Tipologia i cicle de vida de les dades

Pràctica 1

Mireia Calzada i Noemi Lorente

1. Títol del dataset. Cal que poseu un títol que sigui descriptiu.

Muntanyes del món

2. Subtítol del dataset. Agregueu una descripció àgil del vostre conjunt de dades pel vostre subtítol.

Es tracta d'un recurs documental de muntanyes del món, inclou diferents característiques com per exemple, el país, la regió, el continent o l'alçada, que permetran en un posterior tractament fer-ne diferents classificacions.

3. Imatge. Agregueu una imatge que identifiqui el vostre dataset visualment

Quien siente la montaña no necesita explicaciones y mientras existan paredes, agujas y aristas, habrá quien las escale, disfrutando de lo que hace, aunque no comprenda exactamente el por qué. (Josep Manel Anglada)

El verdadero arte de la escalada es el de la supervivencia y la dificultad, que consiste en querer siempre ir más allá de lo que dominamos. Aventurarse allí donde nadie estuvo antes, allí donde nadie te sigue ni te comprende. Lejos de los caminos conocidos es donde los sentimientos y las apariencias resultan más intensas. (Reinhold Messner)

4. Context. Quina és la matèria del conjunt de dades?

És una recull de característiques específiques de diferents de muntanyes del món.

Recull informació útil sobre muntanyes, cims, pics, turons, volcans, massissos, serres i serralades de tot el món. També disposa d'articles de geografia de muntanya, guies, rutes, llistats, classificacions geogràfiques, galeries d'imatges i un glossari especialitzat, que complementen la informació sobre muntanyes i regions de tot el món.

És, per tant, un recurs útil per qualsevol persona que necessiti documentació geogràfica.

5. Contingut. Quins camps inclou? Quin és el període de temps de les dades i com s'ha recollit?

Les dades sobre les diferents muntanyes del món es recullen al web www.montipedia.com des de finals del 2005, en primer lloc amb la intenció de pal·liar la carència de contingut digital sobre aquesta temàtica a Internet tot aprofitant la combinació del perfil professional i l'afició al muntanyisme dels seus editors, i alhora contribuir a la difusió de la informació de manera lliure i gratuïta.

La informació és publicada a montipedia a partir d'aportacions que són verificades i corregides per l'equip editorial. Aquestes aportacions són finalment publicades amb el nom de l'autor que ha aportat la informació.

Per cada muntanya documentada es recull la següent informació:

- **Nom:** Nom de la muntanya.
- **Nom alternatiu**: Altre nom pel qual es coneix la mateixa muntanya
- **Tipus de muntanya**: Classificació segons el tipus de muntanya (cerro, pico, aguja, nevado, monte, volcán, ...)
- **Continent / zona geogràfica** Continent en el que està situada la muntanya (América del Norte, América central, América del sur, Europa, Ásia, África, Oceanía, Antártida)
- **Unitat de relleu**: Cadena muntanyosa a la qual pertany la muntanya (cordilleras, sistemas, sierras, macizos, etc.)

  Per exemple, Corredor ecológico Ajusco-Chichinautzin

- **País: (**o llista de països): En els que es troba la muntanya.
  Per exemple, China, Kirguizistan,..

- **Cimera**: Indica si es tracta d'un pic principal o secundari.

- **Regió:** Regió on es troba la muntanya
  Per exemple, Andes centrales, Andes bolivianos, ...

- **Altitud:** Alçada en metres
  Per exemple, 3.690 m

- **Classificació per alçada**: Camp calculat en funció de l'altitud
  Per exemple, tres mil, quatre mil, cinc mil, sis mil, set mil, vuit mil.

- **Latitud**: Coordenada geogràfica corresponent a la latitud
  Per exemple, 27º 59' N

- **Longitud**: Coordenada geogràfica corresponent a la longitud
  Per exemple, 86º 56' E

- **Descripció**: Breu descripció de la muntanya.

- **Keywords:** Conjunt de paraules clau que ens ajuden a classificar les muntanyes
  Per exemple, Everest, monte, alpinismo, ...

6. Agraïments. Qui és propietari del conjunt de dades? Inclou cites de recerca o anàlisi anteriors.

Com s'ha comentat en l'apartat anterior, les dades mostrades en montipedia són un recull d'aportacions de diferents autors, tot i que cal agrair l'equip editorial per la seva dedicació en la verificació i correcció de la informació, així com de la seva publicació.

Agrair doncs al propietari del web www.montipedia.com Iván G. Guerrero per la iniciativa i dedicació, així com a l'editorial Lexico Tecnia. LT Servicios Lingüísticos y Editoriales, S.L., així com a Vanesa Markovic i el mateix propietari en les seves tasques de disseny.

7. Inspiració. Per què és interessant aquest conjunt de dades? Quines preguntes li agradaria respondre la comunitat?

El muntanyisme i l'alpinisme és un tema molt de moda que cada vegada té més seguidors i adeptes. Per això, el conjunt de dades generat es pot utilitzar com informació base d'estudis d'investigació mediambientals relacionats amb estils de vida saludables o bé estudis d'impacte socioeconòmic que permetin identificar on és més factible impulsar un negoci relacionat amb el muntanyisme tenint en compte les característiques d'aquest.

També es pot creuar amb altres tipus d'informació com activitats d'oci, esport i aventura que es poden realitzar en aquestes muntanyes i zones geogràfiques o bé amb informació sobre els accidents de muntanya que s'hi han produït, per maximitzar la seguretat i precaucions de tots els muntanyencs.

8. Llicència. Cal que seleccioneu una d'aquestes llicències i cal dir perquè l'heu seleccionada:

○ Released Under CC0: Public Domain License
○ Released Under CC BY-NC-SA 4.0 License
○ Released Under CC BY-SA 4.0 License
○ Database released under Open Database License, individual contents under Database Contents License
○ Other (specified above)
○ Unknown License

Hem escollit la llicència CC BY-NC-SA 4.0 per a la publicació d'aquest conjunt de dades per donar compliment als requeriments de la llicència Coloriuris sota la que el propietari de les dades, Montipedia, permet copiar, distribuir i comunicar de forma pública les anotacions del lloc web, tant de forma parcial o total, permanent o provisional, sempre que es realitzi sense ànim de lucre i per a usos no comercials. En altre cas, la cessió del dret de transformació quedarà limitada als supòsits previstos per la normativa vigent en matèria de drets d'autor.

La llicència CC BY-NC-SA 4.0 dona llibertat per:

- compartir, copiar i distribuir el material en qualsevol mitjà i forma
- adaptar, remesclar, transformar i crear a partir dels materials

Seguint els termes de llicència següents:

- Cal reconèixer l'autoria del conjunt de dades, proporcionar un enllaç a la llicència i indicar si heu fet algun canvi.
- No podeu utilitzar el material per a finalitats comercials
- Si remescleu, transformeu o creeu a partir del material, heu de difondre les vostres creacions amb la mateixa llicència que l'obra original.
- No podeu aplicar termes legals ni mesures tecnològiques que restringeixin legalment a altres de fer qualsevol cosa que la llicència permet.

9. Codi: Cal adjuntar el codi amb el que heu generat el dataset, preferiblement amb R o Python, que us ha ajudat a generar el dataset

La carpeta *src* del projecte github conté els fitxers python amb els que s'ha generat el dataset.

S'ha programat en python3, i concretament hem utilitzat la llibreria BeatifulSoup per a parsejar el web.

Beautiful Soup és una llibreria de python per a explorar dades d'arxius en HTML i XML. Aquesta eina construeix l'arbre d'etiquetes de la pàgina HTML de manera que podem extreure la informació desitjada de manera ràpida i eficaç.

Un dels avantatges d'aquesta biblioteca és que totes els documents sortints de l'extracció de dades estan en UTF-8, i s'eviten d'aquesta manera els problemes típics de les codificacions, no obstant, cal dir que tot i utilitzar aquesta llibreria hem hagut de resoldre algun problema que ens hem trobat amb la lletra *ñ*.

També hem utilitzat el format d'arxius csv (doncs es tracta d'un format senzill per importar i exportar dades)

D'altra banda, també s'ha tractat en el mateix script l'atribut 'alçada' per tal de fer una classificació més general de les muntanyes segons els metres que tenen (tres mi, quatre mil, …)

10. Dataset: Dataset en format CSV

La carpeta *csv* del projecte github conté el fitxer de dades generat.

Recursos web:

- Tutorial de GitHub
  URL: https://guides.github.com/activities/hello-world/

- Learn Python for Data Science - Online Course | DataCamp
  URL: https://www.datacamp.com/courses/intro-to-python-for-data-science

- About The Licenses - Creative Commons
  URL: https://creativecommons.org/licenses/

- Open Database License
  URL: https://opendatacommons.org/licenses/odbl/

- Installing and using Git and GitHub on Ubuntu: A beginner's guide
  URL: https://www.howtoforge.com/tutorial/install-git-and-github-on-ubuntu-14.04/

- Web scraping the President's lies in 16 lines of Python
  URL: http://www.dataschool.io/python-web-scraping-of-president-trumps-lies/

- How To Install the Anaconda Python Distribution on Ubuntu 16.04
  URL:https://www.digitalocean.com/community/tutorials/how-to-install-the-anaconda-python-distribution-on-ubuntu-16-04

- KdNuggets
  URL:https://www.kdnuggets.com/2018/02/top-20-python-ai-machine-learning-open-source-projects.html

Puntuació:
Tots els apartat són obligatoris. La ponderació dels exercicis és la següent:
● Els apartats 1, 2, 3 i 4 valen 0,25 punts cadascun.
● Els apartats 5, 6, 7, 8 valen 1 punt cadascun.
● Els apartats 9 i 10 valen 2,5 punts cadascun.

Lliurament
Durant la setmana del 19 de març: entrega parcial opcional. En l'entrega parcial els estudiants hauran de lliurar per correu electrònic (mcalvogonza@uoc.edu) l'enllaç al repositori Github amb el que hagin avançat.

Pel que fa a l'entrega final, cal lliurar un únic fitxer que contingui l'enllaç a Github on hi hagi:
1. Una Wiki on hi hagi els noms dels components del grup i una descripció dels fitxers.
2. Un document Word, Open Office o PDF amb les respostes a les preguntes i els noms dels components del grup.
3. Una carpeta amb el codi Python o R generat per obtenir les dades.
4. El fitxer CSV amb les dades.
Aquest document de l'entrega final de la Pràctica 1 s'ha de lliurar a l'espai de Lliurament i Registre d'AC de l'aula abans de les 23:59 del dia 16 d'abril. No s'acceptaran lliuraments fora de termini.