بسم الله الرحمن الرحيم

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 بالحوسبة 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

الحصول على الصورة

التحويل الثنائي

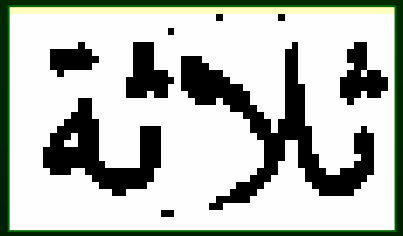
ما قبل المعالجة

التقطيع

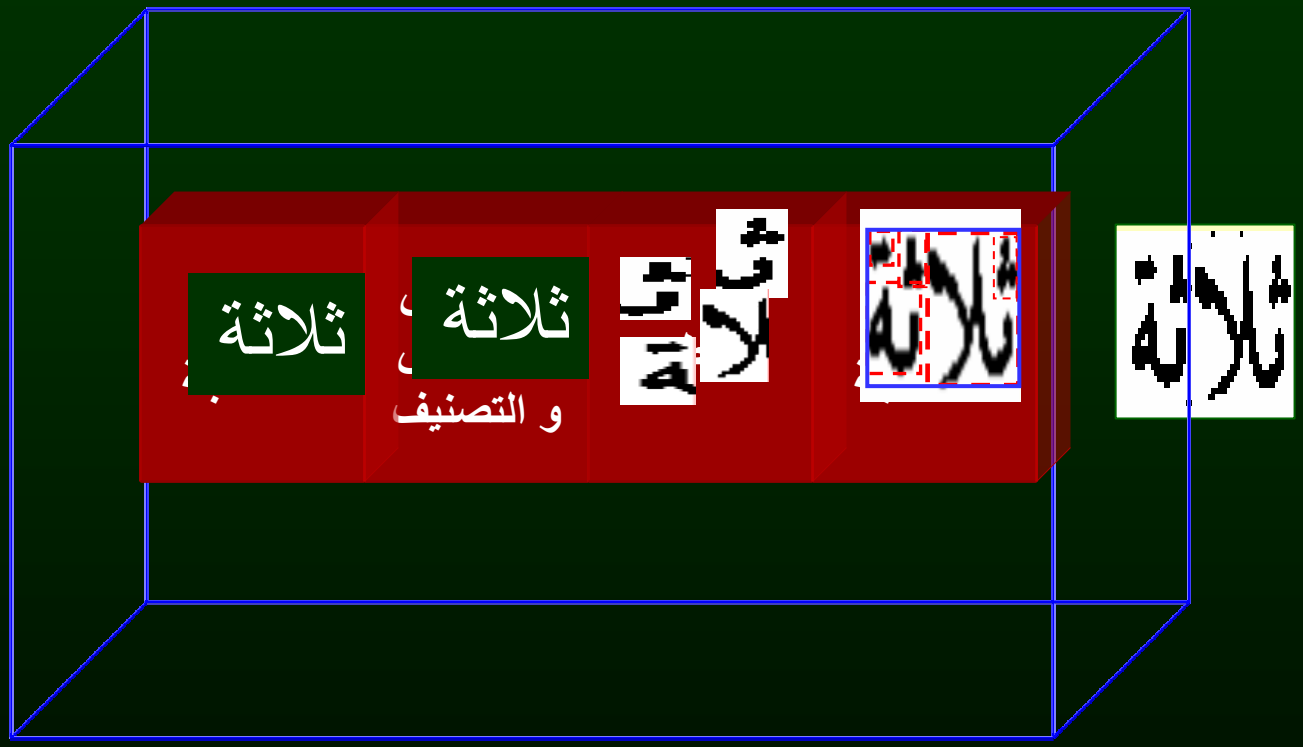
استخلاص النصوص

ما بعد المعالجة

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 مثالا 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 لوحده 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 لوحده 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	لله	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	يو	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

مقطع	نسبة%	مقطع	نسبة%	مقطع	نسبة%
لمستضعفين	٠,٠٠٠٨٦	فليقطعهما	٠,٠٠٠٠٠٩	قسطنطينية	٠,٠٠٠٠٠٩
ليثنينهما	٠,٠٠٠٠٠٥	فلتقتلنهم	٠,٠٠٠٠٠٥	يستبينهما	٠,٠٠٠٠٠٥
مستقبليها	٠,٠٠٠٠٠٥	فليلتمسها	٠,٠٠٠٠٠٥		

كل المقاطع بطول ٨ ونسبتها

مقطع	نسبة %	مقطع	نسبة %	مقطع	نسبة %	مقطع	نسبة %	مقطع	نسبة %
مستخلفكم	٠,٠٠٠٠٠٥	مستخلفين	٠,٠٠٠٠٠٥	ستطعمتها	٠,٠٠٠٠٠٥	فليحملها	٠,٠٠٠٠٠٥	للمطففين	٠,٠٠٠٠٠٥
ملتصقتين	٠,٠٠٠٠٠٥	فسقيتهما	٠,٠٠٠٠٠٥	لمجنبتين	٠,٠٠٠٠٠٥	لمستقبلة	٠,٠٠٠٠٠٥	لمطمئنين	٠,٠٠٠٠٠٥
فهيجتهما	٠,٠٠٠٠٠٥	بحبيبتيه	٠,٠٠٠٠٠٥	فليحلبها	٠,٠٠٠٠٠٥	تستطيعها	٠,٠٠٠٠٠٥	فتستقبله	٠,٠٠٠٠٠٥
فجعلتهما	٠,٠٠٠٠٠٥	تخفيفهما	٠,٠٠٠٠٠٥	يستحسنها	٠,٠٠٠٠٠٥	فغمستهما	٠,٠٠٠٠٠٥	يستكملها	٠,٠٠٠٠٠٥
فقبضتهما	٠,٠٠٠٠٠٥	لقطعتكما	٠,٠٠٠٠٠٥	ليبتليكم	٠,٠٠٠٠٠٥	فليمسكها	٠,٠٠٠٠٠٥	فليمتهما	٠,٠٠٠٠٠٥
يستعملها	٠,٠٠٠٠٠٥	فمنعنيها	٠,٠٠٠٠٠٥	فجمعتهما	٠,٠٠٠٠٠٥	لمقتسمين	٠,٠٠٠٠٠٩	فلبستهما	٠,٠٠٠٠٠٩
يستقبلكم	٠,٠٠٠٠٠٩	ففظعتهما	٠,٠٠٠٠٠٩	ليخلعهما	٠,٠٠٠٠٠٩	فليجعلها	٠,٠٠٠٠٠٩	فليستجمر	٠,٠٠٠٠٠٩
فليستغفر	٠,٠٠٠٠٠٩	فليستنشق	٠,٠٠٠٠٠٩	فليتحلله	٠,٠٠٠٠٠٩	فليمنحها	٠,٠٠٠٠٠٩	يستلمهما	٠,٠٠٠٠٠٩
فتبيعنيه	٠,٠٠٠٠٠٩	لمتخلفين	٠,٠٠٠٠٠٩	بمغنيتين	٠,٠٠٠٠٠٩	مستضعفين	٠,٠٠٠٠٠٩	ستقبلهما	٠,٠٠٠٠٠٩
تسليمتين	٠,٠٠٠٠٠٩	تصلينهما	٠,٠٠٠١٤	لخليفتين	٠,٠٠٠١٤	لجهنميين	٠,٠٠٠١٤	لينعلهما	٠,٠٠٠١٤
لمتشبهين	٠,٠٠٠١٨	فليستتر	٠,٠٠٠١٨	تستعملني	٠,٠٠٠١٨	تستعينها	٠,٠٠٠١٨	للمحلقين	٠,٠٠٠٢٧
لمتكلفين	٠,٠٠٠٢٧	فليستعفف	٠,٠٠٠٢٧	بسيفيهما	٠,٠٠٠٣٢	فنفختهما	٠,٠٠٠٣٦	ليقطعهما	٠,٠٠٠٣٦
فليطلقها	٠,٠٠٠٤٥	للمسلمين	٠,٠٠٠٨٦						

تكرار الحروف في مقاطعها (جزء من الجدول)

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

تكرار الحرف الأول في المقطع مع طول المقطع

(جزء من الجدول)

الحرف الأول	تكراره كأول حرف في مقطع بطول									مجموع
	١ (منفصل)	٢	3	4	5	6	7	8	9	
ء	11900									11900
آ	3552									3552
أ	103951									103951
ؤ	792									792
إ	29677									29677
ئ	184	2912	3602	160	45					6903
ا	213387									213387
ب	15540	106216	24975	6286	2456	483	29	10		155995
ة	17596									17596
ت	6134	9610	13122	4191	1885	471	60	15		35488
ث	2260	9600	39515	743	121	13				52252
ج	2966	14302	5296	2993	1160	64	4			26785
ح	3256	51019	14474	3682	624	67	7			73129
خ	243	8895	11693	1487	636	77	22			23053
د	18951									18951
ذ	13529									13529
ر	56148									56148
ز	8625									8625
س	5836	25053	34475	13131	2897	328	110	3		81833
ش	330	4793	4529	6015	131	15	1			15804

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
ت										1	1
ث			4								4
ح										2	2
د										7	7
ذ										7	7
ر										20	20
س								1			1
ش					2						2
ض										2	2
ع										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									4	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
ف					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

الحرف التاسع	مجموع
ة	2
ا	6
م	1
ن	19

بعض النتائج

- حرف الألف المنفصل يمثل حوالي ١٠% من تعداد المقاطع
- يليه حرف الواو المنفصل
- فالألف المهموزة
- فاللام
- ثم المقطع الثلاثي "لله
- العديد من المقاطع بطول ٢ و٣ و٤ و٥ وتظهر أكثر شيوعا 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 ذاك المقطع

بعض النتائج

- الحروف التي لا تتصل من اليسار لا تظهر كحرف أول إلا في منفصلة (طول المقطع ١)
  - ء، ا، أ، ؤ، إ، ا،ة، د، ذ، ر، ز، و، ى
- 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