DATASET: ARTICLES I COMENTARIS DEL BLOG FITNESS REVOLUCIONARIO

DESCRIPCIÓ:

Les dades obtingudes en aquesta activitat contenen els enllaços a tots els articles del *blog Fitness Revolucionario* (www.fitnessrevolucionario.com). Aquest *blog* està dedicat, generalment, a la salut. Busca millorar la crisi actual d'obesitat i malalties cròniques del món modern, basant-se en la ciència i l'evolució.

IMATGE IDENTIFICATIVA:

IL·LUSTRACIÓ 1 . WORD CLOUD DELS TÒPICS DEL BLOG FITNESS REVOLUCIONARIO

CONTEXT:

Les dades contenen la *url*, la data, el títol, el número de comentaris i el text dels comentaris (anònims). També conté les referències dels articles entre ells (enllaços). Els articles parlen de temes com la nutrició, l'esport, la connexió social i, sobretot, la salut. Amb aquest conjunt de dades es busca, d'una banda, tenir una idea general dels tòpics del *blog* i de la xarxa d'articles que el conformen. De l'altre, analitzar la satisfacció dels usuaris lectors a partir dels seus comentaris.

CONTINGUT:

Cada fila del fitxer '*fitness_revo_full.csv*' conté les característiques d'un article, amb un total de més de 300 articles publicats des de l'any 2011. Aquestes característiques són:

- **Url:** enllaç a la pàgina web de l'article.

- **Date:** data de publicació de l'article.
- **Title:** títol principal de l'article.
- **numComment:** número de comentaris fets pels usuaris sobre l'article.
- **fullComments:** el text de tots els comentaris de l'article, en minúscula i amb filtrat de caràcters no útils. Cada comentari està separat per el caràcter ' * ', per facilitar-ne la lectura amb *split*.

El fitxer *referencies.csv* conté, a cada fila els enllaços als que cada un dels articles fa referència dins el mateix domini. Al *referencies_encoded.csv* es troba el mateix en forma de transacció (TRUE/FALSE), a punt per utilitzar per algorismes de ML.

Finalment, el fitxer *comment_polarity.csv* conté una llista dels sentiments transmesos als comentaris del blog. S'han utilitzat només 784 comentaris per a generar les dades, corresponents als 20 primers articles.

Agraïments:

Vull agrair, primer de tot, a l'autor del *blog Fitness Revolucionario* Marcos Vázquez, per la seva feina i per permetre l'accés al *scraper* i a un *sitemap* ben estructurat.

Tanmateix, agrair a tota la comunitat informàtica per la seva imprescindible ajuda a l'hora de resoldre dubtes i per les fonts d'aprenentatge disponibles navegant per Internet. També, als desenvolupadors dels paquets en llenguatge Python i R que s'han fet servir per l'extracció i anàlisi de les dades.

Finalment, agrair als creadors dels recursos brindats per la UOC, quan han estat de molta utilitat.

Inspiració:

Fitness Revolucionario dóna un enfocament, val a dir, revolucionari, per a millorar la salut de la població moderna en general. Actualment, gran part de la població no té coneixements suficients o està erròniament informada sobre com gaudir d'una bona salut física i mental.

Amb aquest conjunt de dades es pretén, d'una banda, donar una visió general del *blog* i dels temes tractats, així com de la seva evolució al llarg dels anys. De l'altra, analitzar les reaccions dels lectors per mesurar-ne el grau de satisfacció. Els principals interessats serien, per descomptat, l'autor del *blog*, així com els mateixos lectors o futurs lectors del *blog*.

Finalment, i en un àmbit més acadèmic, les dades serveixen per aprofundir en àrees com el *text mining* o l'anàlisi de sentiments. Un exemple seria el *pie chart* que resumeix la satisfacció dels lectors a partir del fitxer *comment_polarity.csv*. El mateix enfocament es podria reutilitzar per a saber, sobre qualsevol tòpic, què en pensa la gent en un moment donat.

Llicència:

Així, aquesta llicència permet compartir i transformar les dades, però sempre donant crèdit al seu autor, al propietari original de les dades i sense fer-ne un ús comercial.

CODI:

El codi font es troba a la carpeta */src* en forma de *notebook* de Python. També hi ha un *script* de R que permet executar l'algorisme *apriori* sobre el conjunt de referències esmentat.

A més, a la carpeta */html* hi ha el mateix codi transformat en format *html*. Si per algun motiu no es pogués accedir als *notebooks*, es recomana utilitzar el següent enllaç per obrir-los:

https://nbviewer.jupyter.org/

Els *notebooks* estan executats i documentats, per facilitar-ne la comprensió.

DATASET:

Les dades extretes i preparades es troben contingudes en els fitxers *.csv* de la carpeta */csv*.

D'altra banda, a la carpeta */img* hi ha les imatges obtingudes durant els diferents processos executats. Finalment, a la carpeta */html* es pot accedir a uns gràfics interactius que descriuen la xarxa d'articles del *blog*.