الجمهورية العربية السورية

جامعة دمشق

كلية الهندسة المعلوماتية

اختصاص الذكاء الصنعي

مادة معالجة اللغات الطبيعية NLP

الفصل الأول 2025/2026

الوظيفة الأولى

الوظيفة الأولى: مسألتا تصنيف الأسئلة الطبية وترقيم النصوص العربية

Arabic Medical Question Classification and Punctuation Prediction

الهدف من الوظيفة والفائدة المرجوة للطالب من إنجازها

1. التعامل مع النصوص العربية.
2. فهم مسألة وضع علامات الترقيم الصحيحة في نص عربي.
3. فهم مسألة تصنيف الأسئلة الطبية
4. تحليل البيانات لفهم محتواها بدون قراءتها كلها.
5. تنظيف وتوحيد النصوص، ودراسة تأثير عمليات المعالجة المسبقة على دقة حل المسألة.
6. تمثيل النصوص باستخدام طرق مختلفة للتمثيل الشعاعي (Vectorization).
7. حل مسألة التصنيفات باستخدام خوارزميات التصنيف الآلي المختلفة (Classification).
8. حل مسألة ترقيم النص العربي باستخدام خوارزميات تنميط السلاسل المختلفة (Sequence Labeling).

ملاحظة: اعتمد على وضع الملاحظات على تحقيق الطلبات إن كان لشرح الأكواد التي حققت بها المطلوب أو ملاحظاتك على الخرج، واحرص أن تكون ملاحظاتك باللغة العربية الفصحى السليمة. أيضاً لا تنسى، حافظ على نظافة الكود لتسهيل قراءته (وبالتالي تسهيل حصولك على علامة جيدة 🙂).

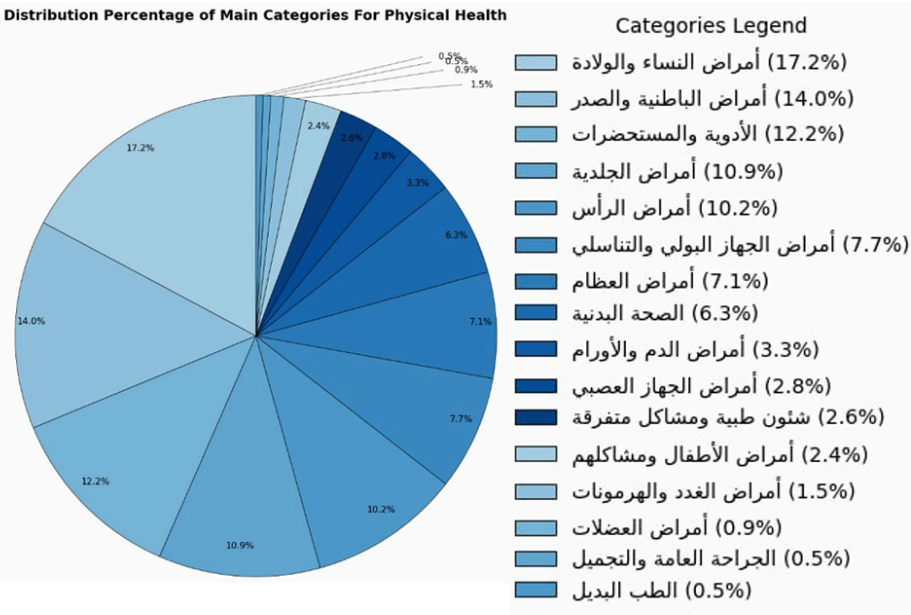
نص الوظيفة

المسألة الأولى: مسألة تصنيف الأسئلة الطبية

تقدم العديد من المواقع منصات للإجابة عن أسئلة المستخدمين في مجالات مختلفة، من هذه المواقع المواقع الطبية التي يجيب فيها أطباء مختصون عن استفسارات السائلين عبر مسائل طبية مختلفة.

يطرح السائل سؤاله عادة دون أن يعرف الاختصاص الدقيق الذي ينتمي له سؤاله، ويكون من مهمة القائمين على الصفحة توجيه السؤال للطبيب المختص.

لتسهيل العملية وتسريع هذه العملية سنقوم ببناء مصنف يقوم بتحديد الاختصاص الدقيق للمسألة المطروحة، بناءً على قاعدة بيانات "شفاء" الموجودة على الرابط التي تحتوي على أسئلة وأجوبة طبية مصنفة إلى عدة تصنيفات هرمية، استخدام عمود السؤال "Question" لبناء المصنف، واعتمد التصنيف الرئيسي فقط.



المطلوب:

قم ببناء ومقارنة عدة نماذج لتصنيف السؤال. ابدأ بتحليل البيانات EDA، وقم بكتابة مجموعة توابع لإجراء عمليات المعالجة المسبقة على النص وذلك لاختبار تأثيرها لاحقاً على المصنف،

قم ببناء مصنف للبيانات ذلك بثلاث طرق (يمكنك الإضافة على هذه الطرق في حال وجدت حلاً أفضل):

- الطريقة الأولى: اعتمد على حل المسألة باستخدام طرق التمثيل الشعاعي للنص (مثل TF-IDF) خوارزميات التعلم التلقائي التقليدية.
- الطريقة الثانية: باستخدام خوارزميات التعلم العميق مع تمثيل embedding مناسب، قم باختبار أثر استخدام تضمينات مدربة مسبقاً على الأداء،

- الطريقة الثالثة: باستخدام طرق نقل المعرفة (transfer learning) بالاعتماد على نماذج لغوية مدربة مسبقاً.

ملاحظات:

استخدم الـ notebook المرفق لتحميل البيانات، استخدم في بناء النموذج وضبط المعاملات مجموعتي التدريب والتحقيق، استخدم مجموعة الاختبار لعرض النتائج النهائية فقط، انتبه من تسريب البيانات.

المسألة الثانية: مسألة ترقيم نص عربي

تعتبر علامات الترقيم مؤشرات مهمة في النصوص لتحديد الجمل مكتملة المعنى، وتختلف علامات الترقيم بحسب المعنى المراد من الجملة فهناك الفاصلة "،" والنقطة "." وإشارة الاستفهام "؟"... الخ. وتأتي أهمية مسألة تحديد علامة الترقيم المناسبة من كونها مسألة جزئية من تطبيقات عديدة لها مثل تحليل وتركيب الكلام وكعملية معالجة لاحقة في مسألة الترجمة، حيث تساهم في جعل النص المستخرج من الكلام قابل للقراءة والفهم أكثر.

في هذه الجزء ستقوم بحل مسألة ترقيم النص المكتوب باللغة العربية، اعتماداً على مجموعة البيانات Arabic punctuation dataset وهي عبارة عن مجموعة ضخمة من النصوص بالعربية الفصحى والمخصَّصة لتدريب نماذج التعلم الآلي على تحديد حدود الجمل والتنبؤ بالترقيم.

المطلوب

أولاً، قم بتنزيل مجموعة البيانات من الرابط، حيث سنقوم باعتماد البيانات المتوفرة في الملف SSAC-UNPC.zip لتدريب المسألة.

سنعتمد علامات الترقيم التالية كهدف للمسألة: (Question marks، Colons، Commas، Semicolons، Exclamation mark، Full stops) وذلك بأشكالها المختلفة (قم بتوحيد أشكالها قبل البدء بالمسألة)

لإنشاء بيانات التدريب قم بإزالة العلامات من النص ليصبح لديك النص بدون علامات الترقيم.

بداية قم بإجراء تحليل لمجموعة البيانات، قم بكتابة مجموعة توابع للمعالجة المسبقة لدارسة أثرها على المصنفات، مستعيناً بشرح البيانات الموجود في الورقة البحثية.

ستتعامل مع المسألة كمسألة تنميط السلاسل sequence labeling. بحيث يكون الدخل سلسلة الكلمات الموافقة للنص ويكون الخرج سلسلة علامات الترقيم التي تلي كل كلمة، نلاحظ المثال في الصورة المرفقة أدناه أن نص الدخل "أكل الولد الخبز، وشرب الماء"، في البداية تم تقطيع النص إلى كلمات (والذي يفترض أن يكون خالٍ من علامات الترقيم)، وعندها يكون الخرج أيضا سلسلة تحوي أرقام علامة الترقيم الموافقة فنلاحظ أن الكلمة الأولى "أكل" توافق الرقم 0 والذي يعني عدم ورود علامة ترقيم، وكذلك للكلمة الثانية، بعد ذلك في الكلمة "الخبز" تقابل علامة الترقيم ذي الرقم 2 والذي يشير إلى الفاصلة "،". اذاً سيكون الدخل هو مصفوفة الكلمات (من نمط string) والخرج سلسلة معرفات علامات الترقيم التي تلي كل كلمة (من نمط معطيات رقمية). نسمي هذه المسألة عادة sequence-2-sequence.

باعتبار أن مجموعة البيانات غير منمطة ستعمل على تنميطها مستخدماً تابع تقطيع مناسب للنموذج الذي تختبره، ومعتمدا على آلية الاستخراج الموضحة في الجدول1 المرفق أدناه.

Original text	أكل الولد الخبز، وشرب الماء.
No punct. tokenized text	[أكل, الولد, الخبز, وشرب, الماء]
Encoded label	._‘__
Numerical label	[1,0,2,0,0]

جدول *1*: يوضح توصيف دخل وخرج مسألة ترقيم النص العربي *(Wazrah .et.al, 2025)*[21]

قم بمقارنة عدة طرق لبناء النماذج

- الطريقة الأولى باستخدام خوارزميات التعلم العميق مع تمثيل embedding مناسب، قم باختبار أثر استخدام تضمينات مدربة مسبقاً على الأداء،
- الطريقة الثانية باستخدام طرق نقل المعرفة (transfer learning) بالاعتماد على نماذج لغوية مدربة مسبقاً.

هام: اكتب تابعاً يأخذ بيانات اختبار (مجموعة نصوص -من عدة أسطر قد تكون طويلة- مع علامات ترقيم) ويقوم بإزالة علامات الترقيم منها ثم يحسب أداء أفضل نموذج على بيانات الاختبار مع طباعة النتائج بشكل واضح، هذه التابع مهم لتقييم مسألتك ولن تقبل المسألة بدونه.

استخدم المصنف في تصنيف الإجابات على الأسئلة الطبية (القسم الأول من الوظيفة) في قسم "أمراض الباطنية والصدر"، قم بحساب الأداء. هل تعتقد أن النموذج قد عمم جيداً على المجال الطبي؟

قم بإعادة تدريب النموذج مع إدخال البيانات الطبية من باقي التصنيفات على النموذج ثم احسب معايير التقييم لكل من الصفوف. هل تحسن الأداء؟

ملاحظات

لتنزيل البيانات على google colab مباشرة عبر رابط يمكنك الاستعانة بالأداء curlwget، اقرأ عنها لتتعلم كيف يمكنك استخدامها.

في حال كان حجم البيانات للتدريب كبيراً (يستهلك وقتاً طويلاً) قم بالتدريب على جزء من البيانات وفقط اعتمد البيانات كاملة لتدريب نموذجك الأفضل.

قم بتقسيم البيانات إلى مجموعتي تدريب وتحقيق، وسيتم تزويدك ببيانات اختبار عند تسليم الوظيفة لحساب الاداء عليها.

[1] Asma Ali Al Wazrah, Afrah Altamimi, Hawra Aljasim, Waad Alshammari, Rawan Al-Matham, Omar Elnashar, Mohamed Amin, and Abdulrahman AlOsaimy. 2025. Evaluation of Large Language Models on Arabic Punctuation Prediction. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 144–154, Abu Dhabi, UAE. Association for Computational Linguistics.

تعليمات العمل:

1- لا تنسى المرور بجميع مراحل مسألة معالجة اللغات الطبيعية.
2- يشترك الطلاب لحل الوظيفة في مجموعات مؤلفة من 5 طلاب فقط (لا أكثر ولا أقل) ولا جدال في العدد المسموح.
3- يقوم طالب واحد فقط من المجموعة بتسجيل أسماء الطلاب والمسألة التي سيعملون عليها على الرابط خلال يومين من موعد صدور الوظيفة. لن لن وألف لن يُسمح بتغيير أفراد المجموعة في الوظائف القادمة.
4- تحتاج الوظيفة لتنتهي في الوقت المحدد تعاون جميع أفراد المجموعة.
5- ينقسم طلاب كل مجموعة الى فريقين، الأول مؤلف من 2 طلاب لحل المسألة الأولى (مسألة التصنيف)، والفريق الثاني من 3 طلاب لحل المسألة الثانية (ترقيم النص العربي). يمكن لطلاب المجموعة الواحدة التعاون فيما بينهم على حل المسألتين لكن كل طالب يأخذ علامة الوظيفة على المسألة الخاصة به فقط.

تعليمات حل الوظيفة:

6- يتم تسليم notebook واضح ونظيف لكل مسألة من المسائل يحتوي حل المسألة بكامل خطواتها ابتداءً من تحليل مجموعة البيانات وانتهاءً بعرض جدول مقارنة النماذج الذي ستعتمد عليه لاختيار النموذج الأفضل الذي ستحصل على علامة دقته. في المسألة الثانية قم بتسليم notebook إضافي يحوي تابع لاختبار النموذج الأفضل الذي قمت باختياره، بحيث يكون اسم التابع evaluate_best_model ويمرر وسيط رابط مجموعة الاختبار ووسيط رابط النموذج المدرّب. يبدأ التابع بتعريف بنية النموذج ثم تحميل الأوزان من الرابط المرفق ثم تحميل مجموعة الاختبار وأخيراً تقييم النموذج وفق معايير تقييم نماذج التصنيف. ويعيد التابع قيمة دقة النموذج وفق المعيار Fscore الذي سيتم اعتماده للمفاضلة بين حلول الطلاب وإعطاء العلامة.
7- يجب أن يرفق كل notebook بخلايا نصية لشرح الكود قبل خلية الكود وخلية نصية أخرى لكتابة ملاحظات الطلاب على خرج خلية الكود بعدها، ولن يصحح أي كود مالم يرفق بهاتين الخليتين النصيتين.
8- اختر لأسماء النماذج وعمليات المعالجة أسماً واضحة ومعبرة، في حال عدم وجود أي إضافة قم بكتابة none.
9- قم بتخزين نتائج كل نموذج تبنيه والنتائج النهائية في كياني فهرس dictionary، يتألف كل منهما من 7 مفاتيح، قيمة كل مفتاح مصفوفة list تضاف الى الفهرس بعملية append. المفاتيح كالتالي:
   - number_step: يعبر عن رقم النموذج.
   - name_model: ويعبر عن خوارزمية التدريب.
   - features: ويعبر عن شكل الدخل أو سمات النموذج.
   - parameters_model: يحتوي أسماء البارامترات الفائقة hyperparameters المعدلة وقيمها الجديدة وفي حال عدم تعديل بارامترات النموذج تضاف كلمة default.
   - methods_preprocessing: يعبر عن خطوات المعالجة التي قمت بها قبل التدريب.
   - accuracy: يعبر عن مقدار دقة الاختبار.
   - F-score: ويعبر عن مقدار دقة التصنيف على بيانات الاختبار.
10- انتبه أنه هناك dictionary يضاف إليه نتائج طلبات التدريب، وهناك جدول نهائي للنماذج الأفضل من كل طلب منها.
11- لن تقبل الوظيفة بدون طباعة جدول مقارنة النماذج.
12- استخدم معايير F-score بالإضافة إلى الدقة Accuracy. وأيضاً قم بطباعة مصفوفة التعارضات confusion matrix لدراسة تضارب الصفوف مع بعضها

13- أي خطوة تقوم فيها (مثلاً نوع معين من المعالجة المسبقة) يجب عليك أن تثبت فائدته لعملية التدريب وذلك من خلال التجربة.

تعليمات تسليم الوظيفة، وآلية التقييم:

14- تاريخ التسليم: الأحد 21 كانون الأول 2024.
15- بالنسبة للمسألة الثانية يعطى الطلاب مجموعة اختبار منفصلة يوم المقابلة لتقييم النموذج الأفضل الذي قام ببنائه، ويحصل الطالب على جزء من العلامة على حل الوظيفة والجزء الآخر دقة النموذج على مجموعة الاختبار،
16- ستقوم بتسليم ملفي كود ورابط النموذج الأفضل المدرّب
17- قم بتسمية ملفي كل مسألة بأسماء الطلاب المشتركين بحلها، وأرفق أسماء الطلاب أيضا كتعليق ضمن الملفين باللغة العربية، وأي ملف لا يحقق هذا المعيار لن يُصحح.
18- تأكد من عدم طباعة كامل البيانات في ملف ipynb، تفاديا لتجاوز حجم الملف الحد المسموح 10M.
19- تأكد من تنفيذ جميع التعليمات البرمجية في كل خلية.
20- تأكد أن الملفين قابلين للفتح والقراءة بوضوح وأن حجمهم لا يتجاوز 10M.
21- قم برفع الملفين ipynb دون ضغط، على رابط تسليم الوظيفة.

تعليمات عامة:

22- لا ننصحك أبداً باستخدام أي نموذج لغوي لكتابة الكود عنك، وفي حال كان هناك أي مؤشر لعدم فهمك لكل جزئية صغيرة في الكود أو عدم قدرتك على تعديله، ستخسر علامته مباشرة بدون أي نقاش.
23- ننصح باستخدام google colab في تدريب النماذج والتعامل مع البيانات عوضاً عن التنفيذ المحلي على جهازك
24- **تحذير**: عند وجود أي تشابه بين وظيفتي مجموعتين ستخسر المجموعتان العلامة معًا دون مراجعتها (هذا خبر وليس تهديد عزيزي الطالب).

مع أمنياتنا لكم التوفيق

مدرسو المادة