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:

```
from aranorm import normalize_arabic_text
from nltk.corpus import stopwords
from nltk import word_tokenize
import pandas as pd


def normlizeComment(file1,file2):

    with open(file1, 'r') as file1:
        with open(file2, 'w',newline='') as file2:
            for row in file1:


                file2.write(normalize_arabic_text(row))
                file2.write('\n')
```

التعليقات :

117. ❤سوداني جوة وجداني
118. والله ماف كلام سوداني تماااااااام
119. عشان بعدين تزيدوا السعر الله يغطس حجركم
120. احلي شي في الصفحه دي انو الادمن برد عليك والله دا في حد زاتو تواضع منكم موفقين يارب
121. الشبكة تعبانه خالص مع انو شريحتي خط لاكن النت ضعف جداً
122. بجد خدماتكم كلهاااا تبارك الله بس رجعوا لينا ف غيابك دي
123. فعلا كلنا سوداني/ربنا يحفظكم ويبارك فيكم جميعا
124. انا سواني وافتخر
125. النت غالي يااااا أهلنا
126. نعمل 500ميغا بس اشرب شاي واجي القاهم 30ميغا الله يخليكم والي الامام ف الغش
127. والله صراحة أفضل أنترنت في العالم من غير منازع؟
128. سودان كل الحوووووووب ياخ
129. النت أصبحت تقيل ومرات لايوجد المكان الابيض شمال كردفان مع العلم كان سريعا جدا
130. نحن نعاني معاناه ما بعدها من معاناه من شبكة سوداني ضيفه جدا لحدي عدم اللاوجود
131. شبكة في الخرطوم بس ❤
132. لا يوجد شبكة في بعض الولايات ❤❤❤
133. حقيقه سوداني وشبكه عجيبيبه ماشاء الله مزيدا من التقدم والنجاح ودائما يارب تكونو في القمه وسودانيه وافتخررررر
134. موفقين يا سوداني .سمعتكم اصبحت فوق كل الشركات
135. باقاتكم مريحة والله وموفقين دايما
136. خدماتكم سمحةة بس وفر لينا الشبكة
137. مستقبل واعد لشركة كبيرة موفقين
138. دائما في المقدمة، إلى الأمام خطوات نجاح يسبقها توفيق من الله
139. مع سوداني أيامنا اجمل ❤

التعليقات بعد المعالجة:

120 سوداني جوه وجداني
121 والله ماف كلام سوداني تماااااااام
122 عشان بعدين تزيدوا السعر الله يغطس حجركم
123 احلي شي في الصفحه دي انو الادمن برد عليك والله دا في حد زاتو تواضع منكم موفقين يارب
124 الشبكه تعبانه خالص مع انو شريحتي خط لاكن النت ضعف جدا
125 بجد خدماتكم كلهاااا تبارك الله بس رجعوا لينا ف غيابك دي
126 فعلا كلنا سوداني ربنا يحفظكم ويبارك فيكم جميعا
127 انا سواني وافتخر
128 النت غالي يااااا اهلنا
129 نعمل ميغا بس اشرب شاي واجي القاهم ميغا الله يخليكم والي الامام ف الغش
130 والله صراحه افضل انترنت في العالم من غير منازع
131 سودان كل الحوووووووب ياخ
132 النت اصبحت تقيل ومرات لايوجد المكان الابيض شمال كردفان مع العلم كان سريعا جدا
133 نحن نعاني معاناه ما بعدها من معاناه من شبكه سوداني ضيفه جدا لحدي عدم اللاوجود
134 شبكه في الخرطوم بس
135 لا يوجد شبكه في بعض الولايات
136 حقيقه سوداني وشبكه عجيبيبه ماشاء الله مزيدا من التقدم والنجاح وداعما يارب تكونو في القمه وسودانيه وافتخررررر
137 موفقين يا سوداني سمعتكم اصبحت فوق كل الشركات
138 باقاتكم مريحه والله وموفقين دايما
139 خدماتكم سمحهه بس وفر لينا الشبكه
140 مستقبل واعد لشركه كبيره موفقين
141 داعما في المقدمه الي الامام خطوات نجاح يسبقها توفيق من الله
142 مع سوداني ايامنا اجمل
143 سوداني جوه وجداني

ثم تم استحدام مكتبة (nltk) لازالة كلمات التوقف :

```
def removeStopWords(file1,file2):

    arabicStopWords= stopwords.words("arabic")

    with open(file1, 'r') as file1:
        with open(file2, 'w',newline='') as file2:

            for row in file1:
                tokenizedRow = word_tokenize(row)
                commentWithNoStopWords= ' '.join([i for i in tokenizedRow if i not in arabicStopWords])
                file2.write(commentWithNoStopWords)
                file2.write('\n')
```

التعليقات بعد ازالة كلمات التوقف :

```
120 سوداني جوه وجداني
121 والله ماف كلام سوداني تماااااااام
122 عشان بعدين تزيدوا السعر الله يغطس حجركم
123 احلي شي الصغحه دي انو الادمن برد والله دا حد زاتو تواضع منكم موفقين يارب
124 الشبكه تعبانه خالص انو شريحتي خط لاكن النت ضعف جدا
125 بجد خدماتكم كلهاااا تبارك الله رجعوا لينا غيابك دي
126 فعلا كلنا سوداني ربنا يحفظكم ويبارك فيكم جميعا
127 انا سواني وافتخر
128 النت غالي ياااا اهلنا
129 نعمل ميغا اشرب شاي واجي القاهم ميغا الله يخليكم والي الامام الغش
130 والله صراحه افضل انترنت العالم منازع
131 سودان الجووووووب ياخ
132 النت اصبحت تقيل ومرات لايوجد المكان الابيض كردفان العلم سريعا جدا
133 نعاني معاناه بعدها معاناه شبكه سوداني ضيفه جدا لحدي عدم اللاوجود
134 شبكه الخرطوم
135 يوجد شبكه الولايات
136 حقيقه سوداني وشبكه عجيبيبه ماشاء الله مزيدا التقدم والنجاح وداعما يارب تكونو القمه وسودانيه وافتخررررر
137 موفقين سوداني سمعتكم اصبحت الشركات
138 باقاتكم مريحه والله وموفقين دايما
139 خدماتكم سمحهه وفر لينا الشبكه
140 مستقبل واعد لشركه كبيره موفقين
141 داعما المقدمه الي الامام خطوات نجاح يسبقها توفيق الله
142 سوداني ايامنا اجمل
```

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

```
import stanza
nlp2 = stanza.Pipeline('ar')


def remove_duplicated_characters(text):
    result=[]
    text=''.join(i for i,_ in itertools.groupby(text))
    result.append(text)
    return ''.join(result)


def remove_duplicated_words(text):
    s1=[]
    text=' '.join(k for k,v in itertools.groupby(text.replace("&lt;/Sent&gt;","").split()))
    s=re.sub(r'b(.+)(\s+\b)+',r'\1',text)
    return ''.join(s)


with open("commentsWithNoStopWords", 'r',encoding="utf8") as file1:
    with open("commentsWithNoDuplicated", 'w',newline='',encoding="utf8") as file2:
        for row in file1:
            phace1=remove_duplicated_characters(row)
            phace2=remove_duplicated_words(phace1)
            file2.write(phace2)
            file2.write('\n')
```

التعليق

```
100. شكراً على خدمة العملاء
101. سوووووداني وافتخرلو اكتب فيك طوووول عمري يا سوداني م بوفيكي حقك وبعجز عنو لساني
102. سوداني الاجمل
103. شبكتكم زفت
104. شكراً على خدمة العملاء
105. مدروشين بس باقاتكم م بنقدر نستغني عنها 😍❤
106. سرقه اموال وازمان
107. باسعاركم الناشفه دي 😒
108. الكهرباء قاطعه والنت ضعيف
109. النت اخر أيام كان ما كويس نهائي
110. خدماتكم سمحة بس وفر لينا الشبكة
111. دائما في المقدمة، إلى الأمام خطوات نجاح يسبقها توفيق من الله
112. مستقبل واعد لشركة كبيرة موفقين
113. مع سوداني أيامنا اجمل
114. مشاركة شنو أنتو أسوأ شبكة في الإنترنت في العالم
115. "بالتوفيق والنجاح ان شاء الله"
116. "
117. سوداني جوة وجداني❤
118. والله ماف كلام سوداني تماااااام
119. عشان بعدين تزيدوا السعر الله يخطس حجركم
120. احلى شي في الصفحه دي انو الادمن برد عليك والله دا في حد زاتو تواضع منكم موفقين يارب
```

التعليق بعد ازالة التكرار

```
107 شكرا علي خدمه العملاء
108 سوداني وافتخرلو اكتب فيك طول عمري سوداني بوفيكي حقك وبعجز عنو لساني
109 سوداني الاجمل
110 شبكتكم زفت
111 شكرا علي خدمه العملاء
112 مدروشين باقاتكم بنقدر نستغني عنها
113 سرقه اموال وازمان
114 باسعاركم الناشفه دي
115 الكهرباء قاطعه والنت ضعيف
116 النت اخر ايام كويس نهاءي
117 خدماتكم سمحه وفر لينا الشبكه
118 داءما المقدمه الي الامام خطوات نجاح يسبقها توفيق اله
119 مستقبل واعد لشركه كبيره موفقين
120 سوداني ايامنا اجمل
121 مشاركه شنو انتو اسوا شبكه الانترنت العالم
122 بالتوفيق والنجاح ان شاء اله
```

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

```
In [12]: import stanza
         nlp2 = stanza.Pipeline('ar')
         doc = nlp2('ولد')

         print("----------------------------------------------------------")
         for sentence in doc.sentences:
             print(sentence.dependencies)
2022-09-27 07:31:14 INFO: Loading: depparse
2022-09-27 07:31:17 INFO: Loading: ner
2022-09-27 07:31:43 INFO: Done loading processors!
----------
[({
  "id": 0,
  "text": "ROOT"
}, 'root', {
  "id": 1,
  "text": "ولد",
  "lemma": "وَلَد",
  "upos": "VERB",
  "xpos": "VP-P-3MS--",
  "feats": "Aspect=Perf|Gender=Masc|Number=Sing|Person=3|Voice=Pass",
  "head": 0,
  "deprel": "root",
  "start_char": 0,
  "end_char": 3
})]
```

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:

```
import stanza
nlp2 = stanza.Pipeline('ar')



with open("commentsWithNoStopWords", 'r',encoding="utf8") as file1:
    with open("commentTextExtraction", 'w',newline='',encoding="utf8") as file2:
        for row in file1:
            doc = nlp2(row)
            for sent in ([f'entity: {ent.text}\ttype: {ent.type}'
                          for sent in doc.sentences for ent in sent.ents]):
                file2.write(sent)
                file2.write('\n')
```

نتيجة استخراج النصوص من التعليقات :

```
entity: مانسال	type: PER
entity: دي	type: MISC
entity: دي	type: MISC
entity: شنو	type: PER
entity: دي	type: MISC
entity: ليك	type: LOC
entity: دي	type: MISC
entity: غيغا شويه	type: MISC
entity: والميقابايت	type: MISC
entity: دي	type: MISC
entity: داتا	type: MISC
entity: دي	type: MISC
entity: اناسوداني	type: PER
entity: هيي هيي	type: PER
entity: دي نت	type: PER
```

- نلاحظ النتائج الخاطئة بالنسبة للغة العربية

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

```
import stanza
nlp2 = stanza.Pipeline('ar')



with open("commentsWithNoStopWords", 'r',encoding="utf8") as file1:
    with open("commentsLemmatizaion3", 'w',newline='',encoding="utf8") as file2:
        for row in file1:
            doc = nlp2(row)
            for sent in doc.sentences:
                for word in sent.words:
                    file2.write(f'word: {word.text+" "}\tlemma: {word.lemma+" "}\tupos: {word.upos+" "}\txpos: {word.xpos+" "}')
                    file2.write('\n')
```

جذور الكلمات :

```
entity: نفسي\lemma: نَفسِيّ
entity: طيب\lemma: طَيِب
entity: قللتو\lemma: قَلَلتو
entity: رمز\lemma: رَمز
entity: الاشتراك\lemma: إشتِرَاك
entity: دي\lemma: دي
entity: حاولو\lemma: حَاوَل
entity: شكرا\lemma: شُكر
entity: السحر\lemma: سِحر
entity: تم\lemma: تَمّ
entity: نشطت\lemma: نَشَط
entity: ب\lemma: بِ
entity: عشان\lemma: عشان
entity: سوداني\lemma: سُودَانِيّ
entity: سحر\lemma: سِحر
entity: خدمات\lemma: خِدمَة
entity: طبعا\lemma: طَبِع
entity: الشبكه\lemma: شَبَك
entity: انا\lemma: انا
entity: ارجو\lemma: رَجَا
```

التعيقات باستخدام الجذور عوضا عن الكلمات الاصلية:

```
117 سُودَانِيّ جوه وَجدَانِيّ
118 وَ أَللّه مَاف كَلَام سُودَانِيّ تماااااااام
119 عشان بعدين تَزَيَّد سِحر الله غَطَّس هُوَ
120 أَحلِيّ شي صَنَع هُوَ دِي انو الادمن برد وَ أَللّه دَا حَدّ زَائُو تَوَاضُع مِن هُوَ مُوَفَّق يارب
121 شَبَك تَعبَان هُوَ خَالِص إِنِى شَرِيحَة خَطّ لاكن نَتّ ضُعف جِدّ
122 يَجد خِدمَة هُوَ كَلَهَا بَارَك أَللّه رَجَّع لينا غَيَّك دِي
123 فِعل كلنا سُودَانِيّ ربنا حَفِظ هُوَ وَ بَارَك فِي هُوَ جَمِيع
124 انا سواني وافتخر
125 النت غالي يَا اهلنا
126 نعمل ميغا اشرب شَاي واجي قَاهِم ميغا الله أَخلَى هُوَ وَ إِلَى إِمَام عَمّ
127 وَ أَللّه صَرِيح هُوَ أَفضَل انترنت عَالَم منازع
128 سودان الحوووووب ياخ
129 نَتّ أَصبَح ثَقِيل وَمَرَّة لَا وَجَد مَكَان أَبيَض كردفان عِلم سَرِيع جِدّ
130 عَالَى مُعَانَاه بَعدَ هُوَ معاناه شَبَك هُوَ سُودَانِيّ ضَيف هُوَ جِدّ لِ حَدي عَدَم لَاوِج
131 شَبَكَة خَرطُوم
132 وَجَد شَبَكَة وَلَايَة
```

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

```
word: الاقوي  lemma: أَقوَى  upos: ADJ  xpos: A-----MS2D
word: سوداني  lemma: سُودَانِيّ  upos: ADJ  xpos: A-----MS1R
word: المزايا  lemma: مَزِيَّة  upos: NOUN  xpos: N------P2D
word: الفريده  lemma: فَرِيد  upos: ADJ  xpos: A-----FS2D
word: سوداني  lemma: سُودَانِيّ  upos: ADJ  xpos: A-----MS1R
word: شيكتنا  lemma: شيكتنا  upos: X  xpos: U---------
word: حاجه  lemma: حَاجِه  upos: NOUN  xpos: N------S1R
word: تخليك  lemma: تَخَلَّى  upos: VERB  xpos: VIIA-2MS--
word: تفتخر  lemma: إفتَخَر  upos: VERB  xpos: VIIA-2MS--
word: أن  lemma: أَنَّ  upos: SCONJ  xpos: C---------
word: ك  lemma: هُوَ  upos: PRON  xpos: SP---2MS4-
word: عاءله  lemma: عاءله  upos: X  xpos: U---------
word: سوداني  lemma: سُودَانِيّ  upos: ADJ  xpos: A-----MS1I
word: بالتوفيق  lemma: تَوفِيق  upos: NOUN  xpos: N------S1D
word: ليكم  lemma: لِيكم  upos: X  xpos: U---------
```

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:

```
from nltk.tokenize import sent_tokenize, word_tokenize
with open("commentsWithNoStopWords", 'r',encoding="utf8") as file1:
    with open("commentsWithtokenization", 'w',newline='',encoding="utf8") as file2:
        for row in file1:
            file2.write(str(word_tokenize(row)))
            file2.write('\n')
```

الترميز الاحادي:

```
['نفسي', 'يوم', 'تكتبو', 'السعر', 'بدون', 'مانسال']
['طيب', 'تشرحو', 'طريقه', 'الاشتراك', 'الباقه', 'دي']
['قللتو', 'الرساءل', 'دي', 'واديتونا', 'ليها', 'ميقات', 'يكون', 'اجمل']
['رمز', 'الاشتراك', 'تنو']
[]
['الاشتراك']
['دي', 'خط', 'يتمشي', 'المركز', 'عشان', 'تتنشط', 'ليك']
['حاولو', 'اعملو', 'كود', 'للخدمات', 'دي', 'متس', 'الزول', 'مره', 'يمشي', 'يحول', 'شريحته', 'خدمه']
['شكرا', 'التوضيح', 'مفيد', 'اكرر', 'الشكر', 'سوداني', 'الابداع', 'والتميز']
['السعر']
['تم']
['نشطت', 'غيغا', 'شويه', 'ونزلت', 'ملف', 'والباقه', 'خلصت', 'عايزكم', 'تفيدوني', 'ليه', 'خلصت', 'الباقه', 'وشكرا']
['بتبالق', 'نشطه', 'قيقا', 'نزلت', 'ملف', 'ديرها', 'تنتهي', 'كمان']
['عشان', 'انت', 'ساريا']
['سوداني', 'جميل']
['سعر', 'قيقا', 'شريحه', 'خط']
['خدماتكم', 'كلها', 'ضابطه', 'انا', 'زعلانه', 'حاجه', 'وحده', 'تلتو', 'مني', 'شاره', 'ابرز', 'المعجبين']
['طبعا', 'خجلانين', 'تكتبوا', 'السعر', 'الزيادات', 'الاخيره', 'صح']
```

الترميز الثنائي :

```
['نفسي يوم', 'يوم تكتبو', 'تكتبو السعر', 'السعر بدون', 'بدون مانسال']
['طيب تشرحو', 'تشرحو طريقه', 'طريقه الاشتراك', 'الاشتراك الباقه', 'الباقه دي']
['قللتو الرساءل', 'الرساءل دي', 'دي واديتونا', 'واديتونا ليها', 'ليها ميقات', 'ميقات يكون', 'يكون اجمل']
['رمز الاشتراك', 'الاشتراك شنو']
[]
[]
['دي خط', 'خط بتمشي', 'بتمشي المركز', 'المركز عشان', 'عشان تتنشط', 'تتنشط ليك']
['حاولو اعملو', 'اعملو كود', 'كود للخدمات', 'للخدمات دي', 'دي متش', 'متش الزول', 'الزول مره', 'مره بمشي', 'بمشي بحول', 'بحول شريحته', 'شريحته خدمه']
['شكرا التوضيح', 'التوضيح مفيد', 'مفيد اكرر', 'اكرر الشكر', 'الشكر سوداني', 'سوداني الابداع', 'الابداع والتميز']
[]
[]
['نشطت غيغا', 'غيغا شويه', 'شويه ونزلت', 'ونزلت ملف', 'ملف والباقه', 'والباقه خلصت', 'خلصت عايزكم', 'عايزكم تفيدوني', 'تفيدوني ليه', 'ليه خلصت', 'خلصت الباقه', 'الباقه وشكرا']
['بتبالق نشطه', 'نشطه قيقا', 'قيقا نزلت', 'نزلت ملف', 'ملف ديرها', 'ديرها تنتهي', 'تنتهي كمان']
['عشان انت', 'انت ساريا']
['سوداني جميل']
['سعر قيقا', 'قيقا شريحه', 'شريحه خط']
['خدماتكم كلها', 'كلها ضابطه', 'ضابطه انا', 'انا زعلانه', 'زعلانه حاجه', 'حاجه وحده', 'وحده تلتو', 'تلتو مني', 'مني شاره', 'شاره ابرز', 'ابرز المعجبين']
['طبعا خجلانين', 'خجلانين تكتبوا', 'تكتبوا السعر', 'السعر الزيادات', 'الزيادات الاخيره', 'الاخيره صح']
['التبكه تعبانه', 'تعبانه شديد', 'شديد حاولوا', 'حاولوا حلو', 'حلو المشكله', 'المشكله عشان', 'عشان نشترك', 'نشترك واشتراكنا', 'واشتراكنا يروح', 'يروح علي', 'علي الفاضي', 'الفاضي ساكت', 'ساكت تحياتي', 'تحياتي ليكم', 'ليكم منطقه', 'منطقه ابو', 'ابو فسيسيه', 'فسيسيه حله', 'حله الدرحمان', 'الدرحمان ابو', 'ابو طباعه', 'طباعه الفجع', 'الفجع الجماعه']
['انا مشترك', 'مشترك عشره', 'عشره قيقا', 'قيقا جاتني', 'جاتني الفاتوره', 'الفاتوره ورفعتها', 'ورفعتها ومازلت', 'ومازلت قيقا', 'قيقا اول', 'اول الشهر']
```

الترميز الثلاثي:

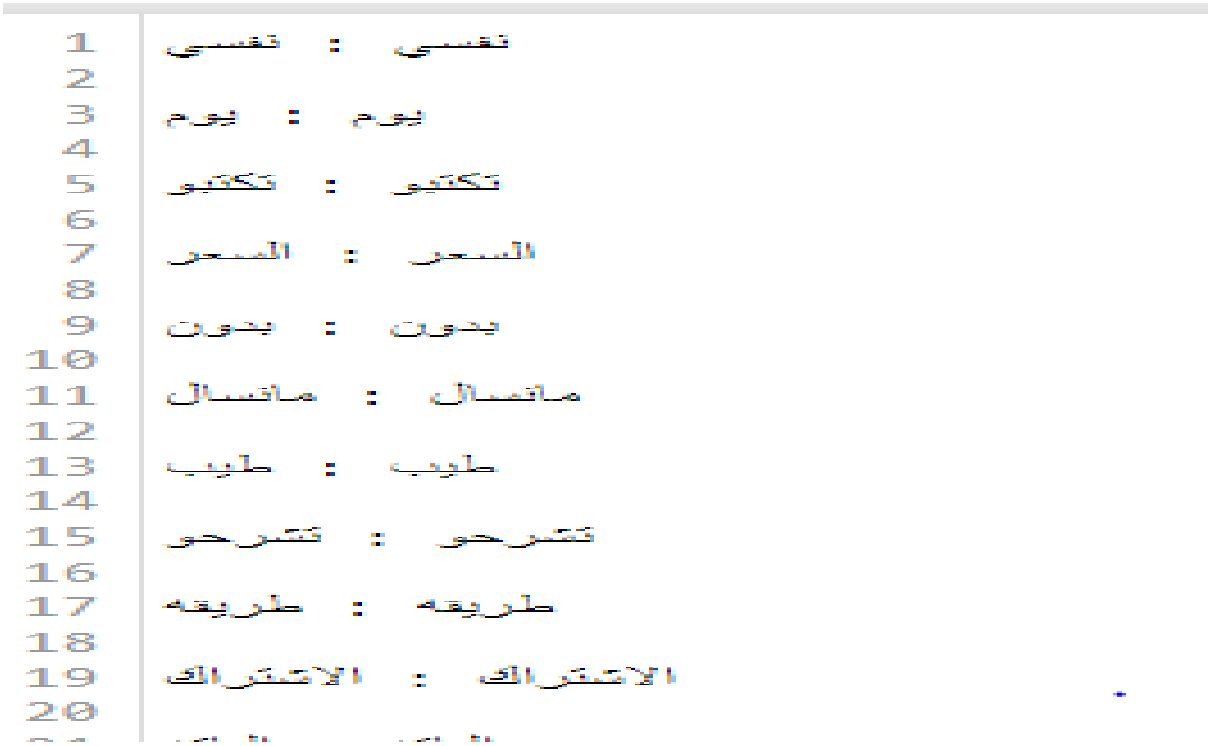
```
['نفسي يوم تكتبو', 'يوم تكتبو السعر', 'تكتبو السعر بدون', 'السعر بدون مانسال']
['طيب تشرحو طريقه', 'تشرحو طريقه الاشتراك', 'طريقه الاشتراك الباقه', 'الاشتراك الباقه دي']
['قللتو الرساءل دي', 'الرساءل دي واديتونا', 'دي واديتونا ليها', 'واديتونا ليها ميقات', 'ليها ميقات يكون', 'ميقات يكون اجمل']
['رمز الاشتراك شنو']
[]
[]
['دي خط بتمشي', 'خط بتمشي المركز', 'بتمشي المركز عشان', 'المركز عشان تتنشط', 'عشان تتنشط ليك']
['و اعملو كود', 'اعملو كود للخدمات', 'كود للخدمات دي', 'للخدمات دي متش', 'دي متش الزول', 'متش الزول مره', 'الزول مره بمشي', 'مره بمشي بحول', 'بمشي بحول شريحته', 'بحول شريحته خدمه']
['شكرا التوضيح مفيد', 'التوضيح مفيد اكرر', 'مفيد اكرر الشكر', 'اكرر الشكر سوداني', 'الشكر سوداني الابداع', 'سوداني الابداع والتميز']
[]
[]
['نشطت غيغا شويه', 'غيغا شويه ونزلت', 'شويه ونزلت ملف', 'ونزلت ملف والباقه', 'ملف والباقه خلصت', 'والباقه خلصت عايزكم', 'خلصت عايزكم تفيدوني', 'عايزكم تفيدوني ليه', 'تفيدوني ليه خلصت', 'ليه خلصت الباقه', 'خلصت الباقه وشكرا']
['بتبالق نشطه قيقا', 'نشطه قيقا نزلت', 'قيقا نزلت ملف', 'نزلت ملف ديرها', 'ملف ديرها تنتهي', 'ديرها تنتهي كمان']
['عشان انت ساريا']
[]
['سعر قيقا شريحه', 'قيقا شريحه خط']
['م كلها ضابطه', 'كلها ضابطه انا', 'ضابطه انا زعلانه', 'انا زعلانه حاجه', 'زعلانه حاجه وحده', 'حاجه وحده تلتو', 'وحده تلتو مني', 'تلتو مني شاره', 'مني شاره ابرز', 'شاره ابرز المعجبين']
['طبعا خجلانين تكتبوا', 'خجلانين تكتبوا السعر', 'تكتبوا السعر الزيادات', 'السعر الزيادات الاخيره', 'الزيادات الاخيره صح']
['تعبانه شديد', 'شديد حاولوا', 'حاولوا حلو', 'حلو المشكله', 'المشكله عشان', 'عشان نشترك', 'نشترك واشتراكنا', 'واشتراكنا يروح', 'يروح علي', 'علي الفاضي', 'الفاضي ساكت', 'ساكت تحياتي', 'تحياتي ليكم', 'ليكم منطقه', 'منطقه ابو', 'ابو فسيسيه', 'فسيسيه حله', 'حله الدرحمان', 'الدرحمان ابو', 'ابو طباعه', 'طباعه الفجع', 'الفجع الجماعه']
['شترك عشره', 'مشترك عشره قيقا', 'عشره قيقا جاتني', 'قيقا جاتني الفاتوره', 'جاتني الفاتوره ورفعتها', 'الفاتوره ورفعتها ومازلت', 'ورفعتها ومازلت قيقا', 'ومازلت قيقا اول', 'قيقا اول الشهر']
```

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:

```
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize

ps = PorterStemmer()

with open("commentsWithNoStopWords", 'r',encoding="utf8") as file1:
    with open("commentsWithstemming", 'w',newline='',encoding="utf8") as file2:
        for row in file1:
            tokens = word_tokenize(row)

            stemmed = []

            for token in tokens:
                stemmed_word = ps.stem(token)
                file2.write(f'{token} : {stemmed_word}'+"\n")
                file2.write('\n')
```

النتيجة:



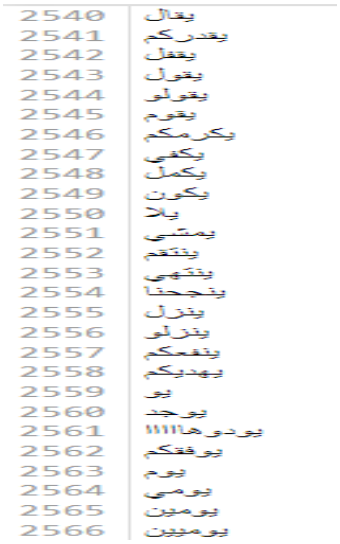
9-تم استخراج ال(TF-IDF)باستخدام مكتبة (sklearn)

```
import pandas as pd
with open("commentsWithNoStopWords", 'r',encoding="utf8") as file1:

    vectorizer = TfidfVectorizer()
    vectors = vectorizer.fit_transform(file1)
    feature_names = vectorizer.get_feature_names_out()
    dense = vectors.todense()
    denselist = dense.tolist()

    with open("TFIDFDENSE", 'w',encoding="utf8") as file2:
        for i in range(len(denselist)):

            file2.write(str(i)+" - "+ str(denselist[0][i])+ "\t"+ str(denselist[1][i]))
            file2.write("\n")
    file2.close()


    df = pd.DataFrame(denselist, columns=feature_names)
```

النتيجة:

```
179 - 0.0	0.0
180 - 0.0	0.0
181 - 0.0	0.0
182 - 0.0	0.0
183 - 0.0	0.0
184 - 0.0	0.0
185 - 0.0	0.0
186 - 0.0	0.0
187 - 0.0	0.0
188 - 0.0	0.0
189 - 0.0	0.3877088831278344
190 - 0.0	0.0
191 - 0.0	0.0
192 - 0.0	0.0
193 - 0.0	0.0
194 - 0.0	0.0
195 - 0.0	0.0
196 - 0.0	0.0
197 - 0.0	0.0
198 - 0.0	0.0
199 - 0.0	0.0
```

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

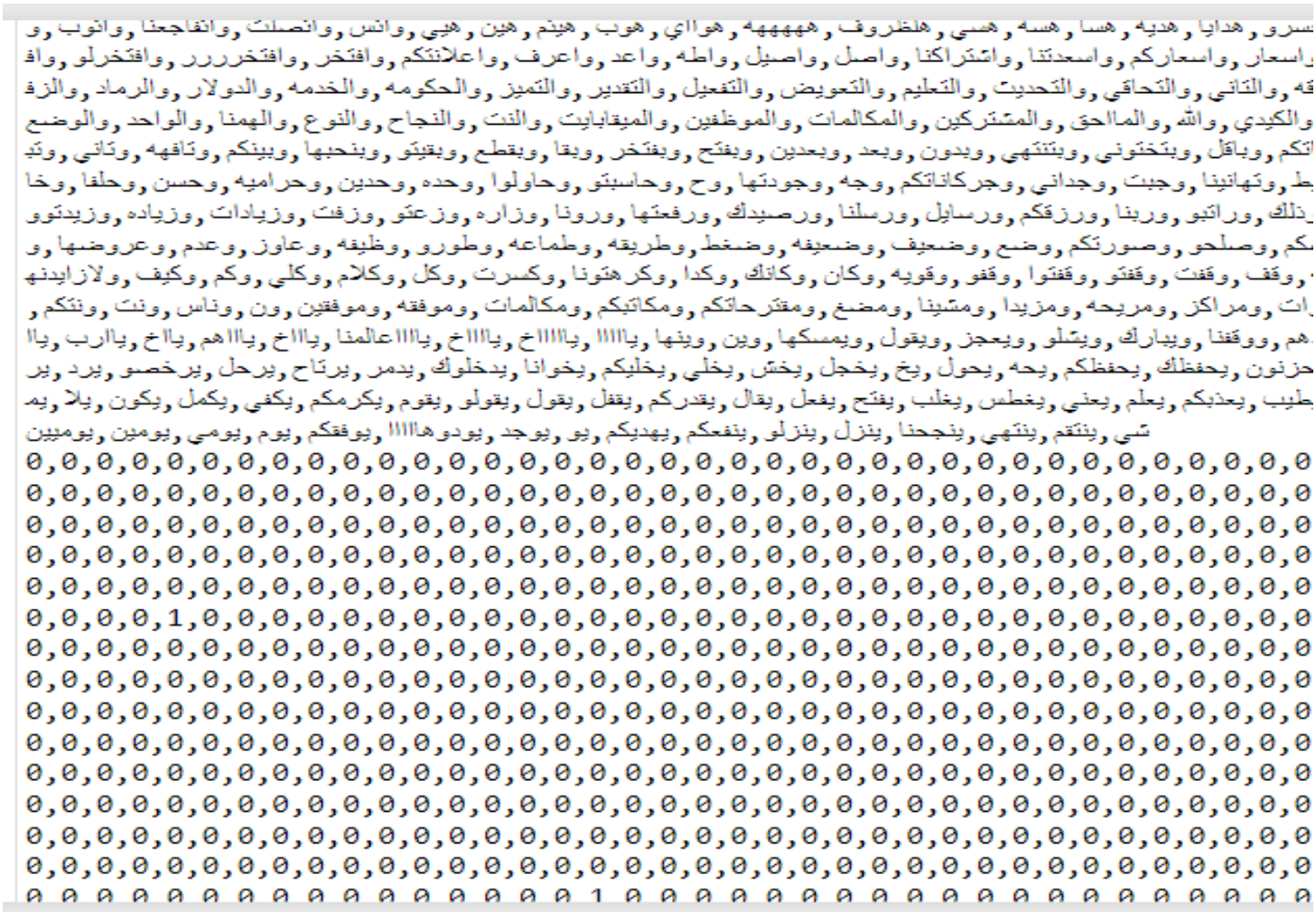


10-تم استخراج(BOW)باستخدام مكتبة(sklearn)

الكود:

```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer

wordset = [

]

with open("commentsWithNoStopWords", 'r',encoding="utf8") as file1:
    for row in file1:
            wordset.append(row)


vectorizer =  CountVectorizer()
X = vectorizer.fit_transform(wordset)

df_bow_sklearn = pd.DataFrame(X.toarray(),columns=vectorizer.get_feature_names_out())
df_bow_sklearn.head()

df_bow_sklearn.to_csv('BOW3.csv', encoding='utf-8')

file2.close()
print(X.shape)
```

النتيجة:



كما تم استخراج الكلمات المفردة المميزة عن طريق(countVectoriser)

```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer

wordset = [

]

with open("commentsWithNoStopWords", 'r',encoding="utf8") as file1:

    for row in file1:

            wordset.append(row)


vectorizer =  CountVectorizer()
X = vectorizer.fit_transform(wordset)
features=vectorizer.get_feature_names_out()
with open("featureWithBOW.csv", 'w',encoding="utf8") as file2:
    for i in features:
        file2.write(str(i+" \n"))


file2.close()
print(X.shape)
```

والنتيجة كانت مشابهة لسابقتها

2548	يكمل
2549	يكون
2550	يلا
2551	يمشي
2552	ينتقم
2553	ينتهي
2554	ينجحنا
2555	ينزل
2556	ينزلو
2557	ينفعكم
2558	يهديكم
2559	يو
2560	يوجد
2561	يودوهاااا
2562	يوفقكم
2563	يوم
2564	يومي
2565	يومين
2566	يومين