كلية الهندسة المعلوماتية - قسم الذكاء الصنعي
عملي معالجة اللغات الطبيعية 2023-2024
الوظيفة الثانية

القسم الأول: Sequence Labeling

مهمة هذا القسم تصنيف الكيانات الاسمية الموجودة في الجملة named entities recognition.

أولاً: تجهيز البيانات: Data Preparation

تتضمن مجموعة البيانات التي ستتعامل معها الآن جدولاً مؤلفاً من 4 أعمدة، Word i قيمه كلمات الجمل، Word i entity tag تصنيف الكلمة بصيغة IOB ولا يوجد بجانب B و I اسم الكيان لأن الداتا صنفت المرض فقط ككيان اسمي فيما عدا ذلك يعتبر O، عمود Word i POS صنف الكلام لكل كلمة، وأخيراً عامود stopword قيمه ثنائية: yes إذا كانت الكلمة كلمة توقف و No فيما عدا ذلك.
عدد الكلمات الكلي 62506 كلمة، وعدد الكلمات الفريدة 4882 كلمة. الجمل مقسمة إلى tokens وحدود كل جملة هي النقطة.
يمكنك تحميل هذه البيانات من الرابط:
https://drive.google.com/drive/folders/1O4DSoum-0FE1peRaz9C-x_Z6tuMr_HIF?usp=drive_link
مرفق مع ملف البيانات ملف بايثون يحتوي تابع التقطيع tokenization بمقاربة طريقة AMIRA في تقطيع النصوص العربية والمستخدمة لتقسيم الداتا أعلاه.
قم بمراجعة الورقة البحثية التي تتحدث عن بناء المجموعة لفهم طريقة تقسيم المفردات tokenization لو أردت: https://sci-hub.se/10.3390/data5030060
والمطلوب:

1 .1- اكتب تابع لتحويل الأسطر إلى قائمة من tuples، كل tuple تحتوي الكلمة وخرج تصنيفها المتوقع entity tag.

1 .2- اكتب تابع لتحويل الأسطر إلى جمل، خرجه قائمة من قوائم, كل قائمة تعبر عن جملة وتحتوي مفردات الجملة.

1 .3- اكتب تابع لتحويل الأسطر إلى قائمة من قوائم, كل قائمة تعبر عن جملة وتحتوي tuples كل tuple تحتوي الكلمة token و خرج التصنيف المتوقع entity tag لها.

1 .4- اكتب تابع خرجه قائمتين، القائمة الأولى هي قوائم مفردات الجمل، القائمة الثانية هي قوائم تصنيف كلمات الجمل.

1 .5- قسم مجموعة البيانات إلى مجموعتي تدريب واختبار بطريقة مناسبة (انتبه إلى كون الكلمات متسلسلة). طبق كل التوابع السابقة على جملة من اختيارك من المجموعة بنهاية كل طلب.

1 .6- احسب عدد الجمل في كل من المجموعتين.

1 .7- احسب عدد الكيانات الاسمية الموجودة في مجموعة التدريب ونسبتها إلى عدد الكلمات الكلي.

1 .8- احسب عدد الكلمات الغامضة في مجموعة التدريب ونسبتها إلى عدد الكلمات الكلي.

1 .9- أوجد الصنف الأكثر تكراراً في مجموعة التدريب.

من أجل جميع الطلبات القادمة, طبق الحل بنهاية كل طلب على الفقرة الآتية:
"التصلب المتعدد مرض يحتمل أن يسبب إعاقة الدماغ والحبل النخاعي (الجهاز العصبي المركزي). عند الإصابة بمرض التصلب المتعدد، يهاجم الجهاز المناعي غمد الحماية (المايلين) الذي يغطي الألياف العصبية، ويسبب مشكلات في الاتصال بين الدماغ وبقية الجسم. في النهاية، يمكن أن يسبب المرض تلفًا أو تدهورًا دائمًا في الألياف العصبية."
ولا تنس استخدام تابع التقطيع tokenization المرفق.
اضف جدول النتائج كما في ملف notebook المرفق، يتضمن اسم النموذج model_name، السمات في حال وجدت features وتكون قيمتها None في حال عدم وجودها، دقة النموذج accuracy.

ثانياً: التصنيف باستخدام الإحصاء: Statistics

2 .1- ستقوم ببناء مصنف حدي لمقارنة أداء النماذج الآتية معه. أوجد أسخف طريقة ممكنة لبنائه واحسب دقة هذه الطريقة.
2 .2- طبق مصنف Lookup Tagger باستخدام أكثر 100 كلمة مكررة واحسب دقته.
طبقه مرة أخرى مع استخدام backoff technique واحسب دقته.
اكتب استنتاجك.
2 .3- طبق كل من unigram tagging و bigram tagging على مجموعة التدريب بعد تحويلها إلى الشكل المناسب واحسب دقة كل منهما.
اكتب استنتاجك.

ثالثاً: التصنيف باستخدام نماذج تعلم الآلة: ML Classification

3 .1- اكتب توابع لاستخراج سمات كلمة في الموقع i من جملة الدخل، يعيد ما يلي:
3. 1. 1- الكلمة + سمات صرفية للكلمة من اختيارك.
3. 1. 2- السمات السابقة + الكلمة السابقة والتالية.
3. 1. 3. السمات السابقة + صنف الكلام POS للكلمة فقط.
3. 1. 4- السمات السابقة + أصناف الكلام للكلمة التالية والسابقة.
3. 1. 5- السمات السابقة + bigram لأصناف كلام الكلمة الحالية والسابقة والتالية.
3. 1. 6- السمات السابقة + تصنيف الكلمة السابقة Consecutive features.
3. 1. 7- السمات السابقة + سمات من اختيارك في حال وجودها. برر اختيارك لها.
3. 2- درب Multinomial naive bayes على السمات السابقة بالتدريج واحسب الدقة لكل منها.
3. 3- درب نموذج Multinomial logistic regression على السمات السابقة بالتدريج واحسب الدقة لكل منها.
3. 4- درب نموذج CRF على السمات السابقة بالتدريج واحسب الدقة لكل منها.
3. 5- درب نموذج HMM واحسب الدقة.
اكتب ملاحظاتك بنهاية كل طلب.

رابعاً: التصنيف باستخدام التعلم العميق: DL Classification

4 .1- اكتب التوابع اللازمة لتحويل مجموعتي البيانات إلى الشكل المناسب لإدخالها على شبكة عصبونية، مع أخذ حل الكلمات غير الموجودة out of vocabulary بعين الاعتبار.

4. 2- صمم شبكة مؤلفة من:

طبقة التضمين embedding layer

طبقة LSTM

طبقة الخرج dense layer

ودربها على مجموعة التدريب مع طباعة منحنيات الدقة والخسارة بعد التدريب، واختبر أداءه على مجموعة الاختبار واحسب الدقة.

انتبه إلى ضرورة معالجة رمز توحيد أطوال النصوص PAD عند التدريب وحساب الدقة.

4. 3- قم بضبط المعاملات الفائقة واطبع النتائج.

القسم الثاني: Sequence to Sequence

مهمتك في هذا القسم هو تعلم إجراء ضبط نموذج لغوي كبير مدرب مسبقاً fine-tune large language model من أجل توليد إجابات لأسئلة طبية.

ستتعلم في هذا القسم:

- التعامل مع huggingface ومكتبة transformers.
- فهم وتطبيق نقل معرفة transfer learning النماذج اللغوية الكبيرة لحل مهام مختلفة.

قم بتحميل مجموعة البيانات من الرابط أدناه والتي تتضمن مجموعتي تدريب واختبار لأسئلة وإجابات طبية سبق أن تعاملت مع قسم منها، البيانات خالية من المفردات غير الهامة (كأسماء الأطباء والتواريخ ... الخ):

https://drive.google.com/drive/folders/1QNipRmidM73UomZnLCSWkdQPR8BQYjX9?usp=drive_link

تتضمن 173055 سؤال وجواب ل 16 اختصاص مختلف.

<u>والمطلوب:</u>

1. (**إضافي**) تتضمن البيانات أخطاء إملائية من نوع space merge حيث تتصل مجموعة من الكلمات الصحيحة مع بعضها مثل:

   سونارعطتني

   ماصارحمل

   معأنو

   هذه الحالات لا يستطيع tokenizer النموذج الذي ستتعامل معه الآن فصلها، قم بتصحيح هذه الأخطاء بطريقة من اختيارك مع شرحها.
2. قم بتحميل مكتبة arabert لإجراء fine tune لنموذج AraGPT2 بحجم من اختيارك على مجموعة التدريب باستخدام خوارزميات فك ترميز decoding مختلفة واطبع النتائج، يمكنك الاستعانة بالروابط التالية:

   https://github.com/aub-mind/arabert/tree/master

   https://huggingface.co/aubmindlab/aragpt2-base

   https://colab.research.google.com/drive/1T5xKRAtKeuUmmokyThvkK8pulHw02CoK?usp=sharing

   https://212digital.medium.com/fine-tuning-the-gpt-2-large-language-model-unlocking-its-full-potential-66e3a082ab9c

   من الأفضل استخدام colab في هذه المرحلة.
3. طبق النموذج الجديد بعد حفظه على مجموعة الاختبار واطبع معيار تقييم الأداء المناسب، واختر خمس أسئلة عشوائية منها وولد لها إجابات واطبعها.

الإرشادات:

- تسلم الوظيفة قبل يوم الثلاثاء 22-12-2023 الساعة 59:11 مساءً. يمكنك تسليم الوظيفة متأخرً ولكن سوف يترتب على ذلك حذف جزء من العلامة (5% عن كل يوم تأخير).
- تتطلب الوظيفة مجموعة مكونة من ثلاثة أعضاء، هم نفس أعضاء المجموعة التي عملت على الوظيفة السابقة ولا مجال لتغيير العدد ولا أفراد المجموعة. علامة الطلاب في المجموعة الواحدة ليست موحدة، يحصل الطلاب على جزء من العلامة على حل الوظيفة والجزء الآخر على مقدار عمل الطالب بها.
- قم بتغيير اسم الملف بكتابة أسماء أفراد الفريق باللغة العربية مكان [your_name]، قم برفع نسختين من الـ notebook الأول بصيغة ipynb والثاني بصيغة html وملف جدول النتائج النهائية على الرابط الآتي بدون ضغطه:

https://forms.gle/LTfutQftKzZiyQYT6

- عندما تقوم باختبار الكود الخاص بك تأكد من أنك لم تقم بطباعة كل البيانات ضمن النوتبوك يكفي أن تطبع حالات الاختبار المطلوبة منك.
- 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 كل البيانات بشكل غير مقصود فيه، في حال تجاوز حجمه الحد المسموح (10M) لن يتم رفعه.
- لن تقبل الوظيفة بدون طباعة جدول مقارنة النماذج.
- مرفق مع الوظيفة ملف 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.

وتذكر أن:

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

عند وجود أي تشابه بين وظيفتي مجموعتين ستخسر المجموعتان العلامة معاً دون مراجعتهما.

مدرسو المادة: زينة الدلال، عُلا طبال، ايليسار بري