علم البيانات واللغة العربية

د. سلطان المجيول
أستاذ لغويات المدونة الحاسوبية المشارك
قسم اللغة العربية–جامعة الملك سعود

25 ذوالحجة 1441هـ

مجالات معالجة اللغة الطبيعية

- تعلم الآلة (الموجه وغير الموجه) والتعلم العميق (تحويل الكلمات على متّجهات دلالية semantic vectors)
- التحليل الإحصائي

مجالات المعالجة

- ✓ استخراج المعلومات information extraction من نص غير مهيكل واستخراج الحقائق fact extraction
- ✓ استرداد المعلومات information retrieval من ملف أو من عدة ملفات
- ✓ تحليل المشاعر sentiment analysis
- ✓ توليد النصوص text generation
- ✓ تلخيص النصوص text summarization
- ✓ الترجمة الآلية machine learning
- ✓ إجابات السؤال question answering
- ✓ نمذجة الموضوع topic modeling
- ✓ تصنيف النصوص text classification
- ✓ ...

هندسة الأنظمة المتعددة الموزّعة

بناء نظام (أنظمة) معالجة البيانات الضخمة

معالجة البيانات باستخدام أدوات البيانات الضخمة المتصلة بشبكة الاتصال

أدوار المهنيّ للبيانات الضخمة

أدوار عالم البيانات

إيجاد واستخراج الأنماط الخفية ذات العلاقات الارتباطية من البيانات

تحليل البيانات من مصادر عديدة

التنبؤ بالاعتماد على الأنماط الإحصائية السابقة باستخدام الذكاء الاصطناعي وتعلم الآلة

طلب الحصول على البيانات وتحليلها ومعالجتها

اكتشاف رؤية واضحة (معلومة بيانية) من البيانات

تصميم التقارير وعرضها: منصة تفاعلية: رسوم بيانية، إلخ.

أدوار محلل البيانات

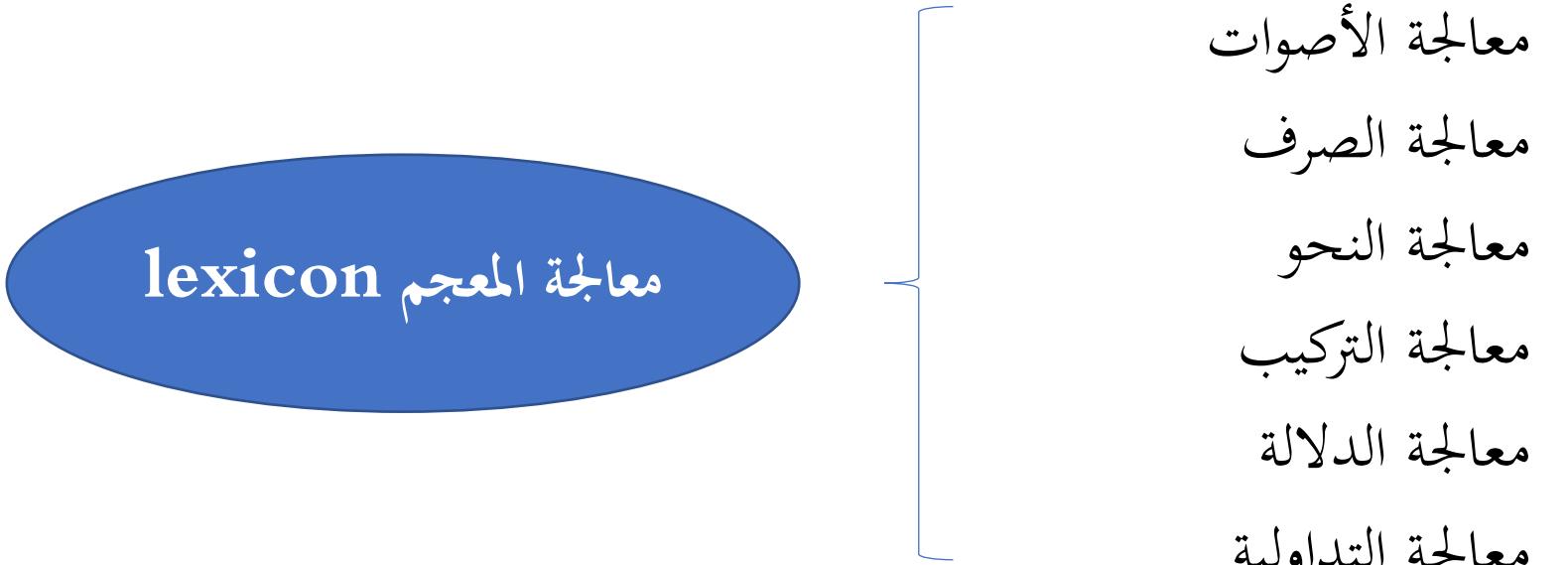
معالجة اللغة العربية الطبيعية

مستويات اللغة العربية

فصيح في مقابل عاميّة (خطأ)

فصاحات (نموذجيات standard levels) في مقابل اللهجات (صحيح)

معالجة اللغة العربية الطبيعية



ضع هذه المعالجات أمام المستويات اللهجية (ما حجم التحديات؟)

معالجة اللغة العربية الطبيعية

ننتقل إلى محيط آخر من المشكلات (المعتمد على القواعد rule-based في مقابل التعليم والتحشية labeling and annotation)

النطق

الكتابة

الصرف (الاشتقاقية والتصريفية)

النحو

الترجمة الآلية

معالجة اللغة العربية الطبيعية (الروايات)

عدة مكتبات معظمها تتطلب بيانات معلّمة labeled

Natural Language Toolkit

spaCy

learn NLP Toolkit

gensim

Pattern

Polyglot

مكتبة الجافا JDK واستعمال محللات وموسمات ومقطعات ستانفورد (المكتبات التي نحتاجها: nltk وpolyglot وrake-nltk). لمستخدمي ويندوز،

هناك حاجة إلى مكتبة PyICU ومكتبة pycld2

معالجة اللغة العربية الطبيعية (الروايات)

تحليل 24 رواية من منتديات غرام (تحليل 20 رواية من منتديات غرام (التكرارات TD-IDF وNamed Entity Recognition وFacts)

##########################

التحليل الدلالي الكامن LSA: تحليل المستند ثم موضوعه ثم الكلمات داخل كل موضوع

##########################

تحليل توزيع دركليه LDA: تحليل توزيع الموضوع في كل مستند، ثم يحدد الموضوع بناء على التوزيع، ثم تحلل توزيع الكلمات وتحدد بعد ذلك

##########################

نـهـايـة الـعـرض

متشوقون لقراءة الأعمال في تحليل الروايات السعودية