



**البيانات الاصطناعية وإعداد بيانات تدريب الذكاء
الاصطناعي: دورة في رسم البيانات وجودة التوسيم
ومجموعات البيانات الحافظة للخصوصية**



البيانات الاصطناعية وإعداد بيانات تدريب الذكاء الاصطناعي: دورة في وسم البيانات وجودة التوسيم ومجموعات البيانات الحافظة للخصوصية

مقدمة:

البيانات الاصطناعية وإعداد بيانات تدريب الذكاء الاصطناعي، من وسم البيانات وجودة التوسيم إلى مجموعات البيانات الحافظة للخصوصية، موضوع هذه الدورة الممتدة 5 أيام لفرق هندسة تعلم الآلة وعمليات التوسيم وهندسة الخصوصية، وتنتهي بإعداد خطة إعداد بيانات التدريب لمنشأة حالة. تتعثر مشاريع نماذج كثيرة لأن الوسوم تتباين بين الموسمين، وتسلم جهات التوسيم دفعات غير متسقة، وتتعدى مشاركة السجلات الحساسة لأغراض التدريب. يعمل المرشحون في بناء مجموعات البيانات أو تنقيحها، وتضيف الدورة قياس جودة التوسيم وتوليد السجلات الاصطناعية واختبار إعادة التعرف على الهوية، عبر دراسات حالة على عينات توسيم حقيقية. وتقدم كور كونسبت هذه الدورة في البيانات الاصطناعية وإعداد بيانات التدريب.

أهداف الدورة:

- كتابة أدلة توسيم وتصنيفات وسوم يطبقها الموسمون باتساق على النصوص والصور والسجلات الجدولية
- قياس الاتفاق بين الموسمين بمعامل كوهين كابا ومعامل فليس كابا ومعامل كرييندورف ألفا ومعالجة الدرجات المنخفضة
- إدارة جهات التوسيم الخارجية وفرق الموسمين عبر المجموعات المرجعية وطوابير الفصل في الخلافات ومعايير قبول الدفعات
- توليد سجلات جدولية ونصية وسجلات تعليمات اصطناعية بالنماذج الإحصائية وشبكات GAN الشبكات التوليدية التنافسية وتوسيع البذور بالنماذج اللغوية الكبيرة
- اختيار مجموعات البيانات المرشحة لمخاطر إعادة التعرف على الهوية باستخدام k-anonymity و l-diversity وميزانيات الخصوصية التفاضلية
- إعداد خطة إعداد بيانات التدريب مع توثيق مجموعة البيانات جاهزة للتسليم إلى فرق النماذج

الفئة المستهدفة:

- فرق هندسة تعلم الآلة التي تجمع مجموعات التدريب والتقييم لنماذج التشغيل الفعلي
- فرق علم البيانات التي تضبط النماذج اللغوية وتحتاج إلى بيانات تعليمات أو بيانات تفضيل
- فرق عمليات التوسيم التي تدير مجموعات الموسمين أو تشرف على جهات توسيم خارجية
- مالكو منتجات البيانات المسؤولين عن نطاق مجموعة البيانات وميزانيات التوسيم وقرارات الإتاحة
- فرق هندسة الخصوصية التي تقيم مخاطر الإفصاح قبل خروج السجلات من بيئة مضبوطة

محاوّر الدورة:

اليوم 1: أساسيات بيانات التدريب ومراجعة الوضع الراهن لمجموعات البيانات

- احتياجات الوسوم في التعلم بالإشراف لمهام التصنيف والاستخراج
- تصميم تقسيم مجموعات التدريب والتحقق والاختبار المحجوب
- مجموعات التعليمات والتفضيل والتقييم لمشاريع النماذج اللغوية
- أثر ضوضاء الوسوم واختلال توازن الفئات على سلوك النموذج
- ورقة جرد مجموعات البيانات لمراجعة فجوات الوضع الراهن

اليوم 2: أدلة التوسيم وتصنيفات الوسوم ومقاييس الاتفاق

- نموذج دليل التوسيم مع قواعد القرار للحالات الحديثة
- تصميم تصنيف الوسوم بالتسلسل الهرمي ومعايير الاستبعاد
- حساب معامل كوهين كابا للمقارنة بين موسمين اثنين
- معامل فليس كابا ومعامل كريندورف ألفا لمجموعات الموسمين
- جولة توسيم تجريبية لمعايرة صياغة دليل التوسيم

اليوم 3: عمليات التوسيم وجهات التوسيم وتوليد السجلات الاصطناعية

- زرع المجموعات المرجعية وسير عمل طابور الفصل في الخلافات
- نطاق عمل جهة التوسيم وعينات قبول الدفعات
- التعلم النشط بأخذ العينات حسب عدم اليقين والاستعلام بلجنة النماذج
- المولدات الإحصائية وشبكات GAN لتوليد السجلات الجدولية
- توسيع بذور التعليمات بالنماذج اللغوية الكبيرة لبيانات الضبط الدقيق

اليوم 4: تقييم البيانات الاصطناعية ومخاطر إعادة التعرف والإتاحة الآمنة

- قياس دقة المحاكاة والمنفعة للسجلات الاصطناعية مقابل الحقيقية
- اختبار التدريب على الاصطناعي والتقييم على الحقيقي لنقل النموذج
- فحوص k -anonymity و l -diversity على تركيبات أشباه المعارف
- ميزانية إبسيلون في الخصوصية التفاضلية لإتاحة البيانات الاصطناعية
- اختبار هجمات استنتاج العضوية والربط قبل مشاركة البيانات

اليوم 5: دراسات الحالة وبناء خطة إعداد بيانات التدريب

- حالة روبوت محادثة لدعم عملاء التجزئة ومراجعة الدليل والاتفاق
- حالة الملاحظات السريية وتقييم السجلات الاصطناعية ومخاطر الإفصاح
- صياغة صحائف بيانات مجموعات البيانات لمدونة الحالة
- بنود اتفاقية مشاركة البيانات وفق نظام حماية البيانات الشخصية
- استكمال خطة إعداد بيانات التدريب ومراجعتها مع المشاركين

المهارات التي ستكتسبها:

- تأليف أدلة التوسيم
- تصميم تصنيفات الوسوم
- تحليل الاتفاق بين الموسمين
- الإشراف على جهات التوسيم
- أخذ العينات بالتعلم النشط
- توليد السجلات الاصطناعية
- اختبار مخاطر إعادة التعرف
- توثيق مجموعات البيانات

لماذا تحضر هذه الدورة:

- تسليم خطة إعداد بيانات التدريب إلى مالك النموذج ومسؤول منتج البيانات، بما يغطي التوسيم والتوليد وضوابط الإنتاج
- الحسم في متى تحل السجلات الاصطناعية محل البيانات الحقيقية أو تكملها أو تقتصر على اختبارها
- تفادي دورات إعادة التوسيم ورفض دفعات الموردين وحوادث الإفصاح الناتجة عن أدلة غامضة أو إتاحة غير آمنة
- توجيه الموسمين وقادة جهات التوسيم في استخدام الدليل وأهداف الاتفاق وممارسة الفصل في الخلافات

الخاتمة:

بعد العودة إلى العمل يسلم المشاركون مالك النموذج ومسؤول منتج البيانات خطة إعداد بيانات التدريب لمشروع قائم في تعلم الآلة أو النماذج اللغوية. وتعتمد الإدارة على الخطة في إقرار ميزانيات التوسيم، وتحديد أهداف الاتفاق لجهات التوسيم، واختيار المواضع التي تقبل فيها السجلات الاصطناعية، والتأكد من جواز مشاركة مجموعة البيانات. وبعد أول دفعة موسومة أو أول إتاحة لبيانات اصطناعية تقارن الوحدة درجات الاتفاق ونتائج نقل النموذج ونتائج اختبارات الإفصاح بما ورد في الخطة، ثم تعدل الأدلة وأساليب أخذ العينات وميزانيات الخصوصية.

الأسئلة الشائعة:

ما الذي يحتاجه المشارك قبل دورة البيانات الاصطناعية وإعداد بيانات تدريب الذكاء الاصطناعي؟

ينبغي أن يعمل المشارك فعلا مع مجموعات بيانات لتعلم الآلة أو النماذج اللغوية، وأن يتعامل بسهولة مع جداول البيانات أو دفاتر Python الأساسية. ويفيد إحضار دليل توسيم أو دفعة موسومة أو وصف مجموعة بيانات من مشروعه، بعد إخفاء ما يدل على أصحابها، في دراسات الحالة.

بم تختلف دورة البيانات الاصطناعية وإعداد بيانات تدريب الذكاء الاصطناعي عن دورات حوكمة البيانات أو جاهزيتها؟

تركز هذه الدورة على إنتاج مجموعة البيانات نفسها: كتابة الأدلة وقياس اتفاق الموسمين وإدارة جهات التوسيم وتوليد السجلات الاصطناعية واختبار مخاطر الإفصاح. أما دورات الحوكمة والجاهزية فتدقق السياسات والملكية والضوابط النظامية على مستوى أصول البيانات، ولا تمسها هذه الدورة إلا عند إتاحة مجموعة بيانات.

هل تغني البيانات الاصطناعية تماما عن السجلات الحقيقية في إعداد بيانات تدريب الذكاء الاصطناعي؟

نادرا. توفر السجلات الاصطناعية وسوما دقيقة وحالات نادرة وعينات أكثر أمانا للخصوصية، غير أن النماذج المدربة عليها وحدها كثيرا ما تضعف عند نقلها إلى الواقع. ويكشف دمج عينة حقيقية صغيرة واختبار النموذج على بيانات حقيقية محدوبة مدى صلاحية المجموعة الاصطناعية.

ماذا يعود به المشارك من دورة البيانات الاصطناعية وإعداد بيانات تدريب الذكاء الاصطناعي؟

يعود المشارك بخطة إعداد بيانات التدريب لمشروع واحد، تضم دليل توسيم وأهداف اتفاق ومعايير قبول لجهات التوسيم ونهجاً لتوليد البيانات الاصطناعية ونتائج اختبارات الإفصاح وصحيفة بيانات، ويستخدمها مالكو النماذج مباشرة لإقرار قرارات التوسيم والإتاحة.