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

تحليل ومعالجة البيانات Data Analysis and Preprocessing

في هذه المرحلة وهي مرحلة تنميط أو تصنيف المعطيات بشكل يدوي سوف تتعلم ما يلي:

- تحليل البيانات التي قمنا بجمعها وتنميطها معاً، وذلك لمحاولة فهم محتوى هذه النصوص بدون قرائتها كلها.
- تنظيف وتوحيد النص بهدف تجهيزه للقيام بعملية التصنيف التي ستقوم بها في الجزء التالي من الوظيفة.

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 التابع ودخلة وخرجه، 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.

أولاً: تحميل البيانات

تم تجميع التغريدات التي قمنا بتصنيفها ,وبلغ عددها 48463 تغريدة، ثم قسمت إلى ثلاث مجموعات، وهي بيانات التدريب (train) و وتحقق(validation) والاختبار (test).
يمكنك تحميل هذه البيانات من الرابط:
https://drive.google.com/drive/folders/1121dzWdV2ZlrQSSSLs0N4P5b3qRwg428?usp=sharing

أو استدعاء التعليمات المكتوبة في بداية الـ notebook المرفق لتحميلها لديك.
قم بتحميل قسم التدريب (train) فقط من البيانات التي لديك، وذلك لكي تتجنب تسريب البيانات (data leakage)، ضع هذه البيانات في data frame، وقم بطباعة الـ head وعدد التغريدات الكلي، قم بحذف التغريدات المكررة والأسطر الفارغة أو التي تحتوي على nan، واطبع عدد التغريدات الكلي.

ثانياً: فهم وتحليل النص Text Analysis

في هذه المرحلة سنقوم بمجموعة من العمليات بهدف فهم المعطيات التي جمعناها، هذه العمليات تمكنك من فهم محتوى التغريدات بصفوفها المختلفة لتتمكن من تنظيفها بشكل أفضل، ومن ثم تطبيق آلية التصنيف المناسبة لها.
والمطلوب:

2. 1- قم بتوحيد أسماء الصفوف labels لأن الصف الواحد يمكن أن يكون مكتوب بأكثر من طريقة، مثال: neutral, nautral, netural ...

2. 2- احسب عدد التغريدات في كل صف، ارسم pie chart لها.

2. 3- قم بتقسيم النص إلى وحدات (tokenization) بطريقة مناسبة، لاحظ أن تقسيم النص على الفراغات وعلامات الترقيم يسبب مشكلة وذلك لاحتواء النص على روابط و mentions ...

قم بطباعة أول 5 تغريدات بعد التقسيم.

احسب عدد الكلمات الكلي وعدد المفردات في بيانات التدريب.

2. 4- ارسم WordCloud لكلمات النص كاملاً، (انتبه أن الكلمات المكتوبة باللغة العربية تحتاج إلى إعدادات خاصة لعرضها، قم بالبحث عنها وتطبيقها في الكود الخاص بك).

2. 5- قم بكتابة تابع يأخذ نصاً مقطعاً ويوجد كل من:

- أكثر 15 كلمة مكررة في النص.
- 10 من الكلمات الأقل تكراراً في النص (hapaxes)
- أكثر 10 ثنائيات كلمات مكررة في النص.
- أكثر 10 ثلاثيات كلمات مكررة في النص.
- أكثر 10 ثنائيات كلمات مهمة في النص (Bigram Collocations)
- أكثر 10 ثلاثيات كلمات مهمة في النص (Trigram Collocations)

طبق التابع السابق على البيانات كاملة وعلى كل من الصفوف على حدى.

هل الكلمات أو العبارات الأكثر تكراراً هي التي تميز الصف؟

قم بتطبيق التابع نفسه على النص بعد حذف كلمات التوقف. هل هناك كلمات أو عبارات مميزة لأي من الصفوف؟ هل ثنائيات وثلاثيات الكلمات الأكثر تكرارً هي نفسها الـ collocations؟ اكتب ملاحظاتك.

2. 6- اكتب تابع يقوم بإيجاد أكثر 10 هاشتاغات شيوعاً في نص،

طبق التابع السابق على البيانات كاملة وعلى كل من الصفوف على حدى.

هل هناك هاشتاغات مميزة لأي من الصفوف؟ اكتب ملاحظاتك.

2. 7- ارسم histogram لطول التغريدات بالنسبة لعدد الكلمات في كل تغريدة.

2. 8- 'طبق خوارزمية LDA وهي من خوارزميات نمذجة النصوص topic modeling بعد البحث عن كيفية تطبيقها على النصوص.

ثالثاً: عمليات تنظيف وتوحيد النص Text Cleaning and Normalization

لا بد أنك لاحظت أن البيانات التي لديك تحتوي على الكثير من المشاكل، سنقوم بمجموعة من العمليات لتجهيز المعطيات لعملية التصنيف، بحيث نوحد طريقة الكتابة فيها قدر الإمكان.

من أجل كل من العمليات التالية قم بكتابة تابع يأخذ نصاً ويعيد النص بعد القيام بالعملية المطلوبة. ثم سوف تقوم باختبار التابع نفسه بطباعة الخرج عند تطبيق التابع على أول 5 تغريدات من البيانات.

(قم بتحقيق هذه التوابع باستخدام التعابير النظامية فقط وليس باستبدال الحروف أو المرور عليها واحداً واحداً)

والمطلوب

3. 1- إزالة الروابط والـ mentions من التغريدات.

3. 2- حذف الـ hashtags من التغريدات

3. 3- حذف المحارف المكررة من النص (مثل رووووووسيا ← روسيا)، (لاحظ أن الكلمات العربية يمكن أن تسمح بتكرار حرفين فقط).

3. 4- التعامل مع الأرقام، قم بتحقيق كل من الطرق التالية، (عند تنظيف بياناتك قد تقوم بأي من هذه العمليات، ستحدد الأفضل منها لاحقاً):

3. 4- 1. توحيد رموز كتابة الأرقام، قم 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، (عند تنظيف بياناتك قد تقوم بأي من هذه العمليات، ستحدد الأفضل منها لاحقاً):

3. 5- 1. حذف علامات الترقيم جميعها (عربية ولاتينية) (انتبه أنك لا يجب أن لا تحذف # و _ من الـ hashtags).

3. 5- 2. حذف الوجوه التعبيرية (emojis).

3. 5- 3. حذف جميع المحارف غير العربية (باستثناء الأرقام وعلامات الترقيم والوجوه التعبيرية أو المحارف اللاتينية الموجودة في الـ hashtags)، انتبه أنه من الأسهل لك هنا أن تفكر بما يجب أن تبقيه من النص بدلاً من التفكير بما يجب أن تحذفه.

3. 6- حذف كلمات التوقف (هل هناك كلمات توقف تظن أنه من الأفضل أن تبقيها؟).

3. 7- تجذيع أو تجذير الكلمات العربية فقط.

3. 8- عمليات توحيد النصوص العربية:

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.

3. 10- قد تحصل بعد تطبيق عمليات التنظيف على تغريدات مكررة (مثلاً كانت تختلف عن بعضها بالمنشنات mentions)، قم بحذف الأسطر التي تحتوي تغريدات مكررة.

<u>إضافي:</u>

- بعض الكلمات وخصوصاً المعربة منها قد تُكتب كل مرة بطريقة مختلفة، مثل: أوكرانيا و أكرانيا وغيرها وقد يكون التجذيع والتجذير غير مفيد هنا في توحيدها، اكتب تابع لتوحيد الكلمات التي يمكن كتابتها بطرق مختلفة.
- لاحظ أن التغريدات مكتوبة بمزيج من اللهجات والفصحى، ولاحظ أيضاً مقدار الأخطاء الإملائية الموجودة فيها، اكتب تابع لتصحيح هذه الأخطاء.

أي عملية قد تجدها مفيدة في هذه المرحلة ولم تذكر بإمكانك تطبيقها مع شرح رؤيتك لفائدتها...

رابعاً: تجهيز بيانات التدريب Prepare Training Data

الآن سنطبق العمليات السابقة على بيانات التدريب، لتصبح هذه البيانات جاهزة للمرحلة التالية من الوظيفة. هذه ليست بالضرورة مفيدة لعملية التصنيف، ولا يمكن معرفة ذلك إلا بالتجريب.

4. 1- قم بكتابة تابع يستدعي كل هذه التوابع، ونفذها بترتيب مناسب، وبطريقة تتيح لك تشغيل أو عدم تشغيل أي عملية منها (أي بطريقة off/on لكل تابع فرعي كما هو موجود في ملف الـ notebook المرفق). اختبر التابع على التغريدات الخمسة الأولى في بيانات التدريب.

4. 2- طبق التابع السابق بتطبيق كل عمليات التنظيف على بيانات التدريب كاملةً، (بدون طباعتها!!) احسب عدد الكلمات الكلي وعدد المفردات في البيانات قارن هذا الرقم بعددها قبل التنظيف، ثبت ملاحظاتك.

4. 3- هل يوجد تغريدات أصبحت فارغة أو مكونة من عدد محارف قليل (أقل من 5) بعد التنظيف؟ قم بحذف هذه التغريدات إن وجدت، ما عددها؟

4. 4- هل يوجد تغريدات مكررة؟ قم بحذف التغريدات المكررة، ما عددها؟ ما عدد التغريدات النهائي؟

الإرشادات

- تسلم الوظيفة قبل يوم الاثنين 12-12-2022 الساعة 59:11 مساءً. 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/tcXMeG1fKphNNxyb9

- 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 (3M) لن يتم رفعه.
- عندما تقوم باختبار الكود الخاص بك تأكد من أنك لم تقم بطباعة كل البيانات ضمن النوتبوك يكفي أن تطبع حالات الاختبار المطلوبة منك.

عدم تقيدك بالإرشادات (كأن لا تتقيد بطريقة كتابة خلايا كل طلب أو كتابة اسمك بالعربية أو تسليم ملف كبير الحجم لأنك طبعت كل التغريدات ...) سينقص من علامتك حتماً -خلص مقفعة معنا من ورا تخبيصات الأجزاء السابقة- لذلك تأكد من قراءتك للتعليمات و تطبيقك لها بحذافيرها.

عند وجود أي تشابه بين وظيفتي طالبين سيحصل الطالبان على الصفر بدون مراجعتهما (هذا خبر وليس تهديد عزيزي الطالب).

مدرسو المادة:
زينة الدلال
علا طبال
ايليسار بري