استخدام ذخائر النصوص لاستخلاص المصطلحات المتخصصة

عبدالمحسن بن عبيد الثبيتي
مركز المعلومات
الرياض – المملكة العربية السعودية

ما لذي تقدمه هذه الورقه

- طريقة إحصائية بسيطة لاستخلاص المصطلحات المتخصصة ذات المفردة الواحدة اعتمادا على ذخائر النصوص.
- تقيس الطريقة مدى ابتعاد استخدام المفردة في مدونة المجالات المتخصصة عنه في مدونة المجالات العامة أو اللغة اليومية.

مالفائدة؟

- يمثل استخلاص المصطلحات المتخصصة آليا الخطوة الأولى نحو التعامل مع النتاج المعرفي المتخصص كقاعدة وأساس لعمليات أخرى مهمة مثل
- بناء المعاجم والقواميس المتخصصة.
- المساعدة في عمليات الترجمة الآلية للنصوص العلمية والفنية.
- استرجاع وفهرسة النصوص العلمية والفنية.
- اختيار الخصائص اللغوية الملائمة لعمليات التصنيف الآلي للنصوص.
- بناء شبكة العلاقات بين المفاهيم في المجالات المتخصصة.

الأساليب المتبعة لاستخلاص المصطلحات

- الاكتشاف الالي للتركيب النحوي للمصطلحات.

  تعاني أدوات معالجة اللغة العربية آليا من شح وندرة في مثل هذا الدراسات وتوافر هذه الأدوات.

- الطرق الإحصائية لاستخلاص المصطلحات في النصوص المتخصصة
- في كلتا الطريقتين يبرز استخدام ذخائر النصوص (المدونات) كطريقة تثبت صلاحية الطرق المستخدمة لاستخلاص المصطلح العلمي آليا كون المدونات تعطي الدليل على التراكيب النحوية الأكثر استخداما التي يظهر فيها المصطلح أو حتى التوزيع الإحصائي لمفرداته.

الفرضية

- تفترض هذه الدراسة اختلاف التوزيع الإحصائي للمفردات العربية في المجالات المتخصصة عنه في المجالات العامة أو اللغة العربية اليومية وذلك اعتمادا على الدراسات السابقة الخاصة بمدونات المجالات المتخصصة في اللغة الانجليزية

توصيف المدونات
(مدونة المجالات العامة)

- تمثل اللغة اليومية وهي مكونة من حوالي ١,١ مليون كلمة مختلفة بهيكلة مشابهة لبنية مدونة (Collins-Birmingham) للنصوص الانجليزية.
- تم جمع نصوص المدونة من شبكة الانترنت من مواقع سعوديه ولكتاب سعوديين في الأغلب موزعة كالتالي:
  - الصحف اليومية (٥٧%).
  - المجلات الأسبوعية (٢١%).
  - الكتب (١٥%).
  - المطويات التعريفية (٦%).
  - المواقع الشخصية (١%).

توصيف المدونات
(مدونة المجالات المتخصصة)

- تمثل المجالات المتخصصة يقارب حجمها ٥٨ ألف كلمة تتكون مدونة المجالات المتخصصة من مدونتين فرعيتين.
  - مدونة الذكاء الصناعي (٤٢%) وقد غطت مواضيع متعددة وخصوصا موضوعي الشبكات العصبية والأنظمة الخبيرة.
  - مدونة الفيزياء (٥٨%). غطت مواضيع فيزياء الكم والفيزياء الذرية والجزيئية.

الكلمات المائة الأكثر تكرارا في مدونة اللغة اليومية.

	المفردة	النسبة
١	و، في، من، على، أن، إلى، التي، لا، ما، عن	١٩,١
٢	هذا، الله، الذي، هذه، أو، مع، ان، ذلك، هو، كان	٣,٤
٣	قد، لم، كل، بين، بن، بعد، هي، إن، كما، الى	١,٩
٤	خلال، العربية، حتى، عبد، له، قال، حيث، المملكة، كانت، غير	١,٣
٥	لكن، بعض، محمد، عليه، السعودية، إلا، قبل، به، بها ، فيها	١,١
٦	يكون، ثم، هناك، لها، أي، تلك، عام، العمل، فيه، إذا	٠,٩
٧	مجلس، أنه، العام، ليس، الدكتور، منها، العالم، رئيس، بل، الذين	٠,٨
٨	عند، مثل، أكثر، يمكن، اليوم، الرياض، تكون، فقد، الملك، الناس	٠,٧
٩	العزيز، عدد، أما، عدم، السعودي، التعليم، عبدالله، عليها، تم، بما	٠,٦
١٠	عبدالعزيز، أخرى، جميع، عندما، فإن، دون، المجتمع، بأن، مما، يجب	٠,٥
	المجموع	٣٠,٣

الكلمات المائة الأكثر تكرارا في مدونة الذكاء الصناعي

	المفردة	النسبة
١	و، في، من، على، التي، أن، إلى، أو، هذه، الذكاء	١٩,٥
٢	عن، هي، هذا، هو، ما، لا، مع، الشبكة، كل، الذي	٤,٨
٣	الصناعي، ذلك، الإنسان، حيث، النظام، قد، الاصطناعي، الخبير، المعرفة، بين	٢,٨
٤	أي، يمكن، أنظمة، لكن، عملية، كما، يتم، الشبكات، بها، كان	٢,٢
٥	بعد، غير، تكون، نظام، التعرف، ثم، المعلومات، يكون، الأنظمة، مثل	١,٨
٦	طريق، البحث، طبقة، مجموعة، الخبيرة، علم، قاعدة، إذا، تقوم، لها	١,٥
٧	الأوزان، الحاسوب، فيها، الخبراء، بعض، معالجة، الطريقة، العصبية، مجال، يقوم	١,٤
٨	فإن، يجب، أنه، حتى، الصوتية، العصبونية، الكلام، المشاكل، ذات، لم	١,٢
٩	التعليم، الحاسب، العمليات، المشكلة، منها، أكثر، إن، بواسطة، حل، مرحلة	١,٢
١٠	هناك، المعالجة، كانت، إلا، الآلية، البشري، الكمبيوتر، المجال، المنطق، النظم	١,١
	المجموع	٣٧,٥

الكلمات المائة الأكثر تكرارا في مدونة الفيزياء.

	المفردة	النسبة
١	و، في،من، على، أن، هذه، إلى، هذا، التي، عن	١٦,٣
٢	ان، لا، النظرية، يمكن، مع، هو، الضوء، هي، النسبية، نظرية	٤,٥
٣	ما، الذي، الكم، بين، سرعة، كل، بالنسبة، أو، الزمن، حيث	٣,٠
٤	ذلك، كان، يكون، الطاقة، لكن، الكون، كما، بعد، أي، الكمومية	٢,٤
٥	الى، فإن، عام، النواة، حول، تكون، الإلكترون، طاقة، غير، الجسيمات	١,٩
٦	لم، عند، الجسيم، جسيمات، حالة، كانت، جسيم، خلال، سوف، قد	١,٦
٧	الأرض، المراقب، الذرة، ذرة، بسرعة، مثل، العالم، الفيزياء، عندما، الهيدروجين	١,٤
٨	نموذج، العلماء، القوة، الموجة، الفا، الخاصة، العامة، هنا، أنه، تفسير	١,٢
٩	الموجية، حركة، مبدأ، معادلة، نفس، الحالة، حتى، هناك، المادة، بشكل	١,١
١٠	ذات، ميكانيكا، احد، الثانية، الجسم، الطبيعة، كتلة، الأبعاد، الذرية، الفضاء	١,٠
	المجموع	٣٤,٤

معامل الغرابة

- يقيس مدى ابتعاد معدل استخدام المفردة في المجال المتخصص عن استخدامها في المجالات العامة أو اللغة اليومية

$$\text{معامل الغرابة} = \frac{\text{معدل استخدام المفردة في المجال المتخصص}}{\text{معدل استخدام المفردة في المجالات العامة}}$$

استخلاص المصطلحات

- يخضع للاعتبارات التالية:
- أن الكلمات التي يطبق عليها معامل الغرابة تنتمي للمجموعة المفتوحة من الكلمات.
- أن تكرار الكلمات مهم جدا في اعتبار الكلمة مصطلحا مقبولا أو متعارفا عليه في المجال المتخصص فهناك علاقة مباشره بين التكرار وقبول الكلمة في المجال (Quirk, 1985).
- أن تكون قيمة معامل الغرابة عالية للكلمات التي تراعي الفقرتين السابقتين أعلاه.

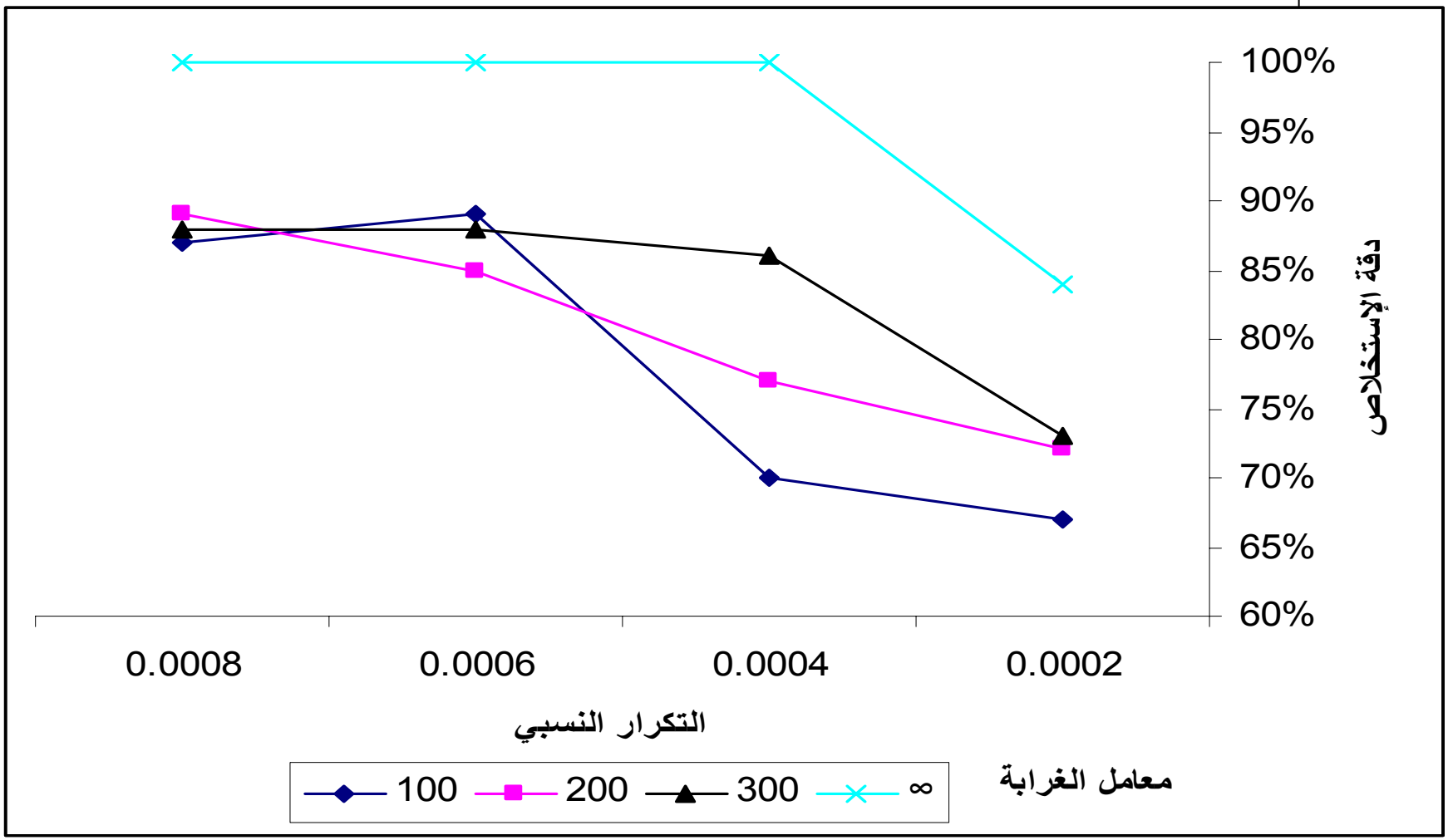
الكلمات العشرين ذات معامل الغرابة الأعلى في مدونة الذكاء الصناعي.

الكلمة	معامل الغرابة	الكلمة	معامل الغرابة
العصبونية/العصبية	∞	درجة	١٠٣٨
الاصطناعي	∞	الاوزان	٥٠٧
الضبابي	∞	محاكاة	٤١٥
المدخلات	∞	الأذرع	٣١١
النظم	∞	البرمجة	٢٩٤
استنتاجات	∞	مسموعة	٢٧٧
الانحيازات	∞	شبكه	٢٧٧
الخوارزميات	∞	صورة	٢١٦
مصفوفة	∞	طبقة	٢٠٦
الخبيرة	١٦٢٦	الأنماط	٢٠٦

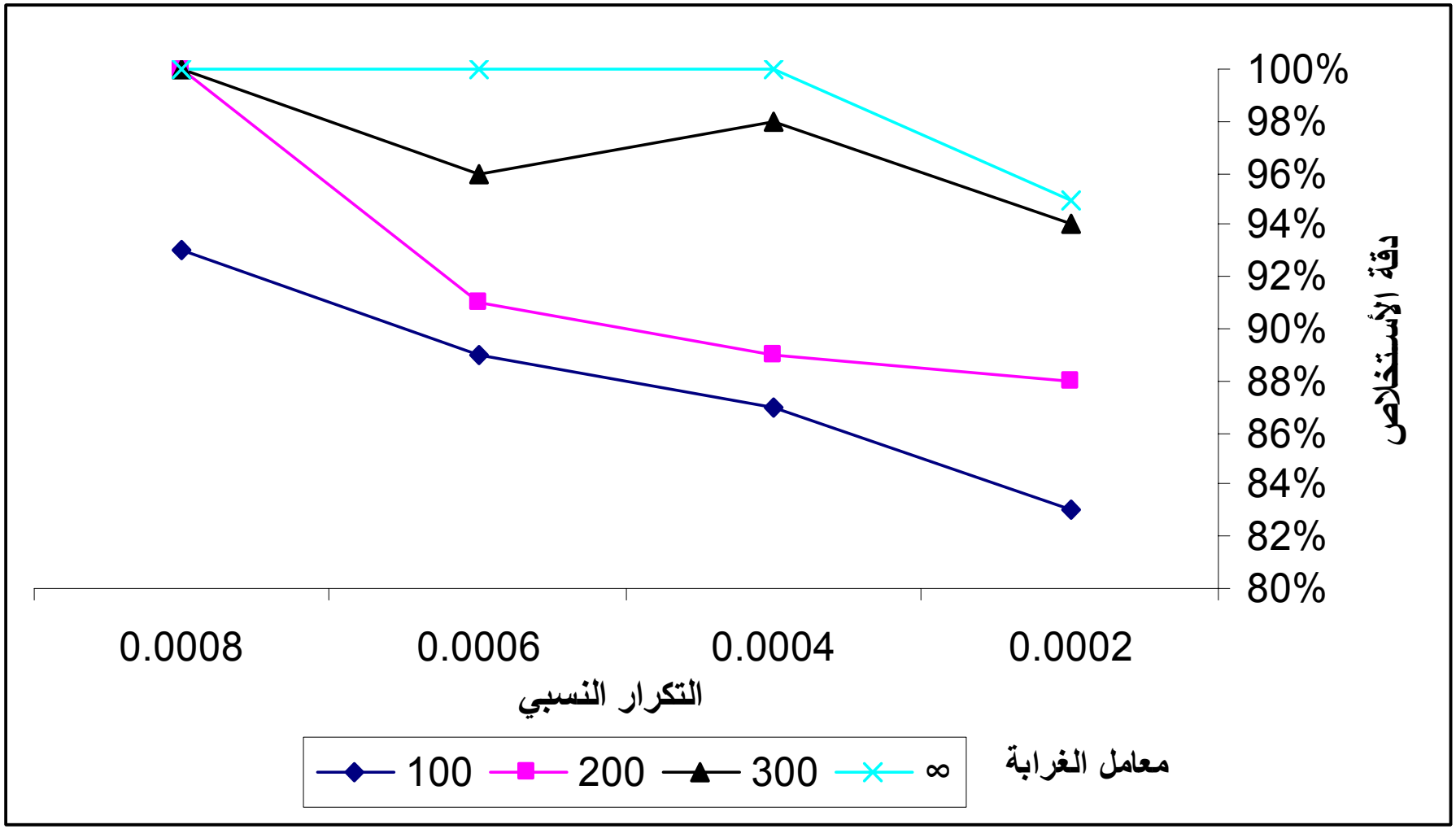
الكلمات العشرين ذات معامل الغرابة الأعلى في مدونة الفيزياء.

الكلمة	معامل الغرابة	الكلمة	معامل الغرابة
الكمومية	∞	بوهر	∞
الإلكترون	∞	ديراك	∞
جسيمات	∞	اينشتاين	∞
الموجية	∞	الزاوي	∞
ميكانيكا	∞	الكواركات	∞
الفوتونات	∞	شرودنغر	∞
الذرات	∞	المدار	∞
الأوتار	∞	الارتياب	∞
الزمكان	∞	الجزيئات	∞
بلانك	∞	الحيود	∞

دقة استخلاص المصطلحات لقيم مختلفة من التكرار النسبي ومعامل الغرابة لمدونة الذكاء الصناعي.



دقة استخلاص المصطلحات لقيم مختلفة من التكرار النسبي ومعامل الغرابة لمدونة الفيزياء.



الخلاصه

- إن دقة الاستخلاص تتناسب طرديا مع معامل الغرابة ومع التكرار النسبي ولكن ازدياد قيم معامل الغرابة والتكرار النسبي يؤدي إلى قلة الكلمات التي ممن الممكن تحديدها كمصطلحات.
- إن حجم المدونة يؤثر على دقة الاستخلاص فكلما ازداد الحجم ازدادت الدقة.
- إن حجم مدونة المجالات العامة ومدى عموميته لهذه المجالات يؤثر على قيمة معامل الغرابة.