



**دورة معالجة اللغة العربية الطبيعية وهندسة الكلام
بالذكاء الاصطناعي: تحليلات النصوص والتعرف الآلي
على الكلام والتطبيقات الصوتية**



دورة معالجة اللغة العربية الطبيعية وهندسة الكلام بالذكاء الاصطناعي: تحليلات النصوص والتعرف الآلي على الكلام والتطبيقات الصوتية

مقدمة:

دورة معالجة اللغة العربية الطبيعية وهندسة الكلام بالذكاء الاصطناعي دورة تمتد 5 أيام لعلماء البيانات ومهندسي معالجة اللغة الطبيعية ومهندسي الكلام، وتنتهي بإعداد خطة حل لتحليلات النصوص والكلام العربي لجهة دراسة حالة. فالجهات تحتفظ بكميات كبيرة من الوثائق والمنشورات والمكالمات المسجلة بالعربية، غير أن الأدوات العامة تخفق في التعامل مع الصرف وتنوع اللهجات وغياب التشكيل، فتخرج التحليلات والنصوص المفرغة غير دقيقة، ويكتب المشاركون شيفرات Python ويدربون نماذج تعلم الآلة في عملهم اليومي، ويتعلمون عبر مختبرات تطبيقية على مدونات نصية وتسجيلات صوتية عربية. وتقدم كور كونسبت هذه الدورة في معالجة اللغة العربية وهندسة الكلام.

أهداف الدورة:

- بناء مسارات معالجة أولية للنص العربي تغطي توحيد الإملاء والتقطيع إلى وحدات والتحليل الصرفي والتشكيل الآلي
- تدريب نماذج تمييز اللهجات وتحليل المشاعر والتعرف على الكيانات المسماة في النصوص العربية وتقييمها
- اختيار تمثيلات الكلمات والجمل العربية وتطبيقها في البحث الدلالي والعنقدة واستكشاف الموضوعات
- ضبط نماذج التعرف الآلي على الكلام العربي وتقييمها بمعدل خطأ الكلمات ومعدل خطأ الحروف
- تجميع مسارات تحويل النص إلى كلام وتحليلات الكلام للمكالمات المسجلة والخدمات الصوتية الموجهة للجمهور
- إعداد خطة حل لتحليلات النصوص والكلام العربي تتضمن مستهدفات التقييم وتصميم النشر

الفئة المستهدفة:

- فرق علم البيانات المسؤولة عن بناء نماذج تصنيف النصوص وتحليلها على المحتوى العربي
- مهندسو تعلم الآلة المسؤولون عن تدريب النماذج اللغوية والصوتية وتغليفيها وتشغيلها
- فرق تقنيات الكلام المسؤولة عن التفريغ النصي والواجهات الصوتية ومسارات البيانات الصوتية
- فرق التحليلات المسؤولة عن تحويل ملاحظات العملاء والمكالمات المسجلة بالعربية إلى رؤى تشغيلية
- القادة الفنيون المسؤولون عن اختيار أدوات اللغة العربية وتحديد مستهدفات جودة النماذج

محاور الدورة:

اليوم 1: بنية اللغة العربية وأسس معالجة النصوص

- الكتابة العربية وتنوع الإملاء وقواعد توحيد الحروف
- الازدواجية اللغوية بين العربية الفصحى المعاصرة واللهجات المحكية
- أدوات التقطيع وتنظيف النصوص في مكتبة CAMeL Tools
- الصرف القائم على الجذر والوزن وفصل السوابق واللواحق
- تقييم الوضع الراهن لأصول النصوص والتسجيلات الصوتية العربية

اليوم 2: معماريات نماذج معالجة اللغة العربية وأساليب تقييمها

- التحليل الصرفي وفك الالتباس باستخدام CAMeL Tools
- التشكيل الآلي وإزالة التشكيل لخدمة المهام اللاحقة
- مقاطعات الوحدات الجزئية ومشفرات Transformer العربية المدربة مسبقاً
- الدقة والاستدعاء ومقياس F1 لمهام الوسم العربية
- أدلة التوسيم ومعامل الاتفاق بين الموسمين للمدونات العربية

اليوم 3: مختبرات تحليلات النصوص العربية

- تدريب مصنف تمييز اللهجات على عينات نصية إقليمية
- تحليل المشاعر على مستوى الجوانب لملاحظات العملاء العربية
- التعرف على الكيانات المسماة للأشخاص والأماكن والجهات في العربية
- تمثيلات الجمل العربية للبحث الدلالي والعنقدة
- استكشاف الموضوعات في الشكاوى وطلبات الخدمة العربية

اليوم 4: التعرف الآلي على الكلام العربي وتوليده ومخاطر التشغيل

- السمات الصوتية ونماذج CTC والانتباه للتعرف على الكلام
- التفريغ بنموذج Whisper وضبطه بإدئة اللغة للتسجيلات العربية
- تشخيص معدل خطأ الكلمات ومعدل خطأ الحروف
- الواجهة الأمامية لتحويل النص العربي إلى كلام بمدخلات مشكولة
- ضوابط التبديل اللغوي وضجيج الصوت الهاتفي وانحراف اللهجات

اليوم 5: بناء مختبري وخطة حل لتحليلات النصوص والكلام العربي

- بناء مختبري لمسار تفريغ المكالمات المسجلة وتحليل مشاعرها
- بناء مختبري لمسار استخراج الكيانات من طلبات الخدمات الحكومية
- تقرير قياس مختبري مقابل مستهدفات معدل خطأ الكلمات وF1
- تصميم تشغيل النماذج اللغوية ومراقبتها وإعادة تدريبها
- عرض خطة حل تحليلات النصوص والكلام العربي أمام لجنة المراجعة

المهارات التي ستكتسبها:

- توحيد النص العربي
- التحليل الصرفي
- تمييز اللهجات
- استخراج المشاعر والكيانات
- اختيار التمثيلات العربية
- ضبط نماذج التعرف على الكلام
- تحليل أخطاء التفريغ الصوتي
- تصميم مسارات تحويل النص إلى كلام

لماذا تحضر هذه الدورة:

- تسليم خطة حل لتحليلات النصوص والكلام العربي إلى قائد علم البيانات ومالك الخدمة مدعومة بنتائج القياس من المختبرات
- المفاضلة بين النماذج الجاهزة والنماذج المضبوطة للنص والكلام العربي بناء على معدلات خطأ مقيسة لا على وعود المزودين
- تجنب إعادة العمل المكلفة الناتجة عن نصوص غير موحدة ولهجات موهمة وصوت هاتفي مشوش قبل وصول النماذج إلى التشغيل
- تدريب فريق العمل على إجراءات التوحيد والتوسيم وتحليل الأخطاء التي طبقت في المختبرات

الخاتمة:

بعد العودة إلى العمل يسلم المشاركون خطة حل لتحليلات النصوص والكلام العربي إلى قائد علم البيانات ومالك الخدمة التي تدعمها. ويعتمدان عليها في إقرار حالات الاستخدام النصية والصوتية التي تنتقل إلى البناء، والنماذج التي تحتاج إلى ضبط، ومستهدفات الجودة التي يجب أن يبلغها كل نموذج قبل الإطلاق. وبعد تشغيل أول مسار على بيانات حية، تقارن الوحدة معدل خطأ الكلمات ونتائج F1 في التشغيل بنتائج المختبر، وتوثق المواضع التي احتاجت فيها تغطية اللهجات أو جودة الصوت أو التوسيم إلى تصحيح.

الأسئلة الشائعة:

ما الذي يحتاجه المشاركون قبل دورة معالجة اللغة العربية الطبيعية وهندسة الكلام؟

يحتاج المشاركون إلى إتقان كتابة شيفرات Python وفهم تعلم الآلة الموجه وخبرة في تدريب نموذج تصنيف واحد على الأقل. وتفيد قراءة العربية في تحليل الأخطاء. ولا تشترط خبرة سابقة في معالجة الكلام، إذ تعرض السمات الصوتية ونماذج التعرف قبل مختبرات الكلام.

بم تختلف دورة معالجة اللغة العربية الطبيعية وهندسة الكلام عن دورات نشر النماذج اللغوية الكبيرة؟

تبنى هذه الدورة نماذج مخصصة لمهام النص والكلام العربي: التقطيع والصرف وتمييز اللهجات والمشاعر والكيانات والتمثيلات والتعرف على الكلام وتوليده. أما دورات النماذج اللغوية الكبيرة فتركز على صياغة الأوامر والتوليد المعزز بالاسترجاع وضبط النماذج التوليدية العامة.

لماذا تحتاج معالجة اللغة العربية الطبيعية وهندسة الكلام إلى معالجة واعية باللهجات؟

لأن العربية المحكية والمكتوبة غير الرسمية تختلف عن الفصحى المعاصرة في المفردات والصرف والنطق، فتفقد النماذج المدربة على النص الرسمي دقتها في المكالمات والمنشورات. ويكشف تمييز اللهجات ومجموعات التقييم الخاصة بكل لهجة مواضع الإخفاق والحاجة إلى بيانات تدريب إضافية.

ماذا يعود به المشاركون من دورة معالجة اللغة العربية الطبيعية وهندسة الكلام؟

يعود المشاركون بخطة حل لتحليلات النصوص والكلام العربي لجهة دراسة حالة، ومسارات مختبرية عاملة للتفريغ وتحليل المشاعر واستخراج الكيانات، وتقرير قياس يعرض معدل خطأ الكلمات ونتائج F1 مقابل المستهدفات المتفق عليها.