وظيفة معالجة اللغات الطبيعية - القسم الرابع

المعالجة المسبقة Text Preprocessing

قمنا في المراحل السابقة بجمع تغريدات عن لقاح فيروس كورونا وتصنيفها إلى أربعة أصناف، والمطلوب منك في هذه المرحلة فهم المعطيات التي حصلت عليها والقيام بعمليات التنظيف والتوحيد الممكنة.

1- تحميل البيانات:

تم جمع كل هذه التغريدات التي قمنا بتصنيفها في ملف موحد، حيث بلغ عدد التغريدات 32262 بعد إزالة التغريدات المكررة، ثم تم تقسيمها إلى بيانات تدريب (train) وتحقق (validation) واختبار (test)، يمكنك تحميل هذه البيانات من الرابط:

https://drive.google.com/file/d/1KepfzAhJ7dloG8XaWQf0ovQipDHYS8al/view?usp=sharing

أو استدعاء التعليمات المكتوبة في ملف النوتبوك المرفق لتحميلها من النوبتوك لديك.

قم بتحميل هذه البيانات وضعها في data frame بهدف معالجتها لاحقاً.

2- فهم وتحليل النص Text analysis:

سنقوم بمجموعة من العمليات بهدف فهم المعطيات التي قمنا بجمعها، ومن أبسط طرق فهم المعطيات الرسم visualization وبمساعدته في هذه المرحلة مع مجموعة عمليات أخرى ستتمكن من معرفة محتوى التغريدات بشكل أفضل لتتمكن من تنظيفها وتوحيدها والعمل على تصنيفها لاحقاً.

ملاحظة مهمة: قم بالعمليات التالية فقط على قسم التدريب train من البيانات، كي لا تقع في فخ تسريب البيانات (data leakage).

ستقوم بما يلي:

1. حساب عدد التغريدات من أجل كل صنف (ويفضل التوضيح برسم بياني).
2. إيجاد الكلمات والتراكيب (n-grams) الأكثر تكراراً والتي تدعى في هذه الحالة collocation بالإضافة إلى الكلمات والتراكيب الأقل تكراراً والتي قد تكون باحتمال كبير مغلوطة إملائياً، بعد إيجادها اكتب ملاحظاتك عن النتائج في ملف النوتبوك بالطريقة الموجودة في قسم الارشادات.
3. كرر نفس العملية (2) ولكن من أجل كل صنف (مثلاً الكلمات الأكثر تكراراً لكل صنف).
4. رسم histogram لطول التغريدات.

5. إيجاد ال trending hashtags أي الهاشتاغات الأكثر شيوعاً مع رسم خط بياني (أو histogram) لهذه الهاشتاغات بالنسبة لعدد التغريدات التي تحتويها.
6. كرر نفس العملية (5) من أجل كل صنف.
7. **طلبات بعلامات إضافية** (لمن استطاع إليها سبيلا):

7.1. طبق إحدى خوارزميات ال topic modeling مثل LDA وقم بشرح ما فهمته منها في بداية الفقرة في النوتبوك حسب قسم الارشادات، بالطبع لن تنظر إلى السقف وتتأمله لتكتشف ما هوة ال topic modeling بل ستقوم بالبحث، يمكنك الاستعانة بمكتبة gensim.

7.2. لاحظ أن النصوص التي لدينا مزيج من نصوص مكتوبة بلغة عربية فصحى وعامية بلهجات مختلفة، ولاحظ أيضاً مقدار الأخطاء الإملائية الموجودة فيها، ابحث عن طريقة لتصحيح هذه الأخطاء وطبقها لأنها قد تساهم في تحسين جودة نتائجك لاحقاً (دقة أعلى = علامات أكثر).

-> يمكنك الإبداع في هذه المرحلة كيفما شئت وإن وجدت أثناء البحث طرق مثيرة للاهتمام لفهم النصوص يمكنك استخدامها أيضاً.

3- تنظيف وتوحيد النص Text cleaning and Normalization:

قم بتنفيذ مجموعة من الخطوات بهدف توحيد طريقة الكتابة في التغريدات قدر الإمكان، وذلك بهدف تحسين دقة التصنيف لاحقاً:

1. حذف المنشنات والروابط والصور. (لماذا لا نحذف الهاشتاغات؟)
2. حذف الأحرف المتكررة في الكلمات، مثل: كوروووونا، هههههههه، ليش يعني؟!!!!!!
3. لاحظ أن بعض الكلمات قد تُكتب كل مرة بطريقة مختلفة، مثل: كورونا قد تُكتب كرونا، كيف يمكنك توحيدها؟ اكتب التابع واشرح بعده فيما إذا كانت هذه الخطوة مفيدة في كل الحالات أم لا.
4. توحيد الأرقام باستبدالها برمز ما، أو حتى يمكنك حذفها مع الانتباه إلى أن وجود الرقم في بعض المواضع هام جداً، أين يمكن أن يكون هام ولا يمكن حذفه؟
5. مجموعة توابع لحذف المحارف التي ممكن أن تكون غير مهمة في النص، مثل علامات الترقيم، والوجوه التعبيرية (الإيموجيز) وكلمات التوقف stop words (قم بكتابة تابع لكل منها) مع ملاحظة أن حذف هذه الأمور ليس أمر جيد دائماً، لماذا؟
6. يوجد في التغريدات كلمات ومحارف ليست من اللغة العربية، انتبه أنه من الأسهل لك هنا حتى تقوم بتنظيف النص أن تفكر بما يجب أن تبقيه من النص بدلاً من التفكير بما يجب أن تحذفه.
7. تجذيع أو تجذير الكلمات stemming/lemmatization مع ملاحظة أن الهاشتاغات لا يجب أبداً أن تجذع أو تجذر، لماذا؟

8. باللغة العربية عموماً نقوم بتوحيد محارف مثل الياء والهمزات ... الخ، على سبيل المثال:

ي --------> ى

أ، إ، آ --------> ا

ؤ --------> ء

ئ --------> ء

ابحث عن هذا الأمر وطبقه، لماذا نقوم بهذه الخطوة برأيك؟

- في ال dataframe قد تجد قيم فارغة null قم بحذف الأسطر التي تحتويها.
- قد تحصل بعد تطبيق هذه العمليات على نصوص مكررة أو مكونة من محرف واحد، قم بحذفها.(مثلاً قد يختلف نص عن آخر من حيث المنشنات، بعد حذفها سيصبح محتوى التغريدة نفسه).
- أي عملية قد تجدها مفيدة في هذه المرحلة ولم تذكر بإمكانك تطبيقها مع شرح رؤيتك لفائدتها...

<u>الإرشادات:</u>

- **لا** تقم بتطبيق عمليات التنظيف والتوحيد على المعطيات مباشرة، بل اكتب لكل خطوة تابع يأخذ تغريدة واحدة كدخل ثم قم بتجريب التابع على تغريدة ما (ك test case) وطباعة التغريدة قبل وبعد تطبيق الخطوة.
- اشرح كل تابع قمت بكتابته قبل الخلية الموجود فيها مباشرة **باللغة العربية**، وضع تعليق يشرح دخل وخرج كل تابع كما في ملف النوتبوك المرفق والذي ستعمل عليه ك template فارغ.
- اكتب إجابات الأسئلة وملاحظاتك بعد الخلية ذات الصلة **باللغة العربية**.
- عند تسليم النوتبوك يجب أن يكون **مُنفذ والنتائج معروضة فيه**.
- بعد الانتهاء من كتابة توابع التنظيف والتوحيد قم بكتابة تابع ينفذ كل هذه التوابع على تغريدة **ولكن بطريقة تتيح لك إلغاء تنفيذ عملية منها دون تعديل هذا التابع لاحقاً (أي بطريقة on/off لكل تابع فرعي)** كما هو موجود في ملف النوتبوك المرفق، لأنك ستكتشف في مرحلة التصنيف أن بعض العمليات لن يكون ضرورياً وحتى أنه لن يعطي نتائج أفضل.
- عندما تقوم باختبار الكود الخاص بك تأكد من أنك **لم تقم بطباعة كل التغريدات** (أو عدد كبير منها) ضمن النوتبوك يكفي أن تطبع حالة اختبار واحدة، أو فقط ال head لل dataframe التي تستعملها.

<u>تسليم الوظيفة:</u>

- آخر موعد لتسليم هذا الجزء هو يوم السبت (4-12-2021) الساعة 11:59 مساءً.
- قم بتغيير اسم الملف بكتابة اسمك مكان [your_name] باللغة العربية.

- قم برفع نسختين من النوتوبك، الأول بصيغة ipynb والثاني بصيغة html، انتبه قبل تسليمك للملف أن الملف يعمل وأنه قابل للقراءة بوضوح.
- يسلم ملف الوظيفة على الرابط التالي

https://forms.gle/AJ6PUUCtJiUmedT5A

ويرجى التأكد من **صلاحية الوصول** إلى ملفك على الرابط المرسل.

- **لا** تقم برفع ملف المعطيات (الداتا) مع النوتبوك، ارفع النوتبوك فقط دون أن تقوم بضغطه، (يجب أن يكون حجم الملف صغير نسبياً في حال لم تطبع البيانات فيه)

تنفذ هذه الإرشادات حرفياً دون إبداعات إضافية.
عند وجود أي تشابه بين وظيفتي طالبين سيخسر الطالبان العلامة معاً

مدرسوا العملي: م.زينة الدلال - م. علا طبال