# خطة المعالجة اللغوية الآلية في محرك البحث العربي «باع»

أ.د. تغريد السيد عنبر
كلية الألسن جامعة عين شمس

# وضع المحتوى الرقمي

- يشغل المحتوى الرقمي ملايين المواقع على الإنترنيت
- تضم هذه المواقع مئات المليارات من الوثائق والمستندات
- هذا الكم يتدفق عليه يوميا ملايين المستندات الجديدة

# سـؤال هـام

- يعتمد الجميع (أفراد ومؤسسات بأنواعها المختلفة) على الإنتريت للحصول على المعلومات

- هذه المعلومات يعتبرها الجميع مؤشرات رأي عام أو أساس لاتخاذ قرارات هامة

- السؤال: هل يمكن لأي شخص أن يبحر منفردا في هذا الخضم المتلاطم ذي الحجم المتفجر
  هل يستطيع فعلا أن يحصل على معلومات دقيقة حول موضوع بعينه؟
  كم من الوقت يمكن أن يستغرق البحث ليحقق هدفه

# إجابة بديهية

- لا بد من الاستعانة بآلية تيسر العمل
- إنها محرك بحث تتوافر فيه صفات خاصة:

الذكاء
سرعة الأداء
شمولية النتائج
دقة المعطيات
حيادية العرض

# كيف يعمل محرك البحث

## 1- مرحلة الفهرسة

- يعتمد محرك البحث على الزاحف الذي يجوب الانترنيت بحثا عن المواقع
- ثم يفرغ وثائقها في مستودع المستندات
- تراجع المستندات آليا وتجمع بينات عنها بحيث يتم التفريق بين مستند وآخر بما يشبه البصمة لكل مستند
- قد تتم تصفيات أخرى مثل إسقاط المفردات العلاقية (لماذا؟)
- تتلو ذلك عملية الفهرسة حيث تتحول النصوص إلى قوائم من المفردات مع كل مفردة مجموعة بياناتها
- ترتب أبجديا
- تجمع المفردات المتطابقة معا مكونة مدخلا واحدا

# كيف يعمل محرك البحث
## 2- الاســتعــلام

- يكتب المستخدم الاستعلام الخاص به
- في وحدة الاستعلام يتم تحليل الاستعلام (بطرق مختلفة وفقا لنظام محرك البحث)
- الهدف هو تحديد المفردات الأساسية للاستعلام
- هذه المفردات هي التي يتم استخراج ما يطابقها في قائمة مفردات المفهرس

# كيف يعمل محرك البحث

## 3- ترتيب نتائج البحث

- من المتوقع أن يكون عدد المستندات التي أسفر عنها البحث كبيرة جدا
- عادة يتعذر على المستخدم الاطلاع عليها كلها
- وحدة ترتيب نتائج البحث تتضمن منظومة بها المعايير الأساسية للترتيب:

المستند الأكثر قربا من الاستعلام فالأقل....

تجنب تكرار نفس المستند

حيادية العرض بعيدا عن أغراض خاصة

# محركات البحث العالمية

- توجد محركات بحث عالمية مثل جوجل و بنج وغيرها
- تشترك جميعا في أداء متميز ودقة بحث مشهود لها
- السؤال التقليدي: لماذا إذن نبذل وقتا وجهدا في تطوير محرك بحث عربي؟
- الإجابة ببساطة: هذه المحركات على جودتها لا تلائم اللغة العربية

# نموذج استعلام

- **Dream**
- The **dream**
- My **dream**
- **Dreams**
- Our **dreams**
- To my **dream**
- To their **dreams**

- **الكلمة المشتركة في جميع السياقات هي dream(s) وبالتالي فهي جميعا مخزنة في الفهرس تحت مدخل واحد**

# اللغة العربية ومحركات البحث التقليدية

- حلم
- الحلم
- حلمي
- حلمنا
- لحلمي
- أحلامنا
- وأحلامهم

**دلاليا المفردة الأساسية واحدة**

**شكليا** (وهو أساس ترتيب مداخل الفهرس) نحن بصدد عشرة أشكال كتابية مختلفة أي 10 مداخل

# اللغة العربية ومحركات البحث التقليدية (2)

- ما تتمتع به اللغة العربية من مرونة تتمثل أساسا في العدد الضخم من الزوائد الذي يمكن أن يلحق بالكلمة، يمثل مشكلة ضخمة في عملية البحث داخل المحركات العالمية.
- إن الزوائد تجعل للكلمة العربية الواحدة أكثر من ألف شكل كتابي
- بالتالي تخزن المفردة الواحدة تحت ألف مدخل مختلف
- كيف نتوقع أن تكون نتيجة البحث

# الهدف من تطويرمحرك بحث عربي

- تجميع المفردات ذات الدلالة الواحدة تحت مدخل واحد في عملية الفهرسة بصرف النظر عن شكلها الكتابي

- ميزة ذلك: تحقيق شمولية نتائج البحث
  تقليل حجم الفرس
  تقليل زمن الفهرسة
  تسهيل تعديل الفهرس وفقا للمستجدات
  منع تكرار إيراد نفس الوثيقة أكثر من مرة

# كيف نحقق هذا الهدف

- بناء معجم خاص ترتب مداخله أبجديا وفقا لمفاهيم دلالية ذات بنية شكلية تجمع الصور الكتابية المختلفة للكلمة
- برمجية ذكية معقدة نسبيا لمعالجة المفردات بحيث تتحول كل مفردة إلى البنية الشكلية في المعجم
- تستخدم هذه البرمجية لمعالجة النصوص المطلوب فهرستها
- وتستخدم أيضا لمعالجة مفردات الاستعلام
- الربط بين ناتج المعالجة وبين مداخل المعجم
- وبالتالي ترجيح أحد احتمالات المعالجة
- استدعاء حصري للمداخل المقابلة في الفهرس المخزن

# متطلبات هذه المعالجة الآلية

- تحديث دوري لمعجم النظام بحيث يغطي ما لا يقل عن 98% من مفردات أي نص
- تحديث دوري لمعجم الأعلام بنفس الشرط السابق
- تطوير البرمجية بهدف التحديث وتصحيح ما يكتشف من أخطاء أو نواقص
- اعتماد التحديث بشكل رئيسي على معطيات فريق الذخيرة اللغوية وبالذات المعطيات الإحصائية وفقا لما يطلبه فريق المعالجة اللغوية
- التعاون بين الفريقين متكامل ويترتب عليه تذليل عقبات كثيرة أمام العمل العلمي

# المطلوب من فريق المعالجة اللغوية

- التعاون الوثيق مع مطوري وحدات الفهرسة والاستعلام وترتيب النتائج
- إمدادهم بعينات كاملة من المعالجة اللغوية لأنها ستكون أساس التطوير
- استقبال تعليقاتهم والتفاهم لتعديل ما يشكل عقبة أمام تطوير وحدات المحرك
- التوقيت المحدد لأول عينة بعد ستة أشهر من بدء العمل في المعالجة اللغوية

# شكرا لحسن استماعكم

أ. د. تغريد عنبر
مقرر فريق المعالجة اللغوية