كلية الهندسة المعلوماتية - قسم الذكاء الصنعي
عملي معالجة اللغات الطبيعية 2022-2023
الوظيفة الأولى - المرحلة الخامسة

التصنيف الآلي للنصوص - Text Classification

في هذه المرحلة سوف تتعلم ما يلي:

- تصنيف النصوص باستخدام طرق مختلفة للتمثيل الشعاعي للنصوص (Vectorization) وخوارزميات التصنيف الآلي (Classification).
- دراسة تأثير عمليات المعالجة المسبقة على دقة عملية التصنيف.

مرفق مع الوظيفة ملف notebook، يطلب منك أن تقوم بملئ الخلايا وفق الطلبات المدرجة أدناه. من أجل كل طلب ضع خلية نصية اكتب فيها رقم الطلب وشرح مبسط عن محتوى الكود، ثم ضع خلية فيها كود الطلب، ثم خلية فيها كود الاختبار، وفي حال طلب منك تثبيت ملاحظاتك قم بإضافة خلية نصية واكتب ملاحظاتك باللغة العربية الفصحى وبطريقة سليمة.

من أجل كل عملية تجريب تقوم بها ستقوم بطباعة الدقة، ووضع رقم الطلب واسم النموذج ومعاملاته وعمليات المعالجة مع الدقة في ال dictionary الذي ستستخدمه في نهاية كل طلب لنطبع جدول المقارنة بين كل النماذج حسب النموذج 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.

أولاً: تحميل البيانات

سنعمل على البيانات نفسها التي عملنا عليها في الجزء الرابع، والتي بلغ عددها 48463 تغريدة.
يمكنك تحميل هذه البيانات من الرابط:
https://drive.google.com/drive/folders/1121dzWdV2ZIrQSSSLs0N4P5b3qRwg428?usp=sharing
أو استدعاء التعليمات المكتوبة في بداية الـ notebook المرفق لتحميلها لديك.
البيانات مقسمة إلى ثلاثة مجموعات، على الشكل التالي:

- قسم التدريب train من أجل تدريب النماذج فقط، *وستعمل مبدأياً فقط على التغريدات الإيجابية 1 والسلبية 0، وتترك باقي الأصناف جانباً*.
- قسم التحقق validation ستختبر عليه نتائج النماذج التي دربتها بهدف توليف/ضبط tuning المعاملات hyperparameter من جهة، وأيضاً للتأكد من عدم وجود overfitting/underfitting أثناء تدريب نماذج الشبكات العصبونية.
- قسم الاختبار test ستختبر عليه دقة النماذج النهائية بعد قيامك بضبط المعاملات، لمقارنة أداء النماذج مع بعضها وحساب الدقة accuracy.

قم بتحميل البيانات وضع كل منها في data frame، قم بحذف التغريدات المكررة والأسطر الفارغة أو التي تحتوي على nan، واطبع عدد التغريدات الكلي في كل جزء.
من أجل كل قسم استخرج مصفوفة التغريدات (X) ومصفوفة الهدف (y)

ثانياً: النموذج الحدودي (Baseline): تمثيل Bag of words

طبق خوارزمية التصنيف Logistic Regression على بيانات التدريب، وذلك بعد تمثيل البيانات بطريقة Bag of Words، لا تقم بأي عملية تنظيف على المعطيات بل استخدمها بشكلها الخام.
اختبر النموذج على بيانات الاختبار، قم بتثبيت النتيجة في جدول النتائج النهائية كما هو موضح في النموذج في الـ notebook المرفق.
(يفضل أن تكتب الكود السابق في تابع لأنك ستقوم باستدعائه أكثر من مرة من أجل الطلب اللاحق.)

ثالثاً: دراسة تأثير عمليات تنظيف وتوحيد النص على عملية التصنيف

قم بتجريب عمليات التوحيد والتنظيف التي عملت عليها في القسم السابق من الوظيفة بالاستعانة بتابع preprocess_tweet الذي يعمل ك on/off على كل تعليمة فرعية والذي ينبغي أن يكون موجوداً في وظيفتك السابقة، حيث ستقوم بعملية توحيد أو تنظيف واحدة كل مرة وتحسب الدقة و تطبعها ثم تدرجها في جدول النتائج لهذا الطلب، وفي حال كانت العملية لم تحسن النتيجة لن تقوم باستخدامها بالخطوة التالية.
ملاحظة: بعد قيامك بعملية التنظيف قم بحذف التغريدات الفارغة والمكررة التي قد تنتج عن العملية التي قمت بها.
وهذه العمليات هي:

3. 1- إزالة الروابط والـ mentions من التغريدات.
3. 2- حذف الـ hashtags من التغريدات.

لاحظ أن الأشخاص المؤيدين أو المعارضين لموضوع معين عادة يستخدمون hashtags معينة في تغريداتهم لتعبر عن آرائهم، ما تأثير هذه الـ hashtags على عملية التصنيف؟، في حال كان وجودها يزيد من دقة التصنيف هل تتوقع أن إبقاء هذه الـ hashtags سيتعطي دقة جيدة في تصنيف بيانات جديدة أخذت في فترة زمنية مختلفة عن الفترة التي تم جمع البيانات فيها؟

3. 3- حذف المحارف المكررة من النص.
3. 4- التعامل مع الأرقام، في كل مرة قم بتجريب إحدى الطرق التالية، ثم استخدم الطريقة التي أعطتك الدقة الأعلى في حال كانت هذه الخطوة تحسن من النتائج:
   3. 4- 1. توحيد رموز كتابة الأرقام، باستبدال الأرقام المكتوبة بالرموز الهندية(١ - ٢ - ٣ ...) بالأرقام المكتوبة بالرموز العربية (1 - 2 - 3 ...)
   3. 4- 2. توحيد كتابة الأرقام باستبدالها جميعها (مهما كانت رموزها) برمز ما من اختيارك.
   3. 4- 2. حذف جميع الأرقام (مهما كانت رموزها) من التغريدات.
3. 5- التعامل مع المحارف غير المرغوب بها، قم بتجريب كل من الطرق التالية، واستخدم جميع الطرق التي تحسن من دقة النموذج:

3. 5- 1. حذف جميع المحارف غير العربية (باستثناء الأرقام وعلامات الترقيم والوجوه التعبيرية أو المحارف اللاتينية الموجودة في الـ hashtags).
3. 5- 2. حذف علامات الترقيم جميعها (عربية ولاتينية).
3. 5- 3. حذف الوجوه التعبيرية (emojis).
3. 6- حذف كلمات التوقف.
3. 7- تجذيع أو تجذير الكلمات العربية.
3. 8- عمليات توحيد النصوص العربية، قم بتجريب كل من الطرق التالية، واستخدم جميع الطرق التي تحسن من دقة النموذج:
3. 8- 1. توحيد طريقة كتابة الهمزات (ء ؤ ئ).
3. 8- 2. توحيد طريقة كتابة الألفات (أ آ إ)
3. 8- 3. حذف التطـــــــــويل
3. 8- 4. حذف علامات التشكيل.
3. 9- احذف الفراغات المكررة، بين الكلمات واستبدلها بفراغ واحد، لماذا لا تؤثر هذه الخطوة على النتيجة؟.
3. 10- قم بتجريب أي عملية إضافية قمت بها في القسم الماضي من الوظيفة. هل كانت هذه العملية مفيدة كما تصورت؟

ما هي مجموعة العمليات التي أعطت النتائج الأفضل؟ قم باختبار النموذج باستخدام على بيانات الاختبار، ثبت النتائج في جدول الاختبار النهائي.

رابعاً: التمثيل باستخدام TF-IDF

4. 1- باستخدام أفضل عمليات التنظيف والتوحيد الناتجة عن الخطوة السابقة، قم بتمثيل البيانات باستخدام نموذج TF-IDF مع logistic regression، بمعاملاته الافتراضية، اختبر نموذجك وثبت النتائج في جدول نتائج الطلب.

4. 2- 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.

4. 3- (**إضافي**) قم بتجريب خوارزميات أخرى بديلاً عن logistic regression من خوارزميات التعلم الآلي، واضبط معاملاتها، في حال أعطت نتائج جيدة قم بتثبيت النتيجة في جدول النتائج النهائي.

ملاحظة: مجموعة العمليات التي أعطتك أفضل نتائج عند تطبيق عملية تمثيل البيانات باستخدام نموذج Bag of words لن تعطي بالضرورة أفضل نتائج عند تطبيق نموذج TF-IDF، وأيضاً تختلف باختلاف خوارزمية التدريب وحتى معاملات كل نموذج وخوارزمية تطبقها ولكن للتسهيل لن نقوم باختبار ذلك هنا.

خامساً: التدريب باستخدام شبكة عصبونية عميقة

5. 1- قم بتدريب شبكة عصبونية عميقة مؤلفة من عدة طبقات متصلة بشكل كامل (fully connected) على التغريدات النظيفة، وثبت النتائج في جدول نتائج الطلب.

5. 2- (**إضافي**) قم بتجريب بنى architecture مختلفة عن البنية المذكورة في المحاضرات، ولا تنسى أن الحصول على بنية جيدة يتطلب تجريب الكثير من البنى والكثير من عمليات الضبط على المعاملات، من أجل كل عملية تجريب لكل بنية (وليس لكل عملية ضبط) ثبت النتائج في جدول نتائج الطلب.

قم باختبار النموذج الأفضل على بيانات الاختبار وتثبيت النتائج في جدول النتائج النهائي.

5. 3- قم بتدريب شبكة عصبونية عميقة من النوع Convolutional Neural Networks (CNN) على التغريدات النظيفة، وثبت النتائج في جدول نتائج الطلب.

5. 4- (**إضافي**) قم بتجريب بنى architecture مختلفة عن البنية المذكورة في المحاضرات، (أو نوع آخر من الشبكات التكرارية RNN) من أجل كل عملية تجريب لكل بنية (وليس لكل عملية ضبط) ثبت النتائج في جدول نتائج الطلب.

قم باختبار النموذج الأفضل على بيانات الاختبار وتثبيت النتائج في جدول النتائج النهائي.

5. 5- (**إضافي**): قم بعملية إظهار (Visualization) لطبقة الـ embedding الناتجة عن عملية التدريب، ثبت ملاحظاتك.

سادساً: التدريب على كل الأصناف

6. 1- قم باختيار أفضل نموذج لديك ودربه على كل الأصناف (classes) وليس فقط على التغريدات السلبية والإيجابية، ثبت النتائج في جدول النتائج النهائي وارفق باسمه جملة with all classes.

ما هو تأثير إضافة الصفوف هذه؟ تبت ملاحظاتك.

6. 2 (**إضافي**) قم بطباعة مصفوفة التعارضات Confusion Matrix ثبت ملاحظاتك.

ملاحظة: المعطيات ليست متوازنة imbalanced بمعنى أن عدد التغريدات في كل صنف ليس متساوي، هذا قد يشكل مشكلة -ولكن ليس بالضرورة-، في حال أردت موازنتها بإمكانك ذلك، ابحث عن هذا الموضوع (resampling) أو قم بموازنتها يدوياً عن طريق أخذ عدد متساوي من التغريدات بشكل عشوائي لكل صنف من الأصناف الأربعة، سنتكلم عن هذا الموضوع بشكل مستفيض في مادة التعلم الآلي.

الإرشادات:

- آخر موعد لتسليم هذا الجزء هو يوم الثلاثاء (2021-12-27) الساعة 11:59 مساءً، ولا يوجد تمديد للموعد على الإطلاق. يمكنك تسليم الوظيفة متأخرً ولكن سوف يترتب على ذلك حذف جزء من العلامة (10% من العلامة في حال سلمت خلال الساعات الـ 12 الأولى من انتهاء الموعد، ومن ثم 5% عن كل يوم تأخير).
- قم بتغيير اسم الملف بكتابة اسمك باللغة العربية مكان [your_name]، قم برفع نسختين من الـ notebook الأول بصيغة ipynb والثاني بصيغة html على الرابط الآتي بدون ضغطه:

https://forms.gle/dKcbWRpdh1v5VLjq8

- لن تقبل الوظيفة بدون طباعة جدول مقارنة النماذج.
- نتائج كل طلب والنتائج النهائية تخزن في dictionary يتألف من 5 مفاتيح، الأول باسم question_step_number ويعبر عن رقم السؤال ورقم الطلب، والثاني باسم model_name ويعبر عن خوارزمية التدريب والثالث باسم parameters ويعبر عن بارمترات النموذج الذي طبقته، والرابع باسم preprocessing_methods ويعبر عن خطوات المعالجة التي قمت بها قبل التدريب، والخامس باسم accuracy ويعبر عن دقة الاختبار، كل مفتاح منها قيمته list يتم ادخلها بتعليمة append.
اختر لأسماء النماذج وعمليات المعالجة أسماءً واضحة ومعبرة، في حال عدم وجود أي عملية قم بكتابة none.
انتبه أنه هناك dictionary لكل طلب، وهناك dictionary نهائي للنماذج الأفضل من كل طلب، قد تختلف أعمدة كل منها حسب الطلب.
- تأكد قبل تسليمك للملف أن جميع الخلايا منفذة بشكل كامل والنتائج معروضة فيه. وتأكد أن الملف يعمل وأنه قابل للقراءة بوضوح، قم بفتح ملف الـ html وتأكد أنه صحيح وأن حجم الملف المسلم صغير، في حال كان كبيراً تأكد من أنك لم تقم بطباعة البيانات خطأً فيه، في حال تجاوز حجمه الحد المسموح (10M) لن يتم رفعه.

وتذكر أن:

تنفذ الإرشادات وتعليمات التسليم حرفياً دون إبداعات إضافية، لا زيادة ولا نقصان.

عند وجود أي تشابه بين وظيفتي طالبين سيخسر الطالبان العلامة معاً دون مراجعتهما

مدرسو المادة:
زينة الدلال
علا طبال
ايليسار بري