

Programació per a la ciència de dades - PAC4

En aquest Notebook trobareu un exercici que suposa la quarta activitat d'avaluació continuada (PAC) de l'assignatura. Aquesta PAC intenta presentar-vos un petit projecte on heu de resoldre diferents exercicis, que englobarà molts dels conceptes coberts durant l'assignatura.

L'objectiu d'aquest exercici serà desenvolupar un **paquet de Python**, fora de l'entorn de Notebooks, que ens permeti resoldre el problema donat. Treballareu en arxius plans `.py`. Aquest paquet haurà d'incloure el corresponent codi organitzat lògicament (separat en mòduls, organitzats per funcionalitat), la documentació de codi (*docstrings*) i tests. A més, s'hauran d'incloure els corresponents arxius de documentació d'alt nivell (`README`), així com els arxius de llicència i dependències (`requirements.txt`) tal i com s'explica a la teoria. Fer un `setup.py` és opcional, de la mateixa forma que incloure un informe (en format `.pdf`) amb el resum dels resultats de la pràctica, però si es fa es valorarà positivament de cara a la nota de la pràctica i del curs.

Se'ns demana que implementem un paquet (mòdul) de Python que sigui capaç de realitzar una anàlisi d'imatges de diferents ciutats europees preses entre 2015 i 2019. Per una banda tindrem les imatges i per altre els objectes presents en aquestes imatges i la seva posició dins de la imatge.

Enunciat:

Ens han encarregat analitzar el contingut d'una base de dades de Twitter per a un projecte de processament del llenguatge natural (NLP) relacionat amb l'anàlisi de sentiments. Per a començar a treballar, tenim un dataset amb 800.000 tuits i sis variables: *sentiment* indica si el sentiment del tuit és positiu o negatiu, *id* és un identificador únic del tuit, *date* indica la data en què va ser publicat en la xarxa social, *query* indica la consulta (si no hi ha mostrarà "NO_QUERY"), *user* és el nom de l'usuari i *text* conté el missatge del tuit. El dataset complet ho podeu trobar aquí.

En aquesta PAC haureu de treballar amb aquest dataset per a processar els textos. Les dades els teniu en **twitter_reduced.csv**, que està comprimit en el fitxer **twitter_reduced.zip**.

Presentació dels resultats:

Per a fer el lliurament més fàcil i homogènia us demanem que organitzeu el codi de tal manera que des del fitxer principal retorni totes les respostes que se us demani en la PAC fent ús de funcions que haureu de definir en mòduls. Per a això, en cada exercici, us indicarem el format que ha de tenir cada resposta. De tal manera que executant 'main.py' es vagi responnent a tota la PAC. Si valoreu que és millor fer-ho d'una altra manera haureu de documentar-ho molt bé en

el README perquè es pugui executar sense problema. Us recordem que en el README també heu d'indicar com executar els test i comprovar la cobertura d'aquests.

Control i revisió del `*dataset`:

Quan comencem a treballar en un projecte d'anàlisi de dades, una bona pràctica és assegurar-nos que les dades són correctes. En altres paraules, és necessari fer una anàlisi exploratòria inicial per a detectar errors o casos especials i prendre decisions sobre com abordar-los. Aquí us proposem fer:

Exercici 1.1. Descomprimiu el fitxer `twitter_reduced.zip` i guardeu el seu contingut en la carpeta `data` del projecte.

Exercici 1.2. Llegiu el fitxer `twitter_reduced.csv` i carregueu el `*dataset` com una llista de diccionaris. Cada fila del fitxer original correspondrà amb un diccionari seguint l'estructura d'aquest exemple:

```
{ 'sentiment': '0',  
  'id': '1467810369',  
  'date': 'Mon Apr 06 22:19:45 PDT 2009',  
  'query': 'NO_QUERY',  
  'user': 'TheSpecialOne',  
  'text': '@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You  
shoulda got David Carr of Third Day to do it. ;D" }
```

Mostreu per pantalla els 5 primers registres del `*dataset` mitjançant `*print`.

Exercici 2.1. Amb molta freqüència, els textos solen contenir elements innecessaris o sorollosos que no aporten informació rellevant per a l'anàlisi. El preprocessament ajuda a eliminar aquests elements i reduir el soroll en les dades.

Realitzeu un preprocessament fent ús de **expressions regulars** que elimini les *URLs*, els caràcters especials no ASCII i els símbols i que converteixi el text a minúscules. Substituiu els textos originals pels modificats en el dataset de l'apartat anterior.

Exercici 2.2. D'altra banda, les `*stopwords` són paraules comunes que no aporten un valor semàntic significatiu a l'anàlisi de text. En eliminar aquestes paraules, es redueix la dimensionalitat del text i s'elimina soroll addicional, permetent centrar-se en les paraules més rellevants per a l'anàlisi. Per exemple, el següent tuit (després del preprocessament de l'apartat anterior): `awww that is a bummer you shoulda got david carr of third day to do it`

quedaría reducido a: `awww bummer got david carr third day` tras eliminar las stopwords.

Per a aquest projecte considerarem que les *stopwords són les següents: ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', 'her', 'hers', 'herself', 'it', 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', 'should', 'now']

Elimineu les *stopwords dels textos dels tuits i mostreu per pantalla les 5 últimes files.

Exercici 3. La tècnica "Bag of Words" (BoW), s'utilitza en el processament del llenguatge natural (*NLP*) per a representar i analitzar textos. La idea bàsica darrere de la BoW és tractar un document de text com una "bossa" (és a dir, una col·lecció no ordenada) de paraules, sense tenir en compte l'estructura gramatical o l'ordre en el qual apareixen les paraules en el text.

En la representació de la *BoW, es crea un vocabulari de totes les paraules úniques que apareixen en el conjunt de documents de text. Després, es compta el nombre de vegades que cada paraula del vocabulari apareix en cada document, la qual cosa es coneix com a freqüència de terme.

Obtingueu les freqüències de termes de cada tuit i emmagatzemeu-les en una llista de diccionaris en la qual cada diccionari indiqui les paraules i la seva freqüència d'aparició en el tuit. Obtingueu també un vocabulari amb totes les paraules úniques del *dataset i guardeu-les en una llista. Mostreu per pantalla els 5 primers elements de la llista de diccionaris obtinguda. Ordeneu alfabèticament el vocabulari i mostreu per pantalla les 10 primeres paraules.

Nota: cada element de la llista de freqüències haurà de tenir la següent estructura:

{'paraula_1': nombre de vegades que apareix la paraula_1,

'paraula_2': nombre de vegades que apareix la paraula_2,

...

'paraula_N': nombre de vegades que apareix la paraula_N,}

Exercici 4.1. Completeu el *dataset original* afegint a cada registre del mateix una nova variable amb el seu diccionari de freqüències de termes associat. Mostreu l'element 20 del dataset.

Exercici 4.2. Guardeu el dataset processament en format *csv*. El nom del fitxer serà *twitter_processed.csv* i se situarà en la carpeta *data* del projecte.

Anàlisi de dades:

L'anàlisi de sentiments juga un paper molt important en la nostra societat cada vegada mes digitalitzada. Actualment hi ha grups de recerca que dediquen tot el seu esforç a analitzar sentiments a través de les xarxes socials i estudiar tendències d'aquests STOP. A causa de la gran importància que tenen aquest tipus d'anàlisi, us demanem que utilitzant el *dataset anterior ens ajudeu a entendre el tipus de grups que hem obtingut fent un clustering*, volem saber si els clústers obtinguts tenen sentit o no.

Per a això deveu:

Exercici 5. Genereu una *word cloud* utilitzant els tuits obtinguts en l'exercici 4.2. Abans de generar els *word clouds* heu de respondre les següents preguntes, que us ajudaran en l'anàlisi:

1. Quants clústers tenim en el nostre *dataset*?
2. Tenim elements buits en les columnes *text*? Si és així, quin és el percentatge? 2.1. En cas de tenir elements nuls en la columna *text*, s'han d'eliminar abans de generar el *word cloud*.
3. Generar un *word cloud* per a cada clúster.

**** Definició de *word cloud*:** En espanyol núvols de *palbras*. Els núvols de paraules o núvols d'etiquetes poden utilitzar-se com a eines d'anàlisi de l'aprenentatge. Són representacions visuals d'un grup de paraules utilitzades pels participants i basades en la seva freqüència 1.

Nota 1: Els clústers els podeu trobar en la columna *sentiment*. **Nota 2:** *Totes les respostes anteriors s'han de resoldre utilitzant codi, mostrant el resultat en un print.*

Exercici 6 Una vegada generat el *word cloud* en l'exercici anterior us demanem que feu una validació dels resultats obtinguts en l'apartat anterior. Per a això heu de generar un histograma amb els valors que heu obtingut en l'exercici 3.

****Nota:** Deveu *generar un histograma per a cada clúster*

Exercici 7 Analitzeu la *word cloud* juntament amb els histogrames i responeu a les següents preguntes:

- a. Quines són les paraules més utilitzades en les crítiques positives?
- b. Quines són les paraules més utilitzades en les crítiques negatives?
- c. Hi ha paraules que apareguin tant en les crítiques positives com en les negatives?

d. A partir de la *word cloud*, què es pot deduir sobre el sentiment general de cada grup?

Nota: Escriviu cada resposta en un text curt no més de 3 línies per resposta.

Cobertura dels tests

La mesura de la cobertura dels tests s'utilitza per a avaluar l'eficàcia dels tests proposats. En particular, serveix per a determinar la qualitat dels tests i determinar les parts crítiques del codi que no han estat testades. Per a mesurar aquest valor us proposem l'ús de l'eina `Coverage.py`. En la documentació, podeu trobar com instal·lar-la i com usar-la.

Per a avaluar els tests desenvolupats en la PEC4, demanem un mínim del 50% de cobertura.

Criteris de correcció

Aquesta PAC es valorarà seguint els criteris següents:

- **Funcionalitat** (5.75 punts): Es valorarà que el codi implementi correctament el que demana l'enunciat.
 - Exercici 1 (0.25 punts)
 - Exercici 2 (0.75 punts)
 - Exercici 3 (1 punt)
 - Exercici 4 (1.75 punts)
 - Exercici 5 (1 punts)
 - Exercici 6 (0.5 punts)
 - Exercici 7 (0.5 punts)
- **Documentació** (0.5 punts): Totes les funcions dels exercicis d'aquesta PAC hauran d'estar correctament documentades utilitzant docstrings (en el format que preferiu).
- **Modularitat** (1 punt): Es valorarà la modularitat del codi (tant l'organització del codi en fitxers com la creació de funcions).
- **Estil** (0.5 punts): El codi ha de seguir la guia d'estil de Python (PEP8), exceptuant els casos on fer-ho compliqui la llegibilitat del codi.
- **Tests** (1.25 punts): El codi ha de contenir una o diverses *suites* de tests que permetin comprovar el bon funcionament de les funcions implementades, obtenint un mínim del 50% de cobertura.
- **Requeriments** (0.5 punts): Hi haurà d'haver un fitxer de requeriments que llisti (només) les llibreries necessàries per a executar el codi.
- **README i llicència** (0.5 punts): Es valorarà la creació d'un fitxer de README, que presenti el projecte i expliqui com executar-lo, així com la inclusió de la llicència sota la qual es distribueix el codi (podeu triar la que vulgueu).

Important

Nota 1: De la mateixa manera que en les PACs anteriors, els criteris transversals es valoraran de manera proporcional a la part de la funcionalitat implementada.

Per exemple, si el codi només implementa la meitat de la funcionalitat demanada, i la documentació d'aquesta part és perfecta, aleshores la puntuació corresponent a la part de documentació seria de 0.25.

Nota 2: És imprescindible que el paquet que lliureu s'executi correctament a la màquina virtual, i que el fitxer de README que inclogueu expliqui clarament com s'ha d'executar el vostre codi per tal de generar les gràfiques resultants de l'anàlisi i tots els resultats, a més de com executar els tests i comprovar la cobertura.

Nota 3: Lliureu el paquet com a un únic arxiu .zip al Registre d'Avaluació Continua. **El codi de Python haurà d'estar escrit en fitxers plans de Python.**