🧩 المدخل المعرفي: لماذا نبرمج الكلمات؟

في السابق، كان يُنظر للحاسوب كأداة مكتبية ثانوية، أما اليوم فقد أصبح امتداداً للعمليات المعرفية والبحثية للأخصائي التربوي. إن الانتقال الرقمي غير "المنظار الاجتماعي" (Socioscope) الذي نراقب من خلاله السلوك البشري.

تعتبر عملية **توحيد النصوص** مرحلة جوهرية في معالجة اللغات الطبيعية (NLP). والهدف منها بسيط لكنه عميق: **تقليل التباين الإملائي** لضمان أن الكلمات المتشابهة دلالياً تُعامل ككيان واحد برمجياً.

💡 **مثال توضيحي:** الكلمات (أحمد، احمد، إحمد) يراها العقل البشري شخصاً واحداً، لكن لغة "بايثون" تراها 3 سلاسل نصية مختلفة. التوحيد يدمجها في صورة نمطية واحدة لتسهيل الإحصاء.

🛠️ خطة العمل التقنية (Unification Pipeline)

تتضمن عملية التوحيد عدة خطوات فنية تُطبق على النصوص الخام:

1. **توحيد الألف:** تحويل كل أشكال الهمزات (أ، إ، آ) إلى الألف العادية (ا).
2. **توحيد التاء المربوطة والهاء:** معالجة اللبس بين (مدرسة) و(مدرسه) بجعلها تنتهي بحرف موحد.
3. **توحيد الياء:** ضبط الفرق بين الياء المنقوطة (ي) والألف المقصورة (ى).
4. **إزالة التشكيل (Diacritics):** حذف الفتحة والضمة وغيرها لأنها تغير شكل الكلمة بالنسبة للحاسوب (مثل: عِلْم وعَلَم).
5. **إزالة التطويل (Tatweel):** حذف علامات المد التجميلية (مثل: مدررررسة).
6. **توحيد الأرقام:** تحويل الأرقام الهندية (١، ٢، ٣) إلى التنسيق الغربي (1، 2، 3) لتمكين بايثون من إجراء العمليات الحسابية وحساب المتوسطات.

🛠️ الأدوات التطبيقية: مكتبة PyArabic

لتبسيط هذه المهمة المعقدة، نستخدم مكتبة متخصصة تُدعى **PyArabic**. توفر هذه المكتبة دوالاً جاهزة تنفذ العمليات بكلمة واحدة:

- استدعاء المكتبة: import pyarabic.araby as araby.
- إزالة الحركات: باستخدام الدالة strip_tashkeel.
- توحيد الألف: باستخدام الدالة normalize_alef.

⚖️ التوحيد (Unification) vs التطبيع (Normalization)

يجب على الباحث التمييز بينهما بدقة:

- **التوحيد:** يركز حصراً على الرموز والحروف المتشابهة شكلاً (مثل الألفات).
- **التطبيع:** عملية أوسع تشمل تنظيف النص من الروابط، الإيموجي، والرموز الخاصة (موضوع المحاضرة القادمة).

🎓 الأثر على جودة البحوث التربوية

إن إهمال هذه المرحلة التقنية ليس مجرد خطأ برمجي، بل هو **خطأ منهجي** يؤدي إلى:

- **أخطاء في التكرارات:** قد تظهر كلمة (قلق) بتكرار منخفض لأن الباحث لم يجمع بين (القلق، وقلق، وبالقلق).

- **صعوبة البحث:** فقدان سجلات الطلاب بسبب أخطاء إملائية بسيطة في الأسماء.
- **ضعف النمذجة:** تتعامل خوارزميات الذكاء الاصطناعي مع الكلمة الواحدة كعدة مفاهيم، مما يضعف التنبؤ بسلوك الطالب.

📝 الخلاصة والتمرين

توحيد النصوص هو "عملية فلترة" تجعل البيانات الخام صالحة للعلم. هي المهارة التي تضمن أن لغة الأرقام تعبر بصدق عن لغة الحروف.

🚴 **تمرين عملي مقترح:** حاول كتابة 5 أسماء لزملائك بتنسيقات مختلفة (بهمزة وبدونها)، وتخيل كيف ستقوم دالة normalize_alef بتحويلها لشكل واحد.