Tarea 1 - Recuperacion de tweets basado en contenidos

May 22, 2019

1. Funcion de limpieza

[1]:
```
import re
sw = open("stopwords.txt", 'r').read().split("\n")
def cleanSW(text):
    global sw
    text = re.sub(r'[^[a-zñA-ZÑáéíóúÁÉÍÓÚüÜ ]]*', "", text).replace("_"," ").
 ↪lower()
    clean = []
    for i in text.split(" "):
        if i not in sw:
            clean.append(i)
    return clean
```

2. Índice invertido

[2]:
```
import json
from pprint import pprint
with open("tweets.json") as f:
    data = json.load(f)
n_doc = len(data)
def invertedIndex(data):
    inv = {}
    ids = {}
    for i in data:
        text = cleanSW(i["text"])
        for word in text:
            if word in inv:
                #frec_words[word] = frec_words[word]+1
                if i["id"] in inv[word]:
                    inv[word][i["id"]] = inv[word][i["id"]]+1
                else:
                    inv[word][i["id"]] = 1
            else:
                inv[word] = {}
                inv[word][i["id"]] = 1
        ids[i['id']] = i["text"]
```

```
    for key in inv:
        inv[key]["df"] = len(inv[key])
    return inv, ids
inv, ids = invertedIndex(data)
```

3. Implementación de TF-IDF

[3]:
```
import math
def tf_idf(n_doc,word, inv,_id):
    if word not in inv:
        return 0
    idf = math.log10( n_doc/inv[word]["df"] )
    return idf * math.log10(1+inv[word][_id])
```

4. Normalize tweets

[4]:
```
def Normalize(vector):
    norm = 0
    for v in vector:
        norm += vector[v]*vector[v]
    norm = math.sqrt(norm)
    for i in vector:
        vector[i] /= norm
    return vector
def generateTable():
    global ids
    table = {}
    for _id in ids:
        text = cleanSW(ids[_id])
        frec = {}
        table[_id] = {}
        for word in text:
            if word in frec:
                frec[word] += 1
            else:
                frec[word] = 0
        for w in frec:
            if w not in table:
                table[_id][w] = {}
            table[_id][w] = tf_idf(n_doc,w,inv,_id)
    for doc in table:
        vector = table[doc]
        vector = Normalize(vector)
    return table
table = generateTable()
```

5. Implementación de la similitud de cosenos

[5]:
```
def cosineScore(Q):
    global n_doc, inv, table
    Q = cleanSW(Q)
    query = {}
    for i in Q:
        if i not in query:
            query[i] = 1
        else:
            query[i] += 1
    query = Normalize(query)
    coss = {}
    for i in query:
        for _id in table:
            coss[_id] = 0
            #print(table[_id])
            for word in table[_id]:
                if i == word:
                    coss[_id] += query[i]*table[_id][word]
            if coss[_id]==0:
                del coss[_id]
    return coss

#ids
```

0.1 Consultas

1. "Lima necesita calidad en la gestión"

[6]:
```
Q1 = "Lima necesita calidad en la gestión"
cs = cosineScore(Q1)
q1s = sorted(cs, key=lambda x: cs[x], reverse=True)[:10]
for i in q1s:
    print("id:", i)
    print("score:",cs[i])
    print("Documento:",ids[i],"\n")
```

```
id: 1046556941034692608
score: 0.2319293631699373
Documento: Por favor! Su gestión fue un desastre

id: 1046560221265645568
score: 0.18074344082846444
Documento: #DitelElChino #LimaKonDitel por su experiencia en gestión municipal
#VamosKonFuerza #VamosDitelVamos

id: 1046442221694791680
score: 0.1733153564104291
```

```
Documento: £En toda su gestión, cuantas veces Muñoz se enfrentó a Castañeda?

id: 1046543568238665729
score: 0.1691606204702285
Documento: Su gran obra será investigar la gestión anterior, lo mismo que
Ollanta y Villaran.

id: 1046588115731861505
score: 0.16865113990170053
Documento: @Capital967 @GomezBacaxLima Hay que investigar su gestión en la
municipalidad de surco...

id: 1046552020646154245
score: 0.15753342864358402
Documento: Ditel Columbus dales clases de gestión municipal

id: 1046465193503584258
score: 0.15255644674449167
Documento: @BrunoGEsc @JorgeMunozAP Y claro le falta mucho sobre gestión de
residuos sólidos ...y será un trabajo del siguiente alcalde o alcaldesa ... sin
embargo creo que se avanzó mucho en gestión municipal ... me gusta #Miraflores y
mucho ...

id: 1046587319296380928
score: 0.14636147299065186
Documento: RT @Fabiolatuitea: @Capital967 Siempre es así. Manuel Velarde durante
toda su gestión iba en bicicleta al trabajo. Y su propuesta es dar pr

id: 1046622142849306624
score: 0.14614861907790988
Documento: #LimaKonDitel #LimaKonDitel porque #PonjaSabe de gestión municipal,
sabe como dar solución a los problemas de Lima olvidada #LimaKonDitel

id: 1046595797482131457
score: 0.14597122631469747
Documento: Julio Gagó dice que el presidente @MartinVizcarraC quiere promover a
la familia gay .. Eso que tiene que ver con gestión municipal !! @juliogagope
#DebateMunicipal
```

2. “Corrupción en Los Olivos”

[7]:
```
Q2 = "Corrupción en los olivos"
cs = cosineScore(Q2)
q2s = sorted(cs, key=lambda x: cs[x], reverse=True)[:10]
for i in q2s:
    print("id:", i)
    print("score:",cs[i])
```

```
    print("Documento:",ids[i],"\n")
```

```
id: 1046584104773447682
score: 0.2667862479079303
Documento: RT @marcelo_cruz_al: Siempre Unidos, partido liderado por Felipe
Castillo, cabeza de un grupo de poder que hace lo que quiere en Los Olivos

id: 1046562839085887488
score: 0.2565621351029574
Documento: @ManuelVelardeD: 12 mil millones de soles, nos roba la corrupcion en
Lima", lo mismo que Felipe Castillo le robó a la Municipalidad de los Olivos.

id: 1046557767392219138
score: 0.24446502883252316
Documento: Siempre Unidos, partido liderado por Felipe Castillo, cabeza de un
grupo de poder que hace lo que quiere en Los Olivos. La doble moral de Velarde
más presente que nunca.

id: 1046557446578343936
score: 0.23634746028923878
Documento: @patarevalo @ManuelVelardeD £Velarde es el candidato de Siempre
Unidos, el partido de Felipe Castillo sentenciado por desvío de fondos en la
Municipalidad de Los Olivos?

id: 1046594248269860865
score: 0.20521184104427453
Documento: @ManuelVelardeD Una duda muy grande, hablas de combatir la corrupción
pero dentro de tu partido tienes a Felipe Castillo quien ha sido procesado por
Malversación de fondos en la Municipalidad de Los Olivos... £Coherencia?

id: 1046565212160249861
score: 0.1908208642777327
Documento: RT @missysv: Amigui ven a lima norte (smp, los olivos, independencia,
etc), anda a SJL, es tu última semana. Créeme que yendo de caravana e

id: 1046568082871865345
score: 0.1908208642777327
Documento: RT @missysv: Amigui ven a lima norte (smp, los olivos, independencia,
etc), anda a SJL, es tu última semana. Créeme que yendo de caravana e

id: 1046482350094397440
score: 0.18830278597457475
Documento: @FlorMariaJimne1 @reynaga_alfredo @Renzo_Reggiardo @pps_peru Lo
Votaron de Los Olivos con denuncias tiene 3 amantes que lo denunciaron por no
darle manutención a sus hijos y es el que quiere ser tu alcalde???

id: 1046584393345781761
```

```
score: 0.1860983327097851
Documento: @Capital967 @RicardoBelmontC Ya existen y se dan cursos , las
piscinas son gigantes y mencionare 2 lloque yupanqui en los olivos y sinchi roca
en comas, no se q habla

id: 1046564621820329984
score: 0.18100995618571394
Documento: Amigui ven a lima norte (smp, los olivos, independencia, etc), anda a
SJL, es tu última semana. Créeme que yendo de caravana en san Isidro, Miraflores
y jockey, no la haces.
```

3. "Mentiras y psicosociales"

[8]:
```
Q3 = "Mentiras y psicosociales"
cs = cosineScore(Q3)
q3s = sorted(cs, key=lambda x: cs[x], reverse=True)[:10]
for i in q3s:
    print("id:", i)
    print("score:",cs[i])
    print("Documento:",ids[i],"\n")
```

```
id: 1046313071625555968
score: 0.31026010662528425
Documento: @rmapalacios @JorgeMunozAP Psicosociales como el celular encontrado a
Montesinos el día anterior de las últimas elecciones presidenciales?

id: 1046487089490141184
score: 0.26239587175470985
Documento: Cada vez más creativos los psicosociales @SolCn @rmapalacios
@deslengua_2 @AlbertoBelaunde @aleja_puente @TatiAleman @DiarioDeCurwen
@JorgeMunozAP https://t.co/CaqpT6Is6g

id: 1046402611136606209
score: 0.21641134371593337
Documento: La guerra sucia otra vez, Lamentablemente los peruanos tenemos muy
malos ejemplos en las últimas décadas de lo que es diarios chicha psicosociales
y demas suciedad de la cual debemos salir algún día

id: 1046388841324728322
score: 0.2068385476167681
Documento: @rmapalacios @JorgeMunozAP Pero siempre es así la prensa es k que al
final manipula a la gente. O acaso cree. Que urresti esta 2 o que belmot 3 son
PSICOSOCIALES que influyen en la decisión de menos pensantes. O acaso DATUM ES
COFIABLE O SU ENCUESTA DE 1000 PERSONAS ES TODO LIMA

id: 1046385539493089280
score: 0.15261562987130176
```

Documento: @salvajedigital @DanielUrresti1 Correcto lis naranjas e infiltrados sudan frio los cuadros de asalto evaluados y bendecidos por la sra K y presentados por la tia SIN no arrugan se espera mejores psicosociales que fotos de Muñoz con Odebrecht vamos fujis si se puede robar es fácil