****

**بسم الله الرحمن الرحيم**

**Sudan University of Science and Technology**

**Collage of Computer Science and Information Technology**

**Department of Information Technology and Software Engineering**

**Graduation Project**

**Build Sentiment Analysis Model for the Sudanese Dialect for the Internet Service (Sudani - Case Study)**

**Preparation:**

**- Athar Abdalgauom (information technology)**

**- Rabab Khalifa (Software Engineering)**

**- Ryan Jafar (Software Engineering)**

**Supervising:**

**A. Muhammad Al-Fateh Othman**

**October 2022**

مقدمة :

سنوضح فيهذا المستند المنهجية التي اتبعت حتى الان بغرض التوضيح و التقييم بعرض الخطوات كل على حدى ( التعليقات قبل و بعد، صور للاكواد و المكتبات المستخدمة ، عرض و شرح و صور للنتائج ) مع مقارنة المنهجية بالمقالات عن المجال .

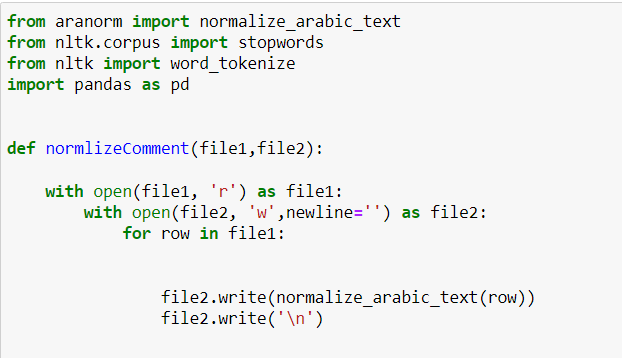
1-جمع التعليقات :

تم جمع التعليقات يدويا من صفحة سوداني للاتصالات لم يتم مراعاة اعتبار لسلبية او ايجابية او حيادية التعليق او علاقتها مع المنشور

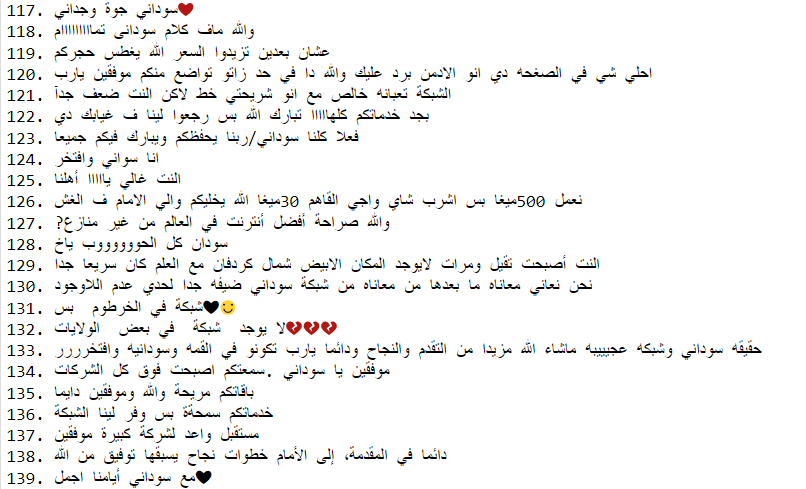
2- تنظيف البيانات :

تم تنظيف التعليقات لازالة الايموجي و التشكيل و الاحرف الانجليزية و الارقام (normalization)باستخدام مكتبة(aranorm)

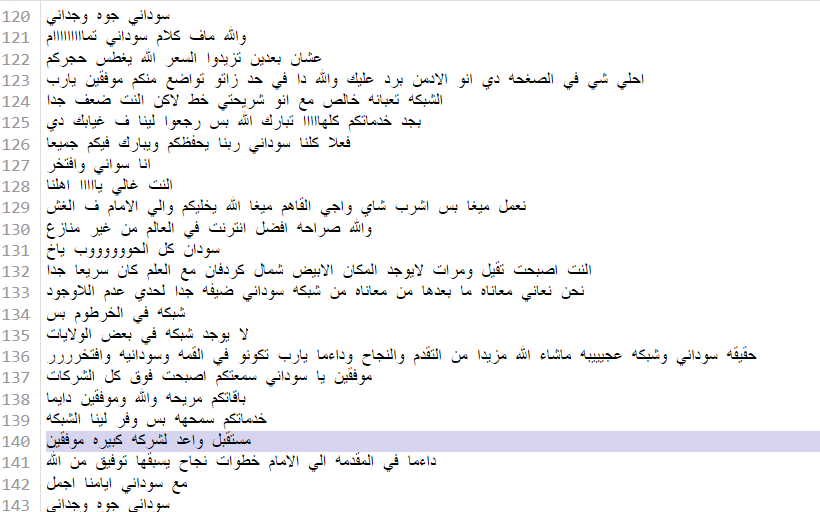
**الكود:**



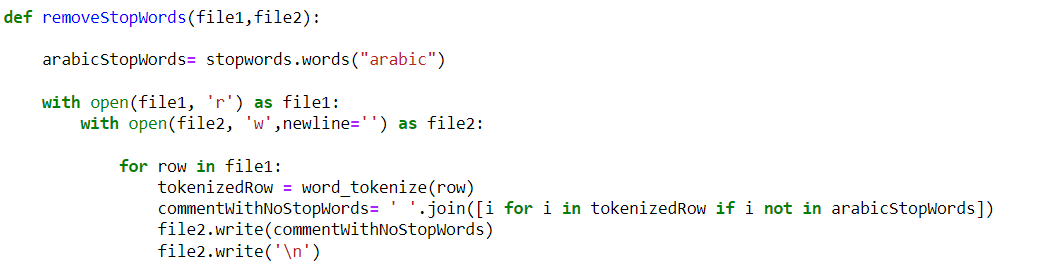
التعليقات :



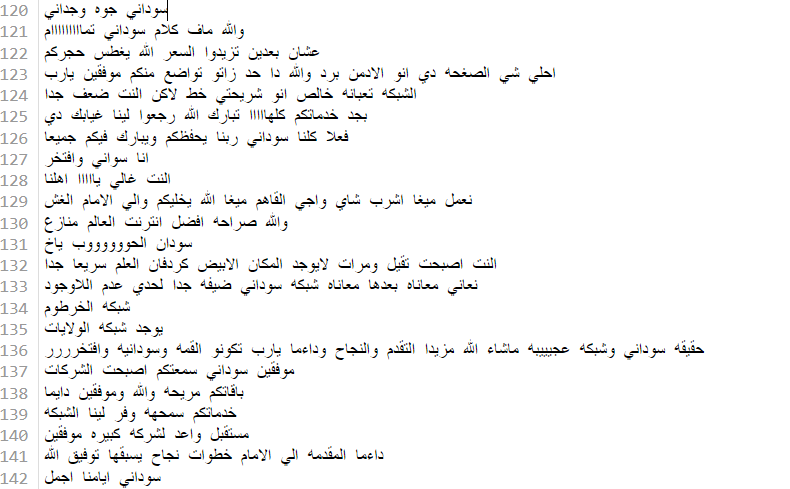
**التعليقات بعد المعالجة:**



ثم تم استحدام مكتبة (nltk) لازالة كلمات التوقف :



التعليقات بعد ازالة كلمات التوقف :

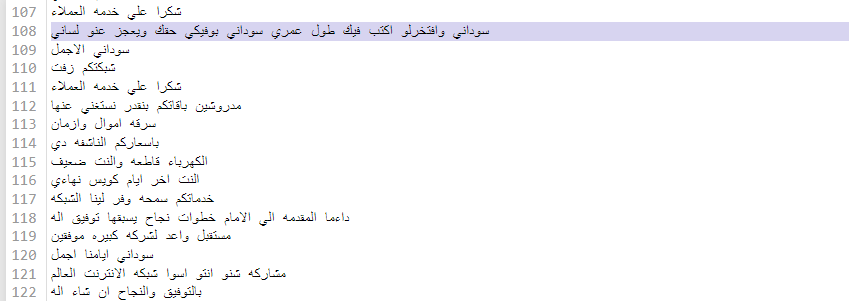
3-ازالة التكرار من الاحرف و الكلمات تم باستخدام مكتبة(re, itertools):



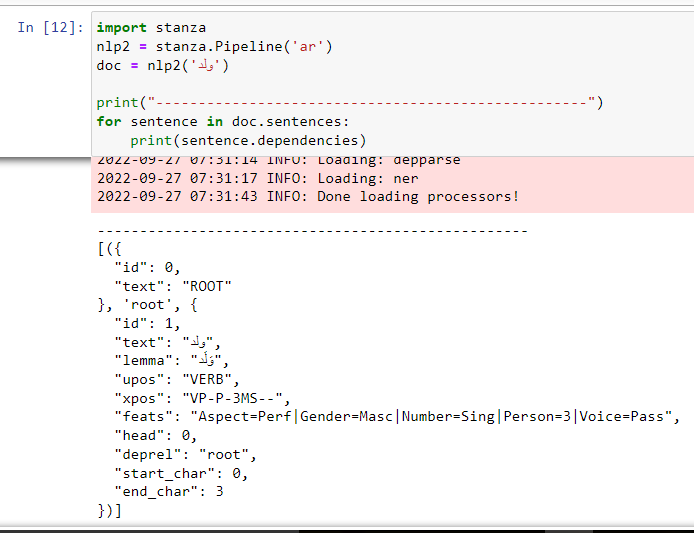
**التعليق**



التعليق بعد ازالة التكرار

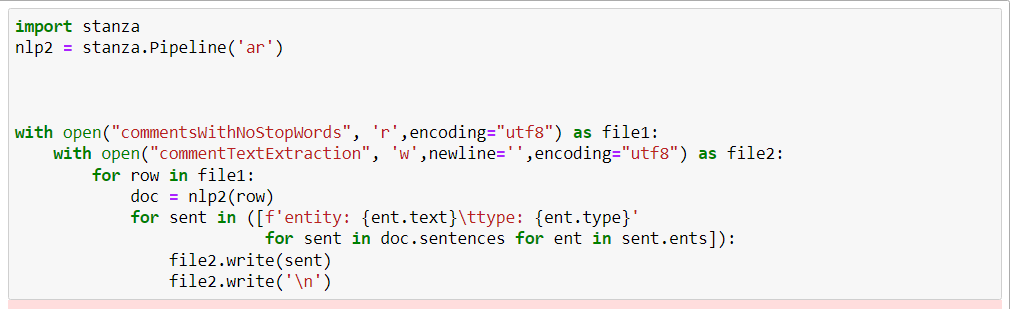


4-استخراج النصوص : تم استخدام ميزة استخراج النصوص في مكتبة (stanza) التي تعطي نتائج دقيقة باللغة الانجليزية

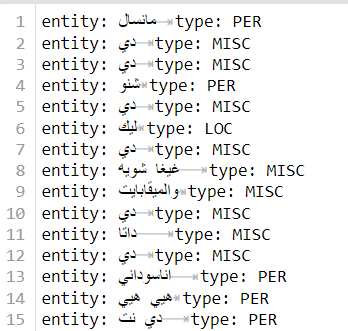


تم اخذ معلومات عن كل وحدة ونوعها اظهرت النتائج 297 كلمة من اصل حوالي 2500 كلمة مميزة كما اخطأت النتائج في اغلب انواع الوحدات .

**الكود:**

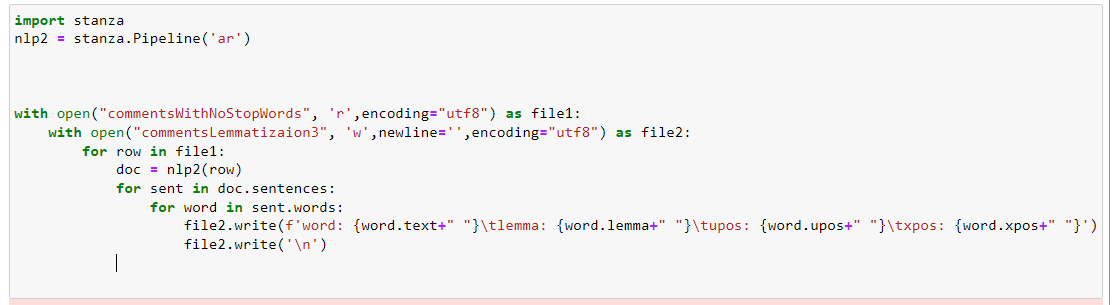


نتيجة استخراج النصوص من التعليقات :

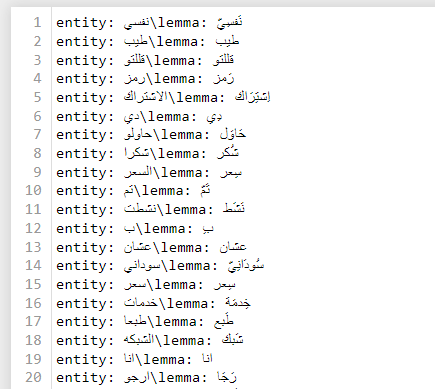


* نلاحظ النتائج الخاطئة بالنسبة للغة العربية

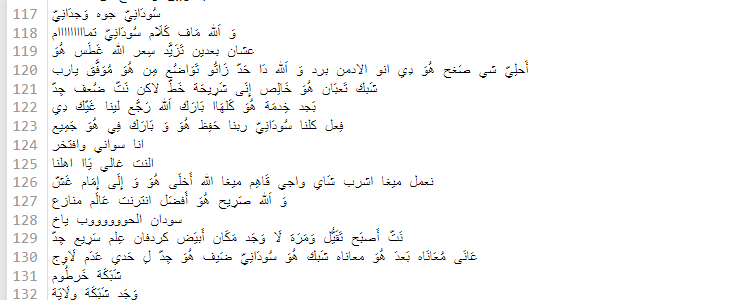
5-استخراج الجذور (lemmatization) تم باستخدام مكتبة(stanza) التي اعطت نتائج جيدة



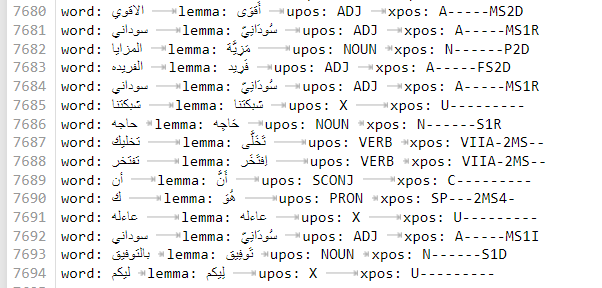
جذور الكلمات :



التعيقات باستخدام الجذور عوضا عن الكلمات الاصلية:



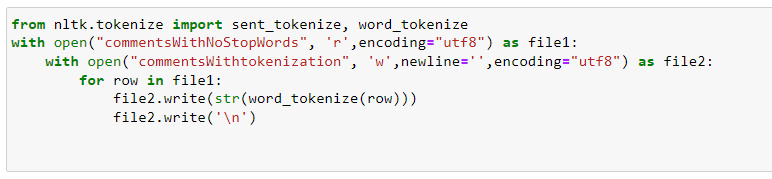
6-احدى خطوات ال (text extraction) هي استخراج ال (Upos , Xpos) تم ذلك باستخدام مكتبة(stanza)



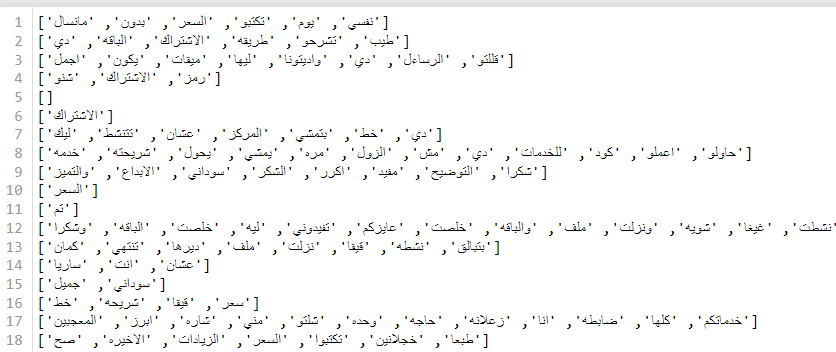
التي اعطت نتائج خاطئة للغة العربية

7-الترميز (tokenization)تم باستخدام مكتبة (nltk) استخراج الترميز الاحادي للتعليقات و الثنائي و الثلاثي كل في ملف منفصل.

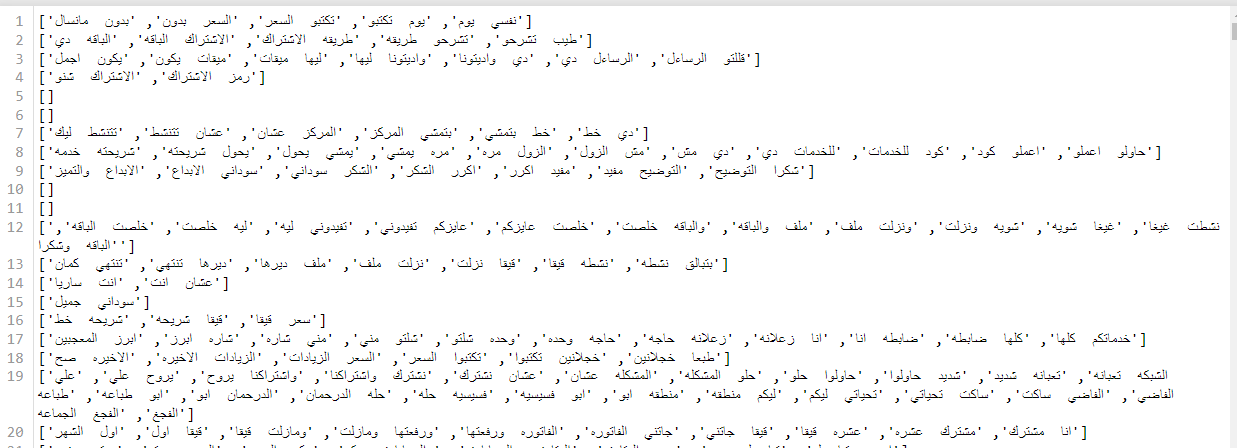
الكود:



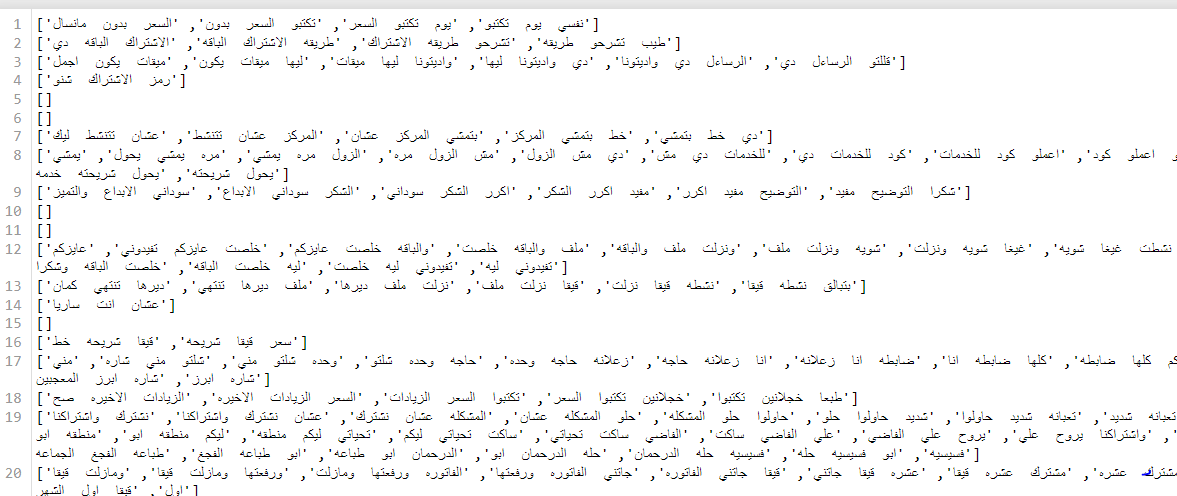
الترميز الاحادي:



الترميزالثنائي :



الترميز الثلاثي:

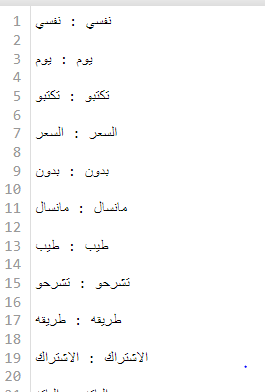


8- الاشتقاق(stemming) تم باستخدام مكتبة (nltk) التي تعطي نتائج صحيحة للغة الانجليزية و لكن النتائج خاطئة للغة العربية حيث انها ترجع نفس الكلمة

الكود:



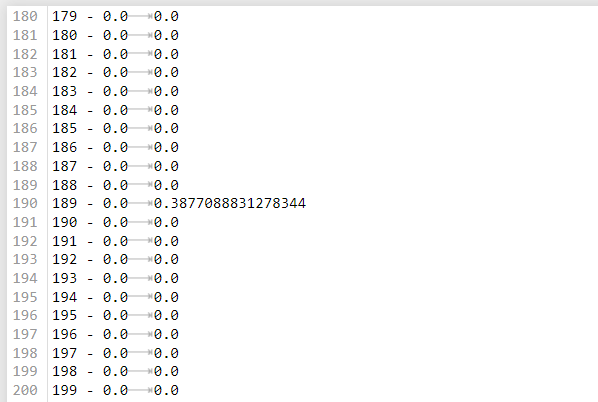
النتيجة:



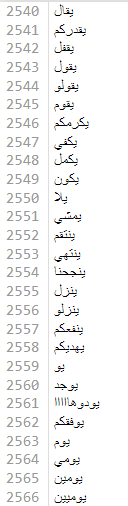
9-تم استخراج ال(TF-IDF)باستخدام مكتبة (sklearn)



النتيجة:



كما تم استخدام (tfidfvectorizer) لاستخراج اسماء المفردات المميزة حيث بلغت 2566مفردة

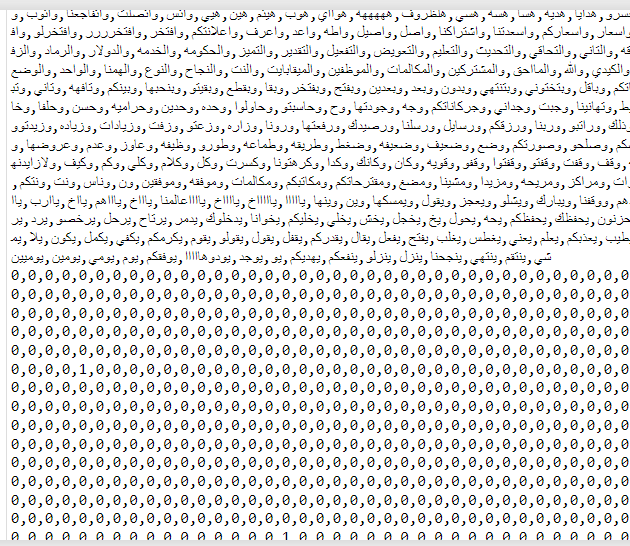


10-تم استخراج(BOW)باستخدام مكتبة(sklearn)

الكود:



النتيجة:



كما تم استخراج الكلمات المفردة المميزة عن طريق(countVectoriser)



والنتيجة كانت مشابهة لسابقتها

