CAT-PAC4-enun

May 17, 2021

1 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 constarà d'un únic exercici a resoldre, 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. Aquest 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`). Fer un setup.py és opcional, 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 de dades amb informació sobre diferents entrevistes en els Estats Units durant l'últim any en relació amb el coronavirus i com l'ex-president Trump ha gestionat la situació. Per una banda tindrem dades sobre les entrevistes (qui les ha dutes a terme, quan, quina era la pregunta, a quin grup de població estava dirigida, ...) i a més a més tindrem informació sobre la credibilitat d'aquestes entrevistes en funció de qui les ha fet.

2 Les dades

Les dades a analitzar ens són proporcionades en dues col·leccions de dades separades: `covid_approval_polls.csv` i `covid_concern_polls.csv`. D'altra banda tenim `pollster_ratings.xlsx` amb la informació sobre la credibilitat de les entrevistes per entrevistador (*pollster*). Aquestes dades provenen del repositori Five Thirty Eight. Us aconsellem usar els arxius que us hem proporcionat nosaltres ja que han sofert alguna modificació respecte les dades originals.

`covid_approval_polls.csv` conté entrevistes sobre l'aprovació o no de l'actuació de Donald Trump durant l'inici de la pandèmia. Mentre que `covid_concern_polls.csv` conté entrevistes sobre la concienciació de la població de l'impacte econòmic o bé de l'impacte a la seva salut o la de la seva família del coronavirus. I com hem mencionat anteriorment `pollster_ratings.xlsx` conté informació sobre la credibilitat de l'agent entrevistador.

Fent una ullada als arxius proporcionats, podreu veure que els diferents arxius contenen força informació. Per a resoldre l'exercici proposat, potser no usareu tota la informació que contenen aquests arxius.

2.1 covid_approval_polls.csv

L'arxiu `covid_approval_polls.csv` conté 2227 línies amb informació sobre 2226 entrevistes:

```
start_date,end_date,pollster,sponsor,sample_size,population,party,subject,tracking,text,
    approve,disapprove,url
2020-02-02,2020-02-04,YouGov,Economist,1500,a,all,Trump,FALSE,Do you approve or
    disapprove of Donald Trumps handling of the coronavirus outbreak?,42,29,https://
    d25d2506sfb94s.cloudfront.net/cumulus_uploads/document/73jqd6u5mv/econTabReport.pdf
2020-02-02,2020-02-04,YouGov,Economist,376,a,R,Trump,FALSE,Do you approve or disapprove
    of Donald Trumps handling of the coronavirus outbreak?,75,6,https://d25d2506sfb94s.
    cloudfront.net/cumulus_uploads/document/73jqd6u5mv/econTabReport.pdf
...
```

La primera línea conté la capçalera i nom de les columnes per cada entrada (línia) del document.

Comentar que les columnes `approve` i `disapprove` estan en tant per cent respecte la resposta a la població (`sample_size`) de cada entrevista.

2.2 covid_concern_polls.csv

L'arxiu `covid_concern_polls.csv` conté 639 línies amb informació sobre 638 entrevistes:

```
start_date,end_date,pollster,sponsor,sample_size,population,party,subject,tracking,text,
    very,somewhat,not_very,not_at_all,url
2020-01-27,2020-01-29,Morning Consult,"",2202,a,all,concern-economy,FALSE,How concerned
    are you that the coronavirus will impact the following? U.S. economy,19,33,23,11,
    https://morningconsult.com/wp-content/uploads/2020/02/200167
    _crosstabs_CORONAVIRUS_Adults_v2_JB-1.pdf
2020-01-31,2020-02-02,Morning Consult,"",2202,a,all,concern-economy,FALSE,How concerned
    are you that the coronavirus will impact the following? U.S. economy,26,32,25,7,https
    ://morningconsult.com/wp-content/uploads/2020/02/200191
    _crosstabs_CORONAVIRUS_Adults_v2_JB-1.pdf
...
```

La primera línea conté la capçalera i nom de les columnes per cada entrada (línia) del document.

Comentar que les columnes `very`, `somewhat`, `not_very` i `not_at_all` estan en tant per cent respecte la resposta a la població (`sample_size`) de cada entrevista.

Notar que els percentatges de vegades sumen una mica més de 100 (o menys). Això és degut a errors en els arrodoniments de les dades, però a vosaltres no us afecta aquest fet en l'anàlisi que heu de fer per aquesta pràctica.

2.3 pollster_ratings.xlsx.

L'arxiu `pollster_ratings.xlsx` conté 454 línies amb informació sobre 453 agents entrevistadores:

```
Pollster,Pollster Rating ID,# of Polls,NCPP / AAPOR / Roper,Live Caller With Cellphones,
    Methodology,Banned by 538,Predictive Plus-Minus,538 Grade,Mean-Reverted Bias,Races
    Called Correctly,Misses Outside MOE,Simple Average Error,Simple Expected Error,Simple
     Plus-Minus,Advanced Plus-Minus,Mean-Reverted Advanced Plus Minus,# of Polls for Bias
     Analysis,Bias,House Effect,Average Distance from Polling Average (ADPA),Herding
    Penalty,latest_poll
```

```
Monmouth University,215,108,yes,yes,Live,no,-1.6,A+,D
    +1.3,81%,21%,5.4,6.7,-1.2,-2,-1.6,71,D +1.8,R +0.4,5.2,0.2,3/13/20
Selzer & Co.,304,48,yes,yes,Live,no,-1.3,A+,D +0.1,79%,25%,4.6,6.1,-1.3,-1.8,-1.1,31,D
    +0.2,D +0.2,5.2,0,11/1/18
ABC News/The Washington Post,3,73,yes,yes,Live,no,-1.3,A+,D
    +0.5,72%,7%,2.8,4.8,-1.7,-1.7,-1.2,68,D +0.8,D +1.3,3.8,0.12,10/31/18
...
```

La primera línea conté la capçalera i nom de les columnes per cada entrada (línia) del document.

3 Exercici

Caldrà que genereu funcions que us permetin fer els següents càlculs:

1. De l'arxiu `covid_approval_polls.csv`:

   1.1 Implementeu una funció que compti **de forma eficient** i mostri per pantalla el nombre de vegades que apareixen els patrons descrits (és a dir, en quantes línies apareix) a continuació en l'arxiu, incloent-hi un missatge explicatiu pels valors que ensenyeu. Els patrons a considerar són:

   - El terme *Huffington Post*
   - Una url (sigui http o https) amb format pdf. Per exemple: *https://d25d2506sfb94s.cloudfront.net/cumulus_uploads/document/73jqd6u5mv/econTabReport.pdf*

   Un exemple d'*output* seria:

   The pattern Huffington_Post appears X times.

   The pattern url_pdf appears Y times.

   1.2 Si tinguéssim un arxiu de 1Gb ho faries igual? Si no és així, implementeu la solució per aquest cas.

   1.3 Si tinguéssim 100 arxius d'1Gb com ho faries? No cal implementar la solució, només una petita descripció de com resoldries el problema.

2. Llegiu els arxius facilitats de la forma més eficient tenint en compte les tasques demanades a continuació i a l'exercici 3, 4 i 5. Justifiqueu la vostra decisió. Prepareu les dades per cadascun dels .csv, obtenint dues taules que anomenarem `approval_polls`(provinent de `covid_approval_polls.csv`) i `concern_polls` (provinent de `covid_concern_polls.csv`) de forma que es compleixin tots els següents requisits:

   - Només estarem interessats en les entrevistes les quals el seu agent entrevistador (*pollster*) estigui a la taula `pollster_ratings.xlsx`
   - Només estarem interessats en aquelles entrevistes sense *tracking*.
   - Només estarem interessats en les entrevistes les quals el seu agent entrevistador no ha estat vetat (*banned*).

   Nota: Per tal de llegir arxius en format *xlsx* podeu instal·lar llibreries addicionals a la màquina virtual.

3. Sobre les dades extretes a l'exercici 2 de la taula `approval_polls`, calculeu i representeu gràficament: 3.1 El nombre de persones que aproven (*approve*) i el nombre de persones que de-

saproven (*disapprove*), per a les preguntes que contenen les paraules *Trump* i *coronavirus* en el text. Representarem aquestes dades per cada partit (*party*) (*D* (demòcrates), *R* (republicans), *I* (independents), *all* (persones sense classificar per partit)).

4. Sobre les dades extretes a l'exercici 2 de la taula `concern_polls`, tenint en compte les següents transformacions sobre el grau a la classificació (*grade*) $*$, calculeu i representeu gràficament (excepte el 4.1):

   4.1 Quanta gent ha participat en les entrevistes. Representar el resultat per pantalla degudament formatat.

   4.2 Quanta gent en la matèria (*subject*) de l'entrevista relacionada amb l'economia (*economy*) està *very* (*concern*, preocupació) i quanta està *not_at_all* (*concern*, preocupació).

   4.3 Quin és el percentatge de gent en la matèria (*subject*) de l'entrevista relacionada amb l'infecció (*infected*) està *very* (*concern*, preocupació) i quanta està *not_at_all* (*concern*, preocupació).

   4.4 Quantes entrevistes hi ha per cada nota classificatòria (*grade*).

$*$ La nota classificatoria serà reduïda a només els valors *A, B, C, D, F* tenint en compte que en cas de tenir una valoració entre dues categories ens quedarem amb l'inferior. *Exemple B/C* $\rightarrow$ *C, B-* $\rightarrow$ *B, B+* $\rightarrow$ *B*.

5. A partir de les dades de l'exercici 4, crearem una nova variable que serà la puntuació (credibilitat) que li donarem a aquella agent entrevistadora. Aquesta puntuació vindrà donada per:

   puntuació = nota avaluada + *Predictive Plus-Minus*

   On la nota avaluada es refereix a que la classificació per *A, B, C, D, F* s'avaluarà segons: A $\rightarrow$ 1, B $\rightarrow$ 0.5, C $\rightarrow$ 0, D $\rightarrow$ -0.5, F $\rightarrow$ -1

   5.1 Calcular (i representar gràficament) per aquelles entrevistes on la seva puntuació sigui superior o igual a 1.5:

   a) El nombre de persones segons nivell de preocupació (*concern very, somewhat,...*) segons si l'entrevista havia finalitzat abans estricte del 2020-09-01 (1 de setembre del 2020), o després.

   b) El percentatge de persones segons nivell de preocupació (*concern very, somewhat,...*) segons si l'entrevista havia finalitzat abans estricte del 2020-09-01, o després. (**Nota:** percentatge respecte del nombre de persones per cada grup abans del 2020-09-01 i després).

   5.2 Què en podeu dir de les dues gràfiques produïdes a l'exercici anterior (5.1)? Quines conclusions podeu treure?

Nota: en el cas que alguna de les variables sobre les entrevistes (files) que cal usar no sigui un camp informat es descararà aquella entrevista. Si hi ha camps no informats però que no els usem al llarg de l'exercici no caldrà eliminar aquella entrada.

A més, haureu de generar codi que permeti **representar els resultats de l'exercici 3, 4 (excepte 4.1) i 5 gràficament**, podeu igualment representar el resultat per pantalla per tal que poguem

comprovar els vostres resultats de manera exacta. Per a cada funció caldrà que penseu quin tipus de gràfica és la més adient per a representar el resultat.

El codi haurà d'estar correctament comentat, incloent-hi documentació de funcions, i correctament testejat usant la llibreria `unittest`. Els testos proporcionats hauran de donar cobertura com a mínim al 50% de la funcionalitat proposada.

3.1 Cobertura dels testos

El mesurament de la cobertura dels testos s'utilitza per avaluar l'eficàcia dels testos proposats. En particular, serveix per determinar la qualitat dels testos desenvolupats i per determinar les parts crítiques del codi que no han estat testejades. Per tal de mesurar aquest valor, proposem l'ús de l'eina `Coverage.py`. A la documentació, podreu trobar com instal·lar-la i com usar-la.

Per a avaluar la qualitat dels testos desenvolupats per la PAC4, demanem un mínim del 50% de cobertura.

4 Ús de Git

Per tal de posar en pràctica el que heu après a la Unitat 6 sobre `Git`, proposem l'ús de `GitHub Classroom` per a desenvolupar el vostre paquet de Python. GitHub Classroom és una eina gratuïta de codi obert que ajuda a simplificar l'ús educatiu de GitHub. Hem usat GitHub Classroom per a crear una aula com aquesta i on hem creat una tasca per a la PAC4. Per a fer ús d'aquest espai que hem creat, us aconsellem seguir els passos indicats en aquesta guia que expliquen com crear un repositori per a treballar en la tasca que hem preparat i trobareu en aquest enllaç.

L'ús d'aquesta eina no és obligatori per l'avaluació de la PAC4, però creiem que és una molt bona oportunitat per a posar en pràctica els vostres coneixements en un entorn vital per a tothom que treballi o vulgui treballar en l'àmbit de la ciència de dades.

4.1 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 (0.75 punts)
  - Exercici 2 (0.3 punts)
  - Exercici 3 (0.7 punt)
  - Exercici 4 (1.5 punts)
  - Exercici 5 (1.75 punts)
  - Visualitzacions (0.75 punt)
- **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 testos 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).

4.1.1 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.**