الجمهورية العربية السورية
جامعة دمشق
كلية الهندسة المعلوماتية
اختصاص الذكاء الصنعي

مادة معالجة اللغات الطبيعية

الفصل الأول 2024/2025

الوظيفة الأولى

مسألة تحديد لهجة نص عربي

Arabic Dialect Identification

الهدف من الوظيفة والفائدة المرجوة للطالب من إنجازها

1. التعامل مع النصوص العربية متعددة اللهجات
2. فهم مسألة تحديد اللهجة وتحدياتها
3. تحليل البيانات لفهم محتواها بدون قرائتها كلها
4. تنظيف وتوحيد النص بهدف تجهيزه للقيام بعملية التصنيف
5. تصنيف النصوص باستخدام طرق مختلفة للتمثيل الشعاعي (Vectorization) وخوارزميات التصنيف الآلي (Classification).
6. دراسة تأثير عمليات المعالجة المسبقة على دقة عملية التصنيف.

ملاحظة: *مرفق مع الوظيفة ملف قالب الحل notebook، يطلب منك أن تقوم بملئ الخلايا وفق الطلبات المدرجة أدناه. في حال طلب منك تثبيت ملاحظاتك قم بإضافة خلية نصية واكتب ملاحظاتك باللغة العربية الفصحى بطريقة سليمة. لا تنسى أن تضع شرح بسيط قبل التوابع في حال كانت معقدة وتستلزم الشرح وحافظ على نظافة الكود لتسهيل قرائته (وبالتالي تسهيل حصولك على علامة جيدة 🙂).*

نص الوظيفة

يعتبر التعرف الآلي على لهجة نص/كلام من المسائل الهامة التي تساعد في تحسين العديد من تطبيقات معالجة اللغات الطبيعية، مثل تخصيص أكثر وأدق للترجمة والتعرف على الكلام وتحليل المشاعر وفقا للهجة المطلوبة وكذلك استرجاع الوثائق الأكثر صلة للاستعلام أيا تكن اللهجة، وغيرها الكثير.
وكما بقية اللغات، فإن التعرف الآلي على اللهجات العربية يواجه تحديات عديدة كون اللهجات لا تبنى على نظام كتابة موحد بل يعتمد على المنطقة وعلى عوامل تتعلق بالمتحدث مثل جنسه وتعليمه وعلى عوامل أخرى سياقية للكلام كأن يكون رسمياً أم لا أو أن يكون ضمن الريف أم المدينة، عدا عن استخدام المتحدث لمفردات من لهجات متعددة وحتى من لغات اخرى (مثل الكلمة "هاي" التي تعني "هذه" باللهجة العراقية, وتعني "مرحبا/Hi" باللغة الإنكليزية, وتعني أيضا باليابانية "نعم/はい". 😭) وهذا يزيد من ظهور مشكلة الغموض في كتابة المفردات، كما أن تدريب نموذج دقيق يعالج كل تلك العوامل يتطلب توفير معاجم خاصة ومحللات صرفية دقيقة والاعتماد على معايير تقييم قياسية قادرة على كشف كل تلك الفروقات، وأنه في حال استطعنا الحصول على نموذج دقيق إلا أن استخدامه ضمن تطبيقات الزمن الحقيقي ك chatbot يواجه تحديات الكلفة الزمنية.
في هذه الوظيفة ستقوم بحل مسألة التعرف الآلي على اللهجات العربية ومواجهة هذه التحديات, من خلال تحليل مجموعة البيانات المطلوبة وتنظيفها واعتماد تمثيل مناسب للنص وتدريب نماذج التعلم وقياس أدائها.

مهمة 01: تحميل البيانات

ستقوم باستخدام مجموعتي بيانات كما يلي:

- المجموعة "مَدَار MADAR" - يمكنك العودة للموقع الرسمي للبيانات عبر الرابط، ويمكنك تحميلها من الرابط - وهي عبارة عن مجموعتي بيانات تضم عبارات مكتوبة بلهجات عربية من 25 مدينة عربية، وقد تم بناؤها عن طريق كتابة كل عبارة من عبارات مجموعة البيانات BTE Corpus الى 25 لهجة (ولهذا تدعى هذه المجموعة بأنها a collection of parallel sentences). اقرأ توصيف البيانات بشكل جيد لمساعدتك على تشكيل وفهم البيانات بشكل جيد.
- المجموعة QADI، المتوفرة على الرابط وهي مجموعة بيانات تم جمعها تلقائيًا لتغريدات تنتمي إلى مجموعة واسعة من اللهجات العربية على مستوى الدول، تشمل 18 دولة مختلفة. تحتوي المجموعة الناتجة على 540 ألف تغريدة من 2,525 مستخدمًا موزعين بالتساوي عبر 18 دولة عربية.

والمطلوب:

مهمة 1-1: قم بتحميل مجموعتي البيانات على dataframes وطباعة أول 10 أسطر منها وعدد العينات الكلي وفي كل مجموعة (تدريب, تحقق واختبار). تحقق من وجود أسطر مكررة أو تحتوي خلايا فارغة أو تحتوي Nan.

مهمة 1-2: في مجموعة البيانات حقل يمثل نوع اللهجة وفقا للمدينة city (مثل اللهجات في دمشق, حلب, الرباط...), إلا أن هذه اللهجات يمكن أن تصنف أيضاً وفقا للمنطقة العامة التي تنتمي لها Region (مثل اللهجات في بلاد الشام, بلاد النيل, المغربية...), ووفقا لدولة اللهجة Country (مثل اللهجات في سورية, مصر, الجزائر...). المطلوب, قم بإضافة حقلين لكل من بقية معايير التصنيف وذلك بشكل آلي (يمكن الحصول على التصنيفات العامة لكل لهجة مدينة من الجدول المخزن في الرابط).

مهمة 1-3: قم بتوحيد مجموعتي البيانات بما تراه مناسباً، حافظ على تقسيم ال train/validation/test الأصلي.
بما أن بيانات QADI ليست مصنفة حسب المدينة يمكنك وضع القيمة NaN في الحقل المقابل للمدينة.
في الطلبات التالية في حال طلب منك أي عملية على مستوى المدينة قم بتطبيقها فقط على البيانات التي تكون فيها قيمة حقل المدينة ليست NaN.

مهمة 02: فهم وتحليل البيانات EDA

في هذه المرحلة سنقوم بمجموعة من العمليات بهدف فهم النصوص بصفوفها المختلفة لتتمكن من تنظيفها بشكل أفضل واقتراح سمات مفيدة عند التصنيف، ومن ثم تطبيق آلية التصنيف المناسبة لها. في هذه المرحلة ستعمل على قسم التدريب (train) فقط من البيانات التي لديك، وذلك لكي تتجنب تسريب البيانات (data leakage).

والمطلوب:

مهمة 2-1: احسب عدد النصوص في كل لهجة من أجل الثلاث أصناف, وارسم chart pie لكل منها, و اكتب ملاحظتك.

مهمة 2-2: قم بتقسيم النص إلى كلمات (tokenization) بطريقة مناسبة، ثم قم بطباعة أول 5 أسطر بعد التقسيم. واحسب عدد الكلمات الكلي (total tokens) وعدد الكلمات الفريدة الكلي (vocab) في بيانات التدريب.

مهمة 2-3: ارسم Word Cloud لكلمات مجموعة التدريب كاملًة، (انتبه أن الكلمات المكتوبة باللغة العربية تحتاج إلى إعدادات خاصة لعرضها، قم بالبحث عنها وتطبيقها في الكود الخاص بك).
اشرح في خلية نصية مختلفة استنتاجك من الرسم.

مهمة 2-4: قم بكتابة تابع يأخذ نصًا مقطعًا tokenized text (نص واحد وليس كل النصوص في ال data frame)
وأوجد كل من:

- أكثر n كلمة مكررة في النص.
- n من الكلمات الأقل تكرارً في النص (hapaxes).
- أكثر n ثنائيات كلمات مكررة في النص.
- أكثر n ثلاثيات كلمات مكررة في النص.
- أكثر n ثنائيات كلمات مهمة في النص (Collocations Bigram).
- أكثر n ثلاثيات كلمات مهمة في النص (Collocations Trigram).

اشرح الطريقة التي أوجدت بها الكلمات المهمة collocations، هل ثنائيات وثلاثيات الكلمات الأكثر تكرارًا هي نفسها الـ collocations؟
من أجل n=10 طبق التابع السابق على البيانات كاملة وعلى كل من المستويات الثلاثة، قم بطباعتها بطريقة مقروءة ومناسبة، لا تطبع كل كلمة على سطر بل اطبع خرج كل تابع على سطر واحد بفواصل مناسبة. واكتب ملاحظاتك.

مهمة 2-5: اكتب تابعاً يقوم بإيجاد الكلمات المميزة للهجة من اللهجات أي أنها فريدة في تلك اللهجة والتي تكرارها يتجازو عتبة معينة مرتبة بحسب تكرارها . نفذ هذا التابع على المستويات الثلاثة, واكتب ملاحظاتك.

مهمة 2-6: قم بتحليل إحصائي مدعوم بالمخططات البيانية اللازمة لدراسة تقارب أو اختلاف اللهجات التابعة لمدن City أو لبلدان Country أو لمناطق Region. (فكر باستخدام heatmap وطرق أخرى للتعبير عن مدى تقارب اللهجات) واكتب ملاحظاتك.

مهمة 2-7: قم بتحميل نماذج تضمين مدربة مسبقا وهي glove و fasttext و word2vec باللغة العربية واحسب نسبة الكلمات في بيانات التدريب والتي تعبتر خارج المفردات OOV لكل منها، ثبت ملاحظاتك واستنتاجاتك.

في حال وجود أمر مستعصي عليك حاول البحث والقراءة عنه من الانترنت ولا تنتظر نزول إلهام من التأمل في سقف الغرفة الخاص بك.

مهمة 03: تنظيف وتوحيد النص Text cleaning and normalization

قم بالبحث عن عمليات التنظيف والتوحيد الأنسب لهذا النوع من البيانات، اكتب بمجموعة من التوابع لتحقيق ذلك من خلال كتابة تابع واحد من أجل كل عملية، دخله نص واحد فقط (وليس كامل البيانات) وخرجه النص المعدّل, قم باختبار كل تابع وطباعة خرجه على أول 5 عينات تدريب

فقط, (تنبيه: قم بتحقيق هذه التوابع باستخدام التعابير النظامية فقط وليس باستبدال الحروف أو المرور عليها الواحد تلو الآخر). والمطلوب:

مهمة 3-1: اكتب مجموعة التوابع التي تراها مناسبة بحيث تكون شاملة،
أعرف أنك ستنسخها من مصدر ما 🙃 لكن المهم أن لا تقم بكتابة التابع لمجرد أنك وجدته على الإنترنت، لا تطبق أي تابع لا تراه مفيداً، وأي تابع تقوم بكتابته يجب عليك التحقق لاحقاً من فائدة تطبيقه أو عدمها.
انتبه أنه من الأسهل لك أن تفكر بما يجب أن تبقيه من النص بدلا من التفكير بما يجب أن تحذفه، وانتبه أيضًا إلى ما تعبر عنه الكلمات غير العربية الموجودة.
اشرح أمام كل تابع تطبقه رؤيتك لفائدته.

مهمة 3-2: بعد انتهائك من تجهيز توابع معالجة النصوص, قم ببناء تابع عام له وسيطين, الأول عبارة عن النص المراد معالجته من نمط string, والوسيط الثاني عبارة عن قائمة list يجب أن تحتوي على أسماء التوابع المراد تنفيذها على النص الممرر. يقوم هذا التابع العام بتنفيذ التوابع الممررة وفقا لترتيبها ضمن القائمة على النص, ثم يعيد النص بعد معالجته.

مهمة 3-3: طبق التابع العام الذي قمت ببنائه في مهمة 3-2 بوجود جميع عمليات المعالجة المطلوبة على مجموعة التدريب كاملًة، واحسب عدد الكلمات الكلي وعدد الكلمات الفريدة وقارنها مع عدد الكلمات الكلية والفريدة قبل التنظيف. اكتب ملاحظتك.

مهمة 3-4: هل يوجد خلايا أصبحت فارغة أو مكونة من عدد محارف قليل أقل من 5 بعد التنظيف؟ قم بحذف هذه الأسطر إن وجدت، ما عددها؟

مهمة 04: بناء النموذج المرجعي Baseline Model

في هذه المهمة سنقوم ببناء نموذج مرجعي على مسألة تصنيف على مستوى البلدان.
من أجل كل عملية تجريب ستقوم بها قم بطباعة الدقة، وضع رقم الطلب واسم النموذج ومعاملته وعمليات المعالجة مع الدقة في ال dictionary الذي ستستخدمه في نهاية كل طلب لطبع جداول المقارنة بين كل النماذج حسب المثال الموجود في الـ notebook المرفق وحسب الإرشادات في آخر الوظيفة. بالإضافة إلى ذلك سيكون هناك جدول نهائي ستثبت فيه نتائج أفضل عملية تجريب قمت بها بكل طلب. والمطلوب:

مهمة 4-1: من أجل كل مجموعة من مجموعات البيانات الثالث (تدريب, تحقق واختبار) بوضعها الخام أي دون تطبيق أي عملية معالجة عليها, استخرج شعاع الدخل X وشعاع الهدف Y الذي يمثل اللهجات multi-class classification.

مهمة 4-2: طبق خوارزمية التصنيف multinomial Naive Bayes على مجموعة التدريب، وذلك بعد تمثيل البيانات بطريقة TF-IDF، لا تقم بأي عملية تنظيف على البيانات بل استخدمها بشكلها الخام.
مهمة 4-3: قم باختبار النموذج المبني في مهمة 4-2 ، ثم قم بتثبيت النتيجة في جدول النتائج النهائية كما هو موضح في المثال في الـ notebook المرفق. لاحظ أن البيانات غير متوازنة data imbalance في الصفوف وهذا قد ينتج دقة غير صحيحة تمامًا، استخدم المقياس F-score من مكتبة sklearn واضيفه إلى الجدول.
يفضل أن تكتب الكود السابق في تابع لأنك ستقوم باستدعاءه أكثر من مرة.

مهمة 4-4: قم بتدريب النموذج السابق باستخدام عملية تنظيف أو توحيد واحدة كل مرة واحسب الدقات و اطبعها ثم ادرجها في جدول النتائج لهذا الطلب، وفي حال كانت العملية لم تحسن النتيجة لن تقوم باستخدامها في الطلب التالي.

مهمة 4-5: درب النموذج على مجموعة العمليات التي حسنت نتيجة الاختبار بالمقارنة مع النموذج المدرب في المهمة 4-2 واختبره ثم اطبع الدقات وثبتها في الجدول.

مهمة 4-6: قم بضبط معاملات TF-IDF للحصول على أفضل نتيجة، وثبت نتيجة الاختبارات في جدول نتائج الطلب.

بعد الانتهاء قم باختبار النموذج الأفضل على بيانات الاختبار، ثبت النتائج في جدول الاختبار النهائي.

مهمة 05: التدريب باستخدام نماذج التعلم العميق Deep learning

مهمة 5-1: قم بتدريب شبكة عصبونية عميقة مؤلفة من عدة طبقات متصلة بشكل كامل (connected fully)على النصوص النظيفة بأفضل مجموعة من العمليات، وثبت النتائج في جدول النتائج. السمات التي ستسخدمها هيTF-IDF.

مهمة 5-2: قم بتدريب شبكة عصبونية عميقة من النوع CNN (Networks Neural Convolutional) على النصوص النظيفة، بحيث تستخدم embedding layer بثلاث طرق:

مهمة 5-2-1:لا تستخدم أشعة مسبقة لتهيئة الطبقة بل دعها تتعدل أثناء التدريب مثلها مثل أية طبقة أخرى.

مهمة 5-2-2: استخدم مصفوفة أشعة الكلمات الناتجة عن تدريب مسبق للنموذج الذي وجدت أنه الأفضل في المهمة 2-7 بدون تعديلها (جمد طبقة ال embedding عند تدريب الشبكة).

مهمة 5-2-3: استخدم مصفوفة أشعة الكلمات الناتجة عن تدريب مسبق للنموذج الذي وجدت أنه الأفضل في المهمة 2-7 مع السماح بتعديلها أثناء التدريب.

وثبت نتائج كل طريقة في جدول نتائج الطلب بعد اختبار نموذج الطلب. ماذا تلاحظ؟ ثبت ملاحظاتك.

مهمة 5-3: لا تنس ضبط معاملات النموذج الفائقة hyper-parameters بعد اختيار أفضل مصفوفة أشعة دلالية للكلمات من تجريب الطرق الأربع في المهمة 5-2، واطبع مقدار مقياس التقييم.
لا تنسى طباعة منحنيات التعلم لاكتشاف جودة التدريب لكل شبكة.

مهمة 5-4: قم بطباعة مصفوفة التعارضات Confusion Matrix ثبت ملاحظاتك.
ما هو النموذج الأفضل؟

مهمة 5-5: قارن النتائج الناتجة عن مصفوفة التعارضات مع نتائج تحليل تشابه اللهجات على مستوى البلدان في الطلب 2-6

مهمة 5-6: استخدم النموذج الأفضل بأفضل إعدادات في تدريب نماذج على مستوى المدن والمناطق. كرر المهمتين 5-4 و 5-5.

مهمة 06: التمثيل الدلالي Semantic representation

الهدف من هذا الطلب دراسة تأثر التمثيل الدلالي عبر اللهجات للكلمات المميزة لكل لهجة والتي قمت باستخراجها في المهمة 2-7 وذلك بإجراء رسم بياني ثنائي البعد للأشعة الممثلة لهذه الكلمات باستخدام النقاط scatter plot، بحيث نلون كلمات كل لهجة بلون مميز. علماً أنه يمكنك رسم الأشعة متعددة الأبعاد بإسقاطها على فضاء ثنائي البعد باستخدام خوارزميات مثل t-SNE التي تعطي إسقاطاً أفضل من PCA.

مهمة 6-1: اطبع الرسم البياني لكل من المستويات الثلاثة باستخدام التضمينات embeddings للنموذج الذي وجدت أنه الأفضل في المهمة 2-7.
قارن وثبت ملاحظاتك
مهمة 6-2: اطبع الرسم البياني لكل من المستويات الثلاثة باستخدام التضمينات embeddings في طبقة التضمينات في النموذج الأفضل الذي دربته في الطلب السابق.
قارن وثبت ملاحظاتك

إرشادات

1. تاريخ التسليم: الخميس 26 كانون الأول 2024.
2. يشترك الطلاب لحل الوظيفة في مجموعات مؤلفة من 2 إلى 3 طلاب فقط (لا أكثر ولا أقل) ولا جدال في العدد المسموح.
3. لا يسمح بتغيير أفراد المجموعة في الوظائف القادمة.
4. تحتاج الوظيفة لتنتهي في الوقت المحدد تعاون جميع أفراد المجموعة.
5. يحصل الطالب على جزء من العلامة على حل الوظيفة والجزء الآخر على مقدار عمل الطالب بها، يجب على كل الطلاب الاشتراك بالعمل في كل الطلبات وفهمها بالتفصيل وسيأخذ علامة فقط في حال فهمه لكل الأجزاء بدون استثناء، معرفة عمل كل طالب ومقداره ستتم بالتأكيد بطريقة غير معلنة.
6. تعليمات حل الوظيفة:
   - التزم بوضع حلك في الملف القالب المرفق.
   - اختر لأسماء النماذج وعمليات المعالجة أسماًء واضحة ومعبرة، في حال عدم وجود أي إضافة قم بكتابة none.
   - قم بتخزين نتائج كل طلب والنتائج النهائية في كياني فهرس dictionary, يتألف كل منهما من 7 مفاتيح قيمة كل مفتاح مصفوفة list تضاف الى الفهرس بعملية append. المفاتيح كالتالي:
     1. number_step_question: يعبر عن رقم السؤال ورقم الطلب.
     2. name_model: ويعبر عن خوارزمية التدريب.
     3. features: ويعبر عن شكل الدخل أو سمات النموذج.
     4. parameters_model: يحتوي أسماء البارامترات الفائقة hyperparameters المعدلة وقيمها الجديدة وفي حال عدم تعديل بارامترات النموذج تضاف كلمة default.
     5. methods_preprocessing: يعبر عن خطوات المعالجة التي قمت بها قبل التدريب.
     6. accuracy: يعبر عن مقدار دقة الاختبار.
     7. F-score: ويعبر عن مقدار دقة التصنيف على بيانات الاختبار.
   - انتبه أنه هناك dictionary يضاف إليه نتائج طلبات التدريب، وهناك جدول نهائي للنماذج الأفضل من كل طلب منها.
   - لن تقبل الوظيفة بدون طباعة جدول مقارنة النماذج.
7. خطوات تسليم الوظيفة:

- قم بتغيير اسم ملف الحل المرفق بعد وضع اجاباتك ضمنه ليصبح باسماء طلاب المجموعة باللغة العربية.
- تأكد من عدم طباعة كامل البيانات, تفاديا لتجاوز حجم الملف الحد المسموح 10M.
- تأكد من تنفيذ جميع التعليمات البرمجية.
- قم بتصدير الملف للاحقة html وipynb
- تأكد أن الملفين قابلين للفتح والقراءة بوضوح وأن حجمهم لا يتجاوز 10M.
- قم برفع الملفين ipynb و html دون ضغط, على رابط تسليم الوظيفة.

8. 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.
9. 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
م. زينة دلّال، م. عُلا طبال،
م. ايليسار بري, م. حاتم بركات