

# PRAC 1-Web Scrapping

Jordi Boldú Millà

## 1. Títol del data set. Cal que poseu un títol que sigui descriptiu.

Objectes de l'espai profund agrupats per constel·lació

## 2. Subtítol del data set


Estrelles, quàsars, cúmuls, galàxies i nebuloses de poca o escassa visibilitat des de la terra agrupats segons la constel·lació a la que pertanyen

## 3. Imatge descriptiva




Hubble Deep UV (HUV) Legacy Survey image (15k galaxies; August 16, 2018)

 More details

 NASA, ESA, P. Oesch (University of Geneva), and M. Montes (University of New South Wales) -  
[https://www.nasa.gov/sites/default/files/thumbnails/image/image1\\_-\\_stsci-h-p1835a-m-2000x1768.png](https://www.nasa.gov/sites/default/files/thumbnails/image/image1_-_stsci-h-p1835a-m-2000x1768.png)

 Public Domain

 File: NASA-Galaxies15k-HubbleHUV-20180816.png

 Created: 16 August 2018

Aug. 16, 2018 - NASA - Hubble Paints Picture of the Evolving Universe<sup>[1]</sup>  
<https://www.nasa.gov/feature/goddard/2018/hubble-paints-picture-of-the-evolving-universe>  
<https://www.cnet.com/news/nasas-hubble-telescope-captures-15000-galaxies-in-one-dazzling-image/> Astronomers using the ultraviolet vision of NASA's Hubble Space Telescope have captured one of the largest panoramic views of the fire and fury of star birth in the distant universe. The field features approximately 15,000 galaxies, about 12,000 of which are forming stars. Hubble's ultraviolet vision opens a new window on the evolving universe, tracking the birth of stars over the last 11 billion years back to the cosmos' busiest star-forming period, which happened about 3 billion years after the big bang. Ultraviolet light has been the missing piece to the cosmic puzzle. Now, combined with infrared

[About this interface](#) | [Discussion](#) | [Help](#)

## 4. Context

Les dades obtingudes fan referència al què s'anomenen objectes d'espai profund<sup>1</sup> (*deep space objects*, en anglès) que és un terme que es fa servir en astronomia amateur per referir-se als objectes celestes que no són del Sistema Solar i que, normalment, no són visibles a simple vista tot i que, els més brillants, es poden veure amb un petit telescopi o fins i tot amb uns binoculars prou potents

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

### NOTA

Existeixen diferents catàlegs (*Messier, NGC, UGC, IC, Caldwell, Sharpless, Herschel, Collinder, Mellote, ...*), que a vegades fan referència als mateixos objectes però amb noms diferents.

Això es deu, principalment, a dos motius:

- alguns d'aquests catàlegs s'han elaborat en èpoques diferents i els posteriors inclouen objectes ja descoberts anteriorment però seguint altra nomenclatura
- alguns d'aquests catàlegs només agrupen objectes d'uns determinats tipus (nebuloses, quàsars, ...) i també segueixen una nomenclatura pròpia.

Sovint, per ajudar a la seva identificació, és costum referir-se a l'objecte concatenant els diferents noms dels catàlegs més importants

Per a cada objecte d'espai profund desat en el nostre data set es proporcionen els següents camps

- **nom complet de l'objecte** : concatenació dels diferents noms que rep l'objecte d'espai profund segons els diferents catàlegs (*si s'escau*)
- **codi catalogació 1** : primer nom que rep l'objecte. No s'especifica el catàleg. Tot objecte té, com a mínim, un nom que l'identifica.
- **codi catalogació 2** : segon nom que rep l'objecte (*si s'escau*)
- **magnitud** : magnitud aparent<sup>2</sup> de l'objecte profund
- **tipus** : el tipus de l'objecte d'espai profund (nebulosa, galàxia, ...)
- **tamany** : tamany de l'objecte d'espai profund mesurat en minuts o segons sexagesimals<sup>3</sup>
- **ascensió recta o rectal (AR)** : coordenades equatorials per a localitzar l'objecte d'espai profund sobre l'esfera celeste<sup>4</sup>
- **declinació** : distància angular d'un astre sobre l'equador celeste i que equival a la latitud sobre la terra<sup>5</sup>
- **constel·lació** : nom en llatí de la constel·lació on està ubicat l'objecte d'espai profund

<sup>1</sup> [https://ca.wikipedia.org/wiki/Objectes\\_de\\_l'espai\\_profund](https://ca.wikipedia.org/wiki/Objectes_de_l'espai_profund)

<sup>2</sup> [https://ca.wikipedia.org/wiki/Magnitud\\_aparent](https://ca.wikipedia.org/wiki/Magnitud_aparent)

<sup>3</sup> [https://es.wikipedia.org/wiki/Segundo\\_sexagesimal](https://es.wikipedia.org/wiki/Segundo_sexagesimal)

<sup>4</sup> [https://ca.wikipedia.org/wiki/Ascensió\\_recta](https://ca.wikipedia.org/wiki/Ascensió_recta)

<sup>5</sup> [https://ca.wikipedia.org/wiki/Declinació\\_\(astronomia\)](https://ca.wikipedia.org/wiki/Declinació_(astronomia))

# PRAC 1-Web Scraping

Jordi Boldú Millà

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

Les dades factuais dels diferents objectes d'espai profund s'han extret del lloc web **Atlas de Astronomía**

<http://atlasdeastronomia.com/>

Les imatges de les diferents constel·lacions han estat extretes del lloc web **International Astronomical Union**

<https://www.iau.org/>

La imatge per il·lustrar el data set ha estat extreta del lloc web **National Aeronautics and Space Administration** (NASA)

<https://www.nasa.gov>

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

Des de sempre m'ha impressionat tot allò que té a veure amb l'exploració de l'espai i els fenòmens físics que s'hi poden trobar: forats negres, quàsars, etc. També m'agrada molt la ciència ficció, ja siguin llibres, documentals, pel·lícules o sèries de televisió que fan divulgació o fantasiegen en viatjar per l'espai, descobrir i colonitzar nous mons, etc.

En aquesta pràctica he trobat una oportunitat per poder apropar dues de les meves aficions: l'espai i la programació

Tot i que aquest data set té una finalitat purament il·lustrativa, podria servir, per exemple, per:

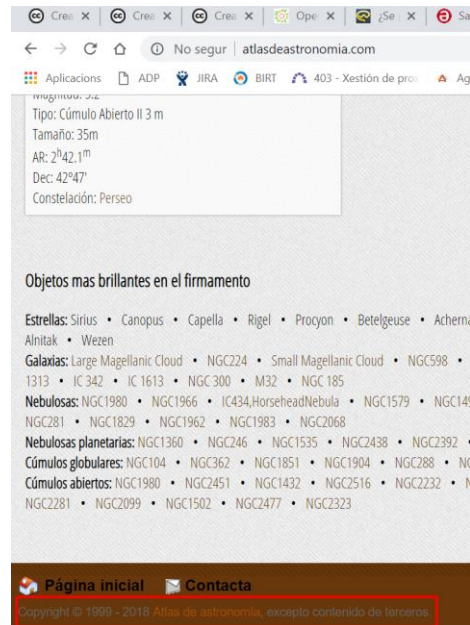
- **Cerca de patrons:** estudiar si la distribució de tipus d'objectes d'espai profund segueix algun patró en funció de les seves característiques i les de la constel·lació a la que pertanyen
- **Estadística descriptiva:** extracció de mitjanes, medianes, ... dels diferents atributs

## 8. Llicència

Malauradament no he pogut determinar un tipus concret de llicència pel nostre data set i, per tant, no em queda altre opció que seleccionar el tipus **Unknown license**

Detallo a continuació els passos que he seguit i que justifiquen aquesta conclusió:

El lloc web **Atlas de astronomia** sobre el què hem dut a terme l'*scraping* (<http://atlasdeastronomia.com/>) conté una llegenda al peu de cada pàgina on hi posa **"Copyright © 1999 - 2018 Atlas de astronomía, excepto contenido de terceros"**, sense cap més referència o crèdits que especifiqui les fonts fetes servir tot i que, pel cal concret de les imatges es referencien les originals que es troben al lloc web **International Astronomical Union** (<https://www.iau.org/>) i que sí que estan catalogades sota llicència **Creative Commons Attribution 4.0 International license** (<https://www.iau.org/copyright/>)



Les dades que apareixen al lloc web semblen extretes d'algun(s) catàleg(s) astronòmic(s) públic(s) (no hi ha informació reservada ni confidencial) i, segons el fitxer **robots.txt** (<http://atlasdeastronomia.com/robots.txt>) no hi ha pràcticament cap restricció per a dur a terme *scraping* sobre el lloc web

```
← → ↻ 🏠 No segur | atlasdeastronomia.com/robots.txt
Aplicacions ADP JIRA BIRT 403 - Xestión de pro Agile Documentaci
User-agent: * # aplicable a todos
Disallow: /error404.php # impide la indexacion de la pagina de error 404.
Sitemap: http://atlasdeastronomia.com/sitemap.xml
```

Per determinar si existia alguna pàgina web dins de tot el lloc web on pogués aparèixer alguna referència a l'autoria dels continguts, als crèdits o la llicència, he analitzat el fitxer **sitemap** (<http://atlasdeastronomia.com/sitemap.xml>) no fos cas que la pàgina existís però l'accés no fos evident sense obtenir tampoc cap resultat concloent

Finalment, he donat un cop d'ull al codi font de la pàgina principal i he vist que existeix una adreça per accedir al lloc web via RSS

```
<!DOCTYPE html>
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta charset="utf-8">
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
<title>Atlas de astronomia</title>
<meta name="title" content="Atlas de astronomia" />
<meta name="description" content="Atlas de astronomia es una guía para poder observar el ci
objetos de cielo profundo, que se actualiza diariamente." />
<meta name="keywords" content="astronomia" />
<meta name="referrer" content="" />
<meta name="viewport" content="width=device-width, initial-scale=1, maximum-scale=1">
<link rel="stylesheet" type="text/css" href="/styles/common.css" />
<link rel="shortcut icon" type="image/ico" href="/favicon.ico" />
<link rev="start" href="http://atlasdeastronomia.com/" />
<link rel="copyright" href="http://atlasdeastronomia.com/about.php" />
<link rel="appendix" href="http://atlasdeastronomia.com/about.php" />
<link rel="alternate" type="text/xml" title="RSS feed" href="/index.xml" />
<link rel="alternate" type="application/rss+xml" title="RSS feed" href="/index.xml" />
</head>
<body>
<header>
<div id="header">
<h1><a href="/" title="AtlasDeAstronomia.com">Atlas de astronomia</a></h1>
</div>
</header>
```

# PRAC 1-Web Scraping

Jordi Boldú Millà

A l'accedir a l'adreça especificada per connectar via RSS (<http://atlasdeastronomia.com/index.xml>), apareix un error en la càrrega del fitxer índex especificat però es pot apreciar un paràgraf de text, darrera la descripció del lloc web que hi posa **Creative Commons (CC) 2018**

This page contains the following errors:

error on line 28 at column 141: Extra content at the end of the document

Below is a rendering of the page up to the first error.

Atlas de astronomía <http://atlasdeastronomia.com> Atlas de astronomía es una guía para poder ob profundo, que se actualiza diariamente. Creative Commons (CC) 2018 Atlas de astronomía info@ 12:45:49 PDT Fri, 2 Nov 2018 12:45:49 PDT Search posts Search posts containing the text... s l: <http://atlasdeastronomia.com/images/logo.jpg> <http://atlasdeastronomia.com> 272 56 Atlas de astr estrellas y los objetos de cielo profundo, que se actualiza diariamente. Invalid Query: SELECT \* 'atlasdeastronomia.posts' doesn't exist

Donat que la llicència del lloc web no s'especifica de manera clara i evident, no ens queda altra que contactar amb els autors del lloc web i demanar-los consentiment per escrit. Mentre aquest punt no quedi aclarit, la llicència triada és de tipus **Unknown license**

## 9. Codi

El codi s'inclou a l'arrel del repositori de *github*<sup>6</sup>

## 10. Data set

El data set s'inclou a la carpeta **dataset** del repositori

S'hi pot trobar el data set en format CSV i en format XLSX (Excel)

Adicionalment, he extret imatges de les diferents constel·lacions treballades (88) i s'han desat a la carpeta **img** del repositori

<sup>6</sup> <https://github.com/JordiBolduMilla/deepSpaceObjectScraper>

# PRAC 1-Web Scraping

Jordi Boldú Millà

## Bibliografia

Minguillon, J. (2016). Fundamentos de Data Science. Editorial UOC.

Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 1. Introduction to Web Scraping.

Masip, David (?). El llenguatge Python. Editorial UOC.

## Llocs web

<http://atlasdeastronomia.com/> (Atlas de Astronomía)

<https://www.iau.org/> (International Astronomical Union)

<https://www.nasa.gov> (National Aeronautics and Space Administration, NASA)

<https://ca.wikipedia.org> Viquipèdia

<https://guides.github.com/activities/hello-world> Tutorial de Github