Tipologia i cicle de vida de les dades

PAC1

ESTUDIANT: David Gil del Rosal (dgilros@uoc.edu)

NOTA: per a facilitar la lectura del document s'han copiat els enunciats dels exercicis (text en blau).

Exercici 1 (60%)

Després de llegir el recurs "Calvo, M., Pérez, D., Subirats, L. (2019). Introducció al cicle de vida de les dades." contesta les següents preguntes:

1. Amb l'ajuda d'un exemple real, expliqui cadascuna de les fases del cicle de vida de les dades necessàries per a la resolució d'un cert problema. Ressalti amb text tipus "Bold" (negreta) el nom de cadascuna de les fases. ¿Aquestes dades requereixen una neteja en particular?. (màxim 250 paraules).

   He triat el sistema MeteoSwiss per a la captació, anàlisi i difusió d'informació meteorològica ja que la documentació disponible al seu lloc web www.meteoswiss.ch permet identificar clarament les fases del cicle de vida de les dades:

   1) **Captura**. Les dades són creades per diverses fonts meteorològiques com estacions, avions i satèl·lits. Consisteixen principalment en lectures numèriques i imatges.
   2) **Emmagatzematge**. Les dades rebudes s'emmagatzemen en un repositori provisional per al seu preprocessat. Una vegada tractades s'insereixen en un magatzem de dades construït sobre una base de dades relacional optimitzada per a fer consultes.
   3) **Preprocessat**. A la documentació es citen les següents tasques:
      - **Integració** de les dades obtingudes de les fonts meteorològiques solucionant inconsistències com diferents unitats de mesurament.
      - **Conversió**. Es calculen nous atributs i s'agreguen les dades per les dimensions geogràfica i temporal.
      - **Neteja**. Les dades perdudes (absència de lectura) s'intenten completar de forma automàtica interpolant els valors de les lectures. Quan no és possible es revisen manualment per experts en meteorologia. Els valors extrems es marquen com invàlids i són revisats per experts.
   4) **Anàlisi**. Les dades s'utilitzen per a la predicció meteorològica emprant tècniques de càlcul numèric. Les prediccions a mig termini es basen en mètodes d'estadística inferencial.
   5) **Visualització**. A la web de MeteoSwiss es permet la visualització de les dades amb diversos nivells de detall i filtratge geogràfic i temporal.
   6) **Publicació**. Les dades es publiquen a la web de MeteoSwiss.

2. En la fase de captura del cicle de vida de les dades, aquestes sempre poden ser creades? Expliqui els dos principals mecanismes per a la captura de dades i desenvolupi un exemple per a cadascun dels mecanismes (màxim 200 paraules).

Les dades no sempre poden ser creades. Molt sovint són creades per tercers i no es disposa de capacitat d'intervenció en els sistemes que les generen. Per aquest motiu, junt a la **creació** l'altre mecanisme principal de captura de dades és l'**extracció**.

La creació consisteix en l'emmagatzemament de les dades d'interès a mesura que es van generant. La creació requereix accés al sistema que genera les dades i la possibilitat d'intervenció per a obtenir les dades. Un exemple són els registres d'accés (*logs*) generats pels servidors web com Apache que contenen dades sobre els accessos als llocs web. Els servidors web incorporen rutines per a emmagatzemar dades sobre cada accés que permeten analitzar el tràfic.

L'extracció s'aplica quan no es possible intervenir en els sistemes que generen les dades, per la qual cosa s'han de fer consultes a les fonts de dades emprant tècniques com utilització d'APIs (si la font de dades disposa d'ells) o el web scraping. Un exemple són els cercadors web com Google, que utilitzen tècniques de web crawling a gran escala per a descobrir i indexar les pàgines web creades per tercers.

3. Explica breument amb les teves pròpies paraules, tres dels factors que poden influir en l'estimació de la qualitat de les dades. [Màxim 150 paraules]

L'**exactitud** és el grau en que les dades es corresponen realment als valors dels atributs de les entitats que descriuen. Per exemple: un error en les dates de naixement ens presenta una imatge distorsionada de la persona a la que fa referència, i això pot tenir conseqüències per al seu tractament.

La **consistència** persegueix que els atributs no tinguin més d'una representació en la font de dades. Les inconsistències són una font d'errors durant l'anàlisi i interpretació de les dades. Per exemple: mescla de dates en format europeu i nord-americà.

La **completesa** pot definir-se com la proporció de dades amb valors diferents de nul, sempre que aquest no sigui un valor legítim per a l'atribut. Per exemple: no totes les persones tenen segon cognom, així que un valor nul no podria considerar-se incomplet. No així en el cas del nom de pila.

Exercici 2 (40%)

Després de llegir el recurs "Subirats, L., Calvo, M. (2019). Web Scraping", capítols 1 i 7. Contesta les següents preguntes:

1. Quan és legal utilitzar web scraping i quan no? Explica cada cas amb un exemple (màxim 100 paraules).

En síntesi el web scraping és legal quan: respecta les normes jurídiques i termes de servei vigents, no causa danys i fa un ús just de les dades. Per exemple: seria legal emprar web scraping per a estudiar la polarització política analitzant fòrums i diaris digitals. A la Unió

Europea seria il·legal emprar-lo per a recopilar dades personals sense consentiment dels afectats ja que vulnera la normativa de la UE[1].

2. Què són els robots.txt i per què és important tenir-los en compte? (màxim 80 paraules).

Els robots.txt són fitxers que especifiquen restriccions al rastreig dels llocs webs. Els robots.txt indiquen quines pàgines i directoris no haurien de ser accedides pels bots com els webs scrapers. Encara que les indicacions no són obligatòries, els bots han de tenir-les en compte ja que ignorar-les es considera un comportament inapropiat que pot ser sancionat amb el bloqueig del bot per part dels administradors del lloc web i inclús pot tenir conseqüències legals.

3. Expliqui breument i amb les seves pròpies paraules, els passos que utilitzaria per avaluar la dificultat de realitzar web scraping a un cert lloc web i per què realitzaria cada pas. Ressalti amb text tipus "Bold" (negreta) una paraula, per cada pas, que funcioni com a títol per a aquest pas. (màxim 150 paraules).

En la meva opinió, podem dividir el procés d'avaluació en dos blocs:

1) Avaluar la **legalitat** del procés de web scraping ja que, com es va dir a la primera pregunta, hi ha el risc de vulnerar-la. Inclourà l'examen dels termes de servei.
2) Avaluar la **complexitat tècnica** del procés de web scraping. Permetrà valorar la dificultat del seu desenvolupament. Inclourà els següents passos:
   - Determinació dels **obstacles** que el script de web scraping haurà de fer front com presència de robots.txt, ús de mecanismes anti-robot com CAPTCHA, necessitat d'inici de sessió i cookies.
   - Estimació de la **grandària** del lloc web que condicionarà els mètodes de rastreig. Per exemple: necessitat de descarregues en paral·lel.
   - Anàlisi de l'**estructura** i les **tecnologies** de disseny de les pàgines del lloc web que permetrà esbrinar si el format i tipus de continguts de les pàgines facilita o dificulta l'extracció d'informació.

4. Expliqui **dues** bones pràctiques, a l'hora de realitzar web scraping, per evitar ser bloquejat. (màxim 100 paraules).

Una bona pràctica és no saturar de peticions als servidors web, ja que això pot portar a la degradació o interrupció del seu servei. Per a evitar-lo es poden introduir retards entre les peticions del script de web scraping.

Altra bona pràctica és modificar la capçalera HTTP User-Agent ja que les llibreries de web scraping solen establir una pròpia de forma predeterminada, facilitant la seva identificació i eventual bloqueig. És convenient canviar-la per una emprada pels navegadors web més habituals.

[1] Particularment el Reglament General de Protecció de Dades de la UE. S'exceptuen les dades personals obtingudes d'un nombre limitat de fonts públiques com mitjans de comunicació o diaris oficials.