايها

كل المقاطع بطول ٨ ونسبتها

مقطع	نسبة %	مقطع	نسبة %	مقطع	نسبة %	مقطع	نسبة %	مقطع	نسبة %
مستخلفكم	٠,٠٠٠٠٥	مستخلفين	٠,٠٠٠٠٥	ستطعمتها	٠,٠٠٠٠٥	فليحملها	٠,٠٠٠٠٥	للمطفين	٠,٠٠٠٠٥
ملتصقتين	٠,٠٠٠٠٥	فسقبيتهما	٠,٠٠٠٠٥	لمجنبتين	٠,٠٠٠٠٥	لمستقبلة	٠,٠٠٠٠٥	لمطمئنين	٠,٠٠٠٠٥
فهيجتهما	٠,٠٠٠٠٥	بحبيبتيه	٠,٠٠٠٠٥	فأيلبها	٠,٠٠٠٠٥	تستطيعها	٠,٠٠٠٠٥	فتستقبله	٠,٠٠٠٠٥
فجعلتهما	٠,٠٠٠٠٥	تخفيفهما	٠,٠٠٠٠٥	يستحسنها	٠,٠٠٠٠٥	فغمستهما	٠,٠٠٠٠٥	يستكملها	٠,٠٠٠٠٥
فقبضتهما	٠,٠٠٠٠٥	لقطعتكما	٠,٠٠٠٠٥	ليبتليكم	٠,٠٠٠٠٥	فليمسكها	٠,٠٠٠٠٥	فليمتهما	٠,٠٠٠٠٥
يستعملها	٠,٠٠٠٠٥	فمنعنيها	٠,٠٠٠٠٥	فجمعتهما	٠,٠٠٠٠٥	لمقتسمين	٠,٠٠٠٠٩	فلبستهما	٠,٠٠٠٠٩
يستقبلكم	٠,٠٠٠٠٩	فقطعتهما	٠,٠٠٠٠٩	ليخلعهما	٠,٠٠٠٠٩	فأيجلها	٠,٠٠٠٠٩	فليستجر	٠,٠٠٠٠٩
فليستغفر	٠,٠٠٠٠٩	فليستنشق	٠,٠٠٠٠٩	فليتحله	٠,٠٠٠٠٩	فليمحها	٠,٠٠٠٠٩	يستلمها	٠,٠٠٠٠٩
فتبنيعه	٠,٠٠٠٠٩	لمتخلفين	٠,٠٠٠٠٩	بمغنيتين	٠,٠٠٠٠٩	مستضعفين	٠,٠٠٠٠٩	ستقبلها	٠,٠٠٠٠٩
تسليمتين	٠,٠٠٠٠٩	تصليينهما	٠,٠٠٠٠٩	لخايفتين	٠,٠٠٠٠٩	لجهنمين	٠,٠٠٠٠٩	لينعلها	٠,٠٠٠٠٩
لمتشبهين	٠,٠٠٠٠٩	فليستنثر	٠,٠٠٠٠٩	تستعملني	٠,٠٠٠٠٩	تستعينها	٠,٠٠٠٠٩	للمحلقين	٠,٠٠٠٠٩
لمتكافين	٠,٠٠٠٠٩	فليستغف	٠,٠٠٠٠٩	بسيفيهما	٠,٠٠٠٠٩	فنفتخهما	٠,٠٠٠٠٩	ليقطعها	٠,٠٠٠٠٩
فليطلقها	٠,٠٠٠٠٩	للمسلمين	٠,٠٠٠٠٩		٠,٠٠٠٠٩				

تكرار الحروف في مقاطعها (جزء من الجدول)

نسبة %	مجموع	تكراره في المقطع الذي بطول									الحرف	
		9	8	7	6	5	4	3	2	1		
0.27	11900										11900	ء
0.11	5042					3	6	22	1459	3552		آ
2.87	126607				2	38	430	2057	20129	103951		أ
0.07	3102					3	55	995	1257	792		ؤ
0.82	35998						1	43	6277	29677		إ
0.23	9941			2	4	71	298	1289	1374	6903		ئ
11.57	509567	6	75	475	2251	10664	29471	113405	139833	213387		ا
5.31	233858			8	44	1199	6578	23260	46774	155995		ب
1.24	54682	2	1	40	625	5350	11377	8824	10867	17596		ة
2.23	98364			1	203	2205	11861	15133	33473	35488		ت
1.48	65373			4	37	77	1540	3554	7909	52252		ث
0.95	41678				18	58	968	3320	10529	26785		ج
2.45	107951			2	50	288	1729	4373	28380	73129		ح
0.69	30406				4	31	248	1186	5884	23053		خ
3.03	133415			7	76	1401	16728	30893	65359	18951		د
0.66	28974			7	18	102	426	3202	11690	13529		ذ
3.91	172062		8	20	420	2392	15879	44166	53029	56148		ر

تكرار الحرف الأول في المقطع مع طول المقطع (جزء من الجدول)

مجموع	تكراره كأول حرف في مقطع بطول								١ (منفصل)	الحرف الأول
	9	8	7	6	5	4	3	٢		
11900									11900	ء
3552									3552	آ
103951									103951	أ
792									792	ؤ
29677									29677	إ
6903					45	160	3602	2912	184	ئ
213387									213387	ا
155995		10	29	483	2456	6286	24975	106216	15540	ب
17596									17596	ة
35488		15	60	471	1885	4191	13122	9610	6134	ت
52252				13	121	743	39515	9600	2260	ث
26785			4	64	1160	2993	5296	14302	2966	ج
73129			7	67	624	3682	14474	51019	3256	ح
23053			22	77	636	1487	11693	8895	243	خ
18951									18951	د
13529									13529	ذ
56148									56148	ر
8625									8625	ز
81833		3	110	328	2897	13131	34475	25053	5836	س
15804			1	15	121	6015	4529	4793	330	ش

جداول وبيانات أخرى

- آخر حرف: حرف متصل في آخر المقطع أو منفصل
- تكرار ظهور الحرف والذي يليه في المقطع
 - الأول مع الثاني
 - الثاني مع الثالث
 - الثالث مع الرابع
 - الرابع مع الخامس
 - الخامس مع السادس
 - السادس مع السابع
 - السابع مع الثامن
- الاتصال بالباحثين

تكرار ظهور الحرف والذي يليه في المقطع: الحرف السابع – الحرف الثامن

الحرف السابع	ا	ة	ر	ف	ق	م	ن	ه	ي	منفصل	مجموع
ئ										2	2
ا										475	475
ب										8	8
ة										40	40
ث										1	1
ش			4							4	4
ح										2	2
د										7	7
ذ										7	7
ر										20	20
س								1		1	1
ش				2						2	2
س										2	2
م										9	9
ف			2	6					19	6	33
ق										1	1
ك					4					10	14
ل		1						3		17	21
م	48		2							98	148
ن								1	6	394	401
ه	27					4				96	127
و										73	73
ي										54	114

تكرار ظهور الحرف والذي يليه في المقطع: الحرف الثامن – الحرف التاسع

الحرف الثامن	ا	ة	م	ن	منفصل	مجموع
ا					75	75
ة					1	1
ر					8	8
فا					6	6
ق					2	2
م			4		4	8
ن					56	56
ه	2		1		6	9
ي		2		19	4	25

تكرار ظهور الحرف التاسع

مجموع	الحرف التاسع
2	ة
6	ا
1	م
19	ن

بعض النتائج

- حرف الألف المنفصل يمثل حوالي ١٠% من تعداد المقاطع
- يليه حرف الواو المنفصل
- فالألف المهموزة
- فاللام
- ثم المقطع الثلاثي "الله"
- العديد من المقاطع بطول ٢ و ٣ و ٤ و ٥ وتظهر أكثر شيوعاً من مقاطع أخرى بأطوال أقل

بعض النتائج

- أكثر من ٩٠% من المقاطع أطوالها ثلاثة حروف فأقل
- ٩٨% من المقاطع أطوالها أربعة حروف فأقل
- المقاطع بطول ٩ قليلة وتبلغ ٨ مقاطع أغلبها ظهر مرة واحدة
- "م" ظهر منفصلا كما ظهر في كل المقاطع ذات الأطوال المختلفة من ١ إلى ٩
- الحرف إذا لم يظهر في مقطع بطول معين، فإنه حتما لن يظهر في مقطع أطول من ذلك المقطع

بعض النتائج

- الحروف التي لا تتصل من اليسار لا تظهر كحرف أول إلا في منفصلة (طول المقطع ١)
- ء، ا، أ، و، إ، اة، د، ذ، ر، ز، و، ي
- في المقطع المنفصل من حرف واحد:
 - الحرف هو الأول وهو الأخير
- الهاء تأتي متصلة بما قبلها عشرة أضعاف ظهورها منفصلة
- تأتي التاء المربوطة منفصلة أكثر من ضعفي ظهورها متصلة بما قبلها

خاتمة

- يمكن استخدام هذه الجداول في تطبيقات الحوسبة العربية المختلفة مثل
 - التعرف الآلي (التعرف الضوئي) على الكتابة العربية
 - كتابة مطبوعة أو بخط اليد
 - فهرسة واسترجاع المعلومات من المخطوطات بناء على المحتوى
 - التدقيق الإملائي
 -
- يمكن لمن يرغب من الباحثين الحصول على هذه الجداول لاستخدامها الاتصال عن طريق البريد الالكتروني:

muhtaseb@kfupm.edu.sa

وشكرا

سبحانك اللهم وبحمدك أشهد أن لا إله
إلا أنت أستغفرك وأتوب إليك