Dataset: Receptari culinari d'Hogarmania (+10.000 receptes disponibles)

Aleix Arnau Soler & Adrià Tarradas Planella

9 de Novembre de 2020

1. Context

La societat actual està marcada per un ritme de vida frenètic, amb horaris laborals, moltes ocupacions i on, en molts casos, mares i pares han de a més, compaginar tot això amb cuidar i atendre els fills. Tot junt fa que en moltes llars sigui pràcticament impossible dedicar gaire temps a la cuina. La falta de temps, preparació, o simplement coneixement, fa que en moltes ocasions s'acabi optant per preparar menjar ràpid o demanant menjar a domicili, amb el risc afegit de que la nostre dieta, i la dels nostres fills, no sigui lo equilibrada que tots desitjariem. Però existeixen un gran número de receptes de cuina que podem elaborar de manera ràpida per als nostres menús diaris. En molts casos, el problema és simplement que es desconeixen els plats que es poden elaborar amb allò que es té a disposició, o quins plats nutritius i equilibrats es poden elaborar si es disposa de poc temps. O simplement que es desconeixen els ingredients i els passos necessaris per a elaborar un plat d'interès. La web **Hogarmania** conté una secció de cuina amb receptes presentades en programes de televisió a Espanya durant més d'una dècada per cuiners reconeguts a nivell nacional com Karlos Arguiñano, entre altres. Aquest lloc web conté totes les dades i material multimèdia necessari per a la creació d'un receptari complet i equilibrat.

2. Títol pel dataset

El títol del dataset és "Receptari Hogarmania" ja que recull informació de totes les receptes de cuina de la pàgina web Hogarmania.

3. Descripció del dataset

El dataset generat és un recopilatori de més de 10.000 receptes de cuines extretes de la secció de cuina de la web **Hogarmania** i elaborades per varis cuiners (la majoria en programes de televisió). Aquest dataset proporciona les eines necessàries per a procedir de forma exitosa en el món de la cuina, amb un registre extens de la nostre memoria i patrimoni culinari. Cadascuna de les receptes enregistrades s'acompanya (si es troba disponible) del vídeo (1 vídeo resum i un vídeo amb la versió completa) que al seu dia va ser emès per televisió on un cuiner de referència a nivell nacional explica tots els passos a seguir per a l'elaboració de la recepta, pas a pas. Per a cadascuna de les receptes (files) es disposa de les següents variables (columnes): nom de la recepta, categoria a la qual pertany la recepta, ingredients de la recepta, temps de preparació, temps total, nom del cuiner, pasos de l'elaboració, composició nutricional, consell medic, imatge de la recepta, video resum amb els passos de l'elaboració, video complet de l'elaboració de la recepta.

4. Representació gràfica

A la Figura 1 es pot veure la representació gràfica escollida, la qual consisteix en un *collage* que agrupa imatges representatives de cada tipus de recepta (carns, sopes, còcktails, etc.).

Figura 1: Mosaic amb imatges d'algunes de les receptes disponibles

5. Contingut

Per a cada recepta (les quals corresponen amb cada fila or registre del conjunt de dades) es disposa de la informació següent en forma de variables o columnes:

- **titol**: Títol/nom de la recepta.
- **ingredients**: Ingredients necessaris per a l'elaboració de la recepta.
- **temps_preparacio**: Temps necessari per a la preparació de la recepta.
- **temps_total**: Temps total necessari per a la preparació i cocció de la recepta.
- **chef**: Cuiner encarregat d'elaborar i explicar la recepta.
- **elaboracio**: Pasos a seguir per a l'elaboració de la recepta.
- **info_nutricional**: Informació nutricional de la recepta.
- **consells_doctor**: Consells per part de personal especialitzat (doctors i dietistes) en relació a la recepta.
- **link_imatge**: Imatge de la presentació final del plat.
- **link_video_resum**: Video resum (3-5 min) dels pasos a seguir per a l'elaboració de la recepta.
- **link_video_complet**: Video complet de l'elaboració de la recepta (tal qual es va emetre per televisió).
- **tipus_recepta**: Categoria a la que pertany la recepta (i.e. carns, pasta, postres, etc).
- **persones**: Quantitat de persones que es podran alimentar amb l'elaboració del plat que explica la recepta.

Tota la informació recopilada s'emmagatzema a un fitxer csv amb el nom "recipes_dataset_AAAAMMDD_HHMMSS", on la part final del nom es correspon al timestamp de la data en la que ha estat creat. Això fa que sigui possible identificar a quin moment es van recopilar les dades i, per tant, saber la data en la que la informació del dataset es correspon exactament al contingut de la web.

6. Agraïments

La titularitat de les dades correspon exclusivament a COMUNICACIÓN S.A. Totes les dades han sigut extretes a partir de la informació disponible en diferents *sitemaps* de l'apartat de receptes de la secció de cuina del web **Hogarmania** Per a l'extracció de dades s'ha fet ús del llenguatge de programació Python i els mòduls disponibles per a dur a terme tècniques de *Web Scraping* com són *requests* o *beautifulsoup4*.

7. Inspiració

Un receptari com el que es troba disponible en aquest dataset té un gran ventall d'aplicacions ja que permet responde a un gran número de preguntes. Per un costat és una guia molt àmplia i detallada de la memoria culinaria del nostre país, amb la informació detallada i necessària per a l'elaboració de cadascuna de les receptes recopilades. Però, i això és el que el fa més important, les dades enregistrades ens permeten filtrar i seleccionar d'entre les més de 10.000 receptes disponibles, aquelles receptes que poden ser d'interès en un moment concret. Així doncs, el dataset ens permetria, per exemple, filtrar i seleccionar aquelles receptes que podríem elaborar només amb allò que tenim disponible a la nevera. De manera similar, empreses de neveres intel · ligents amb implementació de IoT poden implementar un sistema on la teva propia nevera et suggereix plats a elaborar basant-se amb els ingredients que tens disponibles. De forma similar, es pot filtrar per plats d'elaboració curta, si es disposa de poc temps, o partir de referències mèdiques o nutricionals. De forma similar, dietistes i/o usuaris poden utilitzar les dades disponibles en aquest dataset per a elaborar dietas nutricionalment equilibrades, filtran la informació a partir de la informació nutricional de la recepta. Totes les receptes, a més, venen amb vídeos que ajuden als usuaris a elaborar les receptes de forma fàcil i entretinguda.

Per posar un cas pràctic. Una pare o mare pot arribar a casa i preguntar-se: quin plat podria elaborar per a menjar jo i els meus dos fills, que fos nutritiu i gustos, amb menys de 40 minuts que tinc disponibles abans de tornar a la feina, i a partir dels ingredients que tinc a mà a casa? I és més, en l'escenari de que es sàpiga el plat que es vol o es pot cuinar, quins són els passos per a cuinar aquesta recepta? El dataset proposat permetria trobar una o varies receptes d'entre les disponibles que s'adequa als requeriments plantejats i els ingredients disponibles. A més, disposa dels passos i vídeos per a l'elaboració del plat de manera senzilla i satisfactòria.

8. Llicència

Aquest dataset està publicat sota la llicència **Creative Commons Reconeixement-NoComercial-SenseObraDerivada** (CC BY-NC-ND 4.0 License). Els termes legals que fan referència a la llicència es poden consultar a https://creativecommons.org/licenses/by-nc-nd/4.0/.

Aquesta llicència és la més restrictiva de totes. El motiu principal que ens ha portat a aplicar una llicència tant restrictiva és degut a que la "Propiedad intelectual y Derechos de Copyright" de la secció d'"aviso legal" del web especifica que:

"*Queda prohibida la reproducción, distribución, puesta a disposición, comunicación pública y utilización total o parcial, de los contenidos (imágenes, textos y vídeos) de Hogarmania.com, en cualquier forma o modalidad, dispositivo analógico o digital, sin previa, expresa y escrita autorización, incluyendo, en particular, su mera reproducción y/o puesta a disposición como resúmenes, reseñas o revistas de prensa con fines comerciales o directa o indirectamente lucrativos, a la que BAINET COMUNICACIÓN S.A. manifiesta oposición expresa.*"

Vam intentar comunicar-nos amb el propietari de les dades, però a dia d'avui encara no em rebut resposta i per tal d'evitar un impacte negatiu del nostre treball cap al propietari de les dades n'oferim uns usos limitats.

Els punts més importants que caracterítzen aquesta llicència són:

- **BY - Reconeixament**: En cas de redistribució del dataset, caldrà donar crèdit als autors i al propietari de les dades
- **NC - No comercial**: El dataset no pot ser utilitzat amb fins comercials
- **ND - Sense obres derivades**: No es pot publicar cap obra derivada del dataset

9. Codi

El codi utilitzat per a l'extracció de dades del lloc web es troba disponible a:

- **Repositori Aleix:** https://github.com/AleixArnauSoler/recepy
- **Repositori Adrià:** https://github.com/latp/recepy

10. Dataset

El dataset generat es troba disponible a la carpeta corresponent del **Github** i a **Zenodo**.

- El DOI del dataset és **10.5281/zenodo.4265137**:

11. Contribució

Taula 1: Tots els membres han contribuït de manera igualitaria en l'elaboració de la pràctica.

Contribucions	Signa
Recerca prèvia	AAS & ATP
Redacció de les respostes	AAS & ATP
Desenvolupament codi	AAS & ATP

12. Bibliografia

- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
- Masip, D. El lenguaje Python. Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons.
- Tutorial de Github https://guides.github.com/activities/hello-world.