Tipologia i cicle de vida de les dades

Màster universitari de Ciència de Dades

Pràctica 1: Com podem capturar les dades de la web

Autor: Adrià Setó Balcells

Autor: Miquel Arisa Fuente

Professorat: Mireia Calvo González

Universitat Oberta de Catalunya

Setembre, 2022

ÍNDEX

1. Context

Tothom sap que les cases d'apostes esportives, igual que tots els altres jocs d'apostes, estan dissenyats perquè el jugador normalment perdi. Això s'aconsegueix a través de l'estadística i de la psicologia de les persones, en aquest treball ens centrarem en la part on el jugador surt perdent per culpa de l'estadística. El que acostumen a fer els jocs d'apostes és fer veure al jugador que està ficant els diners a una aposta "justa", on per exemple les possibilitats de guanyar són del 50%, i els beneficis són multiplicar per dos els diners de l'aposta del jugador, un tracte bastant just. Però aquesta és una falsa il·lusió, ja que les possibilitats sempre són una mica inferiors al cinquanta. Per exemple la ruleta, que està formada pels números del 0 al 36, on hi ha divuit nombres vermells i divuit nombres negres, on apostant al color dona la sensació de apostar a una aposta cinquanta a cinquanta, però el que normalment no es pensa és que el 0 és un color diferent, no és ni parell ni senar, ni de cap fila ni de cap regió, així que a no sé que el jugador aposti al 0 aquest fa perdre tot arreu, convertint així les possibilitats de guanyar d'un 50% a un 48,64%. Així que al llarg del temps la casa d'apostes guanya un 1,36% per a cada aposta teòrica de cinquanta a cinquanta, i tot això només sense comptar amb la mentalitat d'un jugador que tendeix a ficar més diners quan perd i continuar apostant quan guanya, però això ja són temes psicològics.

A les cases d'apostes el tracte és el mateix, podem veure les possibilitats en tant per u de que una aposta sigui guanyadora (segons la casa d'apostes) dividint 1 entre la quota de cada aposta. Aprofitant l'inici del mundial de futbol posem un exemple d'alguna de les apostes a la Figura 1.

Figura 1. Captura de pantalla d'una aposta de la casa d'apostes Bet365

$$\frac{1}{3{,}40} = 0{,}294 \quad \frac{1}{3{,}20} = 0{,}313 \quad \frac{1}{2{,}25} = 0{,}444$$

Podem veure com si fem el sumatori d'1 entre de totes les tres possibilitats que hi ha dins del partit de futbol (1, X, 2) ens donarà el percentatge que guanya la casa d'apostes.

$$\frac{1}{3{,}40} + \frac{1}{3{,}20} + \frac{1}{2{,}25} = 1{,}051$$

El resultat de la següent operació l'anomenarem viabilitat, i en aquest cas és d'un 1,051. Sabem que el punt d'equilibri (on la casa d'apostes no guanyaria ni perdria res teòricament) és 1, el que haurien se sumar tots els percentatges de possibilitats perquè el joc fos just, així que tot el que estigui per sobre d'aquest llindar seran beneficis per la casa, i tot el que estigui per sota seran pèrdues. En aquest cas la casa rep uns beneficis teòrics d'un 5,1%.

El nostre objectiu d'aquesta pràctica és buscar errors o diferències de criteris entre diferents cases d'apostes per tal de trobar apostes contràries com la vista anterior, on hi ha el 100% de possibilitats que una d'elles sigui guanyadora i que el resultat de la suma del percentatge les seves quotes sigui inferior a 1. D'aquesta manera podrem assegurar uns beneficis sense córrer cap mena de risc. Com a cas teòric, si trobéssim un cas on la viabilitat fos d'un 0,98, tindríem un benefici d'un 2% dels diners apostats assegurats.

Aquest és un projecte gros, que necessita d'unes molt bones anàlisis de múltiples cases d'apostes i de diversos esports dins de cada una d'elles, ja que cada un funciona d'una forma diferent. Per aquest motiu hem decidit començar per una casa d'apostes anomenada Marathonbet, la qual es caracteritza per oferir unes quotes bastant interessants als clients, i dins d'aquesta casa començar a extreure dades d'un esport, el tenis. El motiu pel qual hem decidit començar a extreure dades per aquest esport és perquè hi ha moltes apostes dins del mon del tennis i la gran majoria d'elles són binaries, com per exemple si hi ha més o menys de n jocs, sets, etc. Per això comencem fent el scraping de la següent url: https://www.marathonbet.es/es/popular/Tennis

2. Títol

El títol escollit és el següent:

bet_quotes_dataset

3. Descripció del dataset

El conjunt de dades extret correspondrà als mercats d'apostes de tennis disponibles a Marathonbet en el moment de l'execució de l'script. D'aquests se n'obtindrà el nom de l'esport, per si en un futur es volguessin tractar esports addicionals a banda del tennis; els camps descriptius de l'esdeveniment esportiu, que serien els dos participants (ja siguin equips o individuals); i finalment la informació dels mercats, amb les apostes i quotes corresponents que ofereix la casa d'apostes.

L'objectiu d'aquest *dataset* és comparar-lo amb altres similars, però de cases d'apostes diferents, com per exemple Winamax, Bet365, William Hill, etc. en busca del nostre objectiu, trobar diferències de criteris de les diferents cases d'apostes i poder aconseguir la viabilitat explicada anteriorment per sota de 1.

Amb l'estructura actual del *dataset*, només s'admetrien esdeveniments esportius amb dos participants (jugadors o equips), ja que aquests s'inclouen com a informació de l'esdeveniment en els camps local/visitant.

Un mercat englobarà el conjunt de resultats possibles per una aposta en concret (per exemple *1*, *X*, *2* en la Figura 1) i el *dataset* contindrà una fila per cada possibilitat d'aposta. Degut a això, un mercat estarà representat per *n* files al dataset, on *n* serà el nombre corresponent d'apostes possibles per aquest mercat.

Com a exemple, a la Figura 2 es pot veure un conjunt de mercats de total de jocs en un partit de tennis, que es tracta d'una aposta binària amb dos valors possibles. Cada fila d'aquesta taula correspon a un mercat, i cadascun té dues apostes. El *dataset* generat contindria una fila per cada aposta, on cada parell contindria el mateix identificador de mercat per poder-les identificar com a conjunt. El mercat enquadrat en vermell seria "Més/Menys 19.5 jocs" i generaria dues files al *dataset*, una per l'aposta de "Menys 19.5 jocs" amb quota 4,45 i l'altra per la de "Més 19.5 jocs" amb quota 1,153.

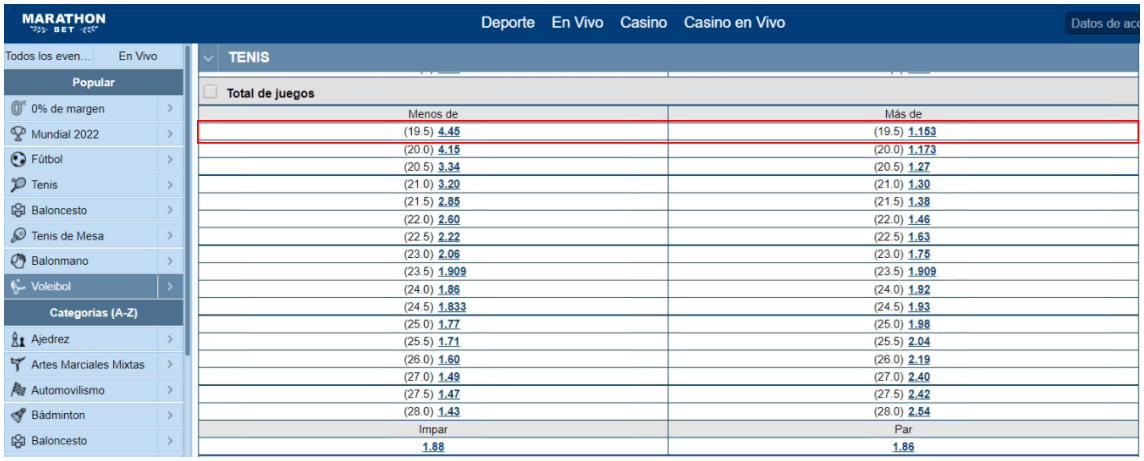

Figura 2. Captura de pantalla d'un exemple de mercat binari a la pàgina Marathonbet

4. Representació gràfica

En la Figura 3 es pot veure un diagrama de flux de com es mou el programa a través de la pàgina web buscant les classes i els atributs dels elements i iterant per a cada un d'ells. Aquest diagrama conté des de que s'entra a la url donada, on agafa totes les altres urls de cada un dels partits i entra dins d'aquests links buscant els mercats amb les diferents apostes i quotes.

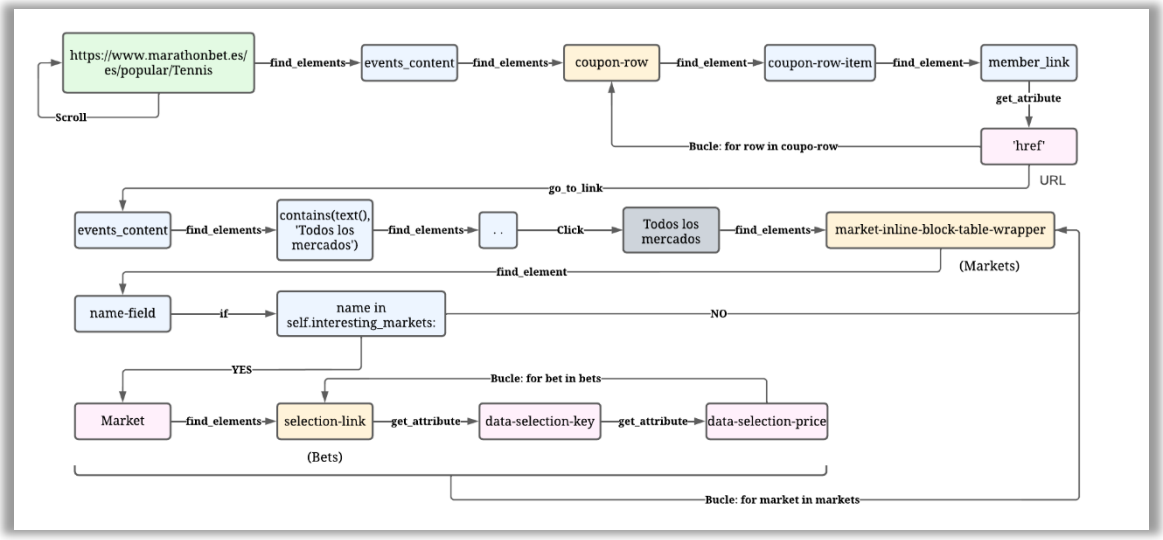

Figura 3. Diagrama de flux de l'script

En la Figura 4 es mostra el diagrama del projecte final amb el *dataset* complert, que consisteix en un set de cases d'apostes (BettingSites), on cada una d'elles té un set d'esports (Sports). Cadascun d'aquests esports té un set de partits o esdeveniments (Events), tots aquests tenen un set dels mercats (Markets) i per cada mercat hi ha diferents apostes (Bets) amb la seva quota associada (Quote).

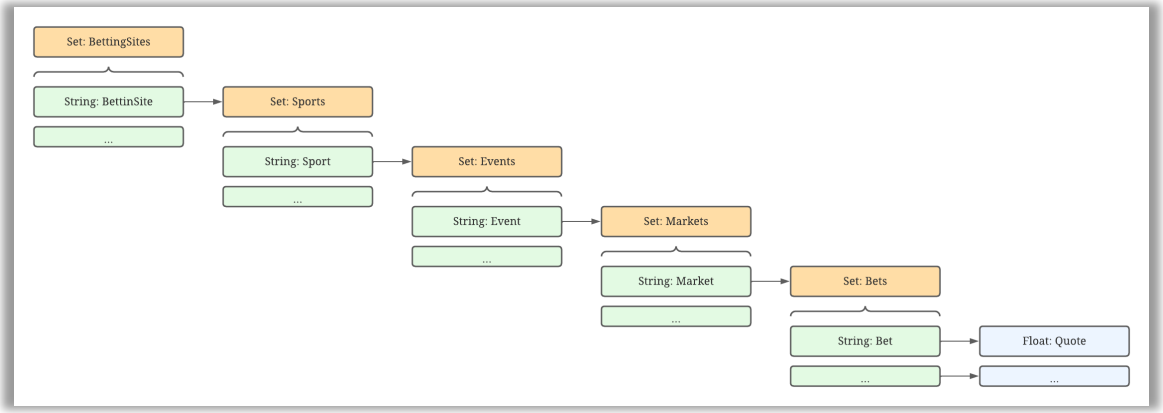

Figura 4. Diagrama de l'estructura de dades del dataset

5. Contingut

El *dataset* inclou els següents camps.

- **BettingSite**: Nom de la casa d'apostes d'on s'han extret les dades.
- **Sport**: Nom de l'esport.
- **Event**: Nom descriptiu de l'esdeveniment esportiu al qual correspon l'aposta. Està format pel nom del local, la paraula "vs" i el nom del visitant.
- **Local**: Nom del jugador/equip que actua com a local al partit.
- **Visitor**: Nom del jugador/equip que actua com a visitant al partit.
- **Market**: Mercat de l'aposta, que agruparà tots els possibles resultats.
- **Bet**: Identificador de l'aposta dins del mercat.
- **Quote**: Quota que ofereix la casa en el cas d'aposta guanyadora.
- **ScrapDateTime**: Data i hora en la que es va fer l'scrapping de l'aposta, amb el format YYYY-MM-DD hh:mm:ss.nnnnnn (UTC).

Les dades del *dataset* proporcionat són obtingudes al moment d'execució de l'script BetScraping, que també apareix com a sufix del nom de l'arxiu csv (22/11/2022 19:42:52 hora local). Aquesta informació s'inclou també al dataset, amb l'hora exacta detallada fila per fila al camp **ScrapDateTime**.

El període de temps d'aquestes dades és relativament curt, ja que les quotes de les apostes poden anar variant depenent de les situacions ocorregudes. En apostes relatives a esdeveniments més pròxims al present, els valors solen ser més variables, i en apostes relatives a esdeveniments més futurs solen ser més estables. A més a més un cop ha acabat l'esdeveniment aquestes dades ja no canvien sinó que desapareixen.

6. Propietari

Pel projecte el qual estem desenvolupant estem agafant dades de cases d'apostes esportives, començant per la primera que és Marathonbet. L'activitat que es pretén dur a terme és coneguda dins el món de les apostes com a Surebets. Algú ja ha dut a terme aquest projecte i s'ha dedicat a vendre aquestes apostes segures a tercers pel seu propi lucre, cosa que ens sembla reprovable. Aquest no és el nostre objectiu, sinó que el nostre és merament una investigació escolar per veure l'estadística de les cases d'apostes i com hi pot haver una manera de que aquestes no sempre guanyin.

Per actuar amb els principis ètics i legals, no ens lucrarem d'aquestes dades obtingudes d'altres companyies que a través dels seus estudis estadístics defineixen aquestes quotes, ni tampoc farem la publicació d'aquestes dades perquè algun tercer les pugui utilitzar i

lucrar-se ell sí, amb aquestes dades a través de la informació de les cases d'apostes i de la nostra obtenció. D'aquesta manera aquest estudi es quedarà simplement com a un projecte universitari.

7. Inspiració

L'obtenció i estudi de tot aquest conjunt de dades és realment interessant ja que pot arribar a respondre una pregunta que moltíssima gent s'ha fet: com es pot guanyar a les cases d'apostes? Perquè sempre tenen l'as de guanyar?

Des de la nostra anàlisi hem pogut veure que en el cas de les cases d'apostes esportives se'ls pot girar la truita. També és veritat que els beneficis obtinguts serien d'un percentatge bastant baix, que calculem que oscil·larien entre un u i un cinc per cent. Així doncs, només podria funcionar de manera exitosa disposant d'un capital econòmic considerablement gran o efectuant moltes repeticions d'apostes i fent créixer el capital invertit exponencialment, cosa difícil perquè les cases d'apostes alentirien el procés a través dels ingressos i les retirades.

Hem vist com la gent que disposa d'aquesta informació el que sol fer és vendre les dades a través de plataformes com Telegram, cosa que fa sospitar que si és que han de vendre la informació no en deuen treure tan benefici.

8. Llicència

Pel *dataset* generat s'ha escollit la llicència **Released Under CC BY-NC-SA 4.0** de Creative Commons.

Figura 5. Released Under CC BY-NC-SA 4.0 License

Aquesta elecció es justifica pel fet que les dades extretes de la casa d'apostes són propietat d'aquesta, per tant és bastant probable que hi haguessin problemes legals en cas de fer-ne un ús comercial venent aquest conjunt de dades, principalment pel fet que un dels objectius seria treure rendibilitat econòmica en detriment de les pròpies cases d'apostes utilitzades per l'extracció de dades.

Amb el terme "**NC – NonCommercial**" s'indica que no se li pot donar un ús comercial al conjunt de dades, amb la qual cosa una tercera persona no podria destinar el *dataset* generat en aquest projecte a aquesta finalitat.

Addicionalment, el terme "**SA – ShareAlike**" determina que, si un tercer vol fer un *dataset* derivat a partir del que s'ha generat en aquest projecte, haurà de compartir-lo amb els mateixos termes, és a dir, sota la mateixa llicència. Així doncs, una alteració del conjunt de dades no dona peu a poder-lo comercialitzar.

9. Codi

L'arxiu robots.txt de la pàgina de Marathonbet ens mostra que el user-agent que utilitzarem no té cap llista específica de bloqueigs, i a la llista per tots els user-agents no apareixen les pàgines del llistat d'esdeveniments i apostes d'esdeveniments, per tant no es realitza cap acció que no es permeti segons indica la web.

```
User-agent: *
Disallow: */toto/files/ # toto
Disallow: *ticketinfo.htm # toto
Disallow: *periodGroupAllEvents # time filter
Disallow: *wintickets.htm # toto
Disallow: *allbets.htm # toto
Disallow: *batchhome.htm # packet bets toto
Disallow: *drawinfo.htm # toto
Disallow: *pooldistr.htm # toto
Disallow: *draws.htm? # toto
Disallow: *betinfo.htm # toto
Disallow: *myaccount # restricted
Disallow: */toto/*/home.htm?
Disallow: *search*
Disallow: *username*
Disallow: *forgottenpassword.htm
Disallow: *?r*
Disallow: *loginpage.htm
Disallow: *login.htm
Disallow: *logout.htm
Disallow: *payment
Disallow: *finstakes.htm?r
Disallow: *withdraw.htm
Disallow: *withdraw2.htm
Disallow: *viewtopic.php
Disallow: *live/*?
Disallow: *printbet.htm
Disallow: *sportstatext.htm
Disallow: *viewprofile
Disallow: *&u=*
Disallow: *?u=*
Disallow: *folder=
Disallow: *[object%20Object]
Disallow: *view=
Disallow: *start=
Disallow: *deposit.htm
Disallow: *deposit2.htm
Disallow: */extensions.htm?
Disallow: */client.htm?
Disallow: *forgottenpassword.htm
Disallow: *updateseance.htm
Disallow: *changepassword.htm
Disallow: *jsessionid
Disallow: */live/stream/
Disallow: */live/animation/
```

El codi per fer web scraping s'ha desenvolupat amb l'assistència de la llibreria **Selenium**, que permet automatitzar navegadors i reproduir el comportament d'una persona real interactuant directament amb una pàgina web a través del navegador.

El **user-agent** que està utilitzant el codi és el que té parametritzat per defecte el WebDriver de **Selenium**, mostrat al terminal mitjançant el mètode *print_user_agent*.

```
user-agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/107.0.0.0 Safari/537.36
```

Figura 6. User-Agent utilitzat pel codi

En primer lloc, es carrega la pàgina principal de l'esport escollit, tennis en el cas del *dataset* resultant, que és on es troben el llistat amb tots els esdeveniments propers.

Aquesta primera pàgina té la peculiaritat que no carrega tots els elements inicialment, sinó a mesura que es va fent l'scroll descendent. Degut a això, s'ha generat el mètode *scroll_to_load_all_data,* la funció del qual és utilitzar les funcions del *WebDriver* per anar a buscar l'element *footer* de la pàgina, que està sempre situat a la part inferior de la secció que conté el llistat, i fer l'scroll fins a la seva posició. S'ha utilitzat aquesta alternativa perquè l'opció més directa de descendir utilitzant les funcions bàsiques de scroll de **Selenium** no eren funcionals per aquesta pàgina, ja que el llistat es troba dins d'una secció amb un scroll propi que no és un objecte, sinó que està desenvolupat mitjançant codi javascript.

Aquesta acció es repeteix fins que s'arriba al final de la pàgina, que es comprova analitzant si hi ha hagut canvi de posició vertical després de fer l'scroll, o fins que s'arriba al timeout indicat per evitar un scroll infinit. Entre cada acció d'scroll es va deixant un marge de temps, principalment perquè doni temps a carregar les noves dades però també per tal d'espaiar les peticions HTTP que es faran al servidor al canviar de pàgina web, per evitar que l'script sigui detectat com a procés automatitzat, que podria comportar un bloqueig de l'accés a la web.

Un cop carregats tots els esdeveniments, s'obtenen tots els enllaços individuals als esdeveniments a partir del mètode *get_all_match_links*, que utilitza les funcions del driver per trobar els elements que contenen els enllaços i guardar-los en una llista. D'aquesta forma, l'script pot realitzar la navegabilitat per buscar les apostes dels esdeveniments d'aquest esport de forma autònoma.

A continuació, s'itera sobre la llista de links obtinguda per anar a cada un dels enllaços dels esdeveniments individualment. Al mètode que realitza el canvi de pàgina (*go_to_link*) se li ha afegit un temps d'espera per assegurar l'espaiat de peticions HTTP.

Un cop entrat al link corresponent el primer que es fa és buscar si la web hi ha un botó on hi ha el text *Todos los mercados.* El driver prem aquest botó i a continuació procedeix a obtenir el nom dels dos participants del partit per tal de reemplaçar el nom d'aquest per local o visitor als noms dels mercats, cosa que serveix per estandarditzar els noms dels mercats per tots els partits.

Fet això es crida la funció *get_betting_dataset*, a dins d'aquesta el primer que fa el codi és obtenir la aposta i la quota de el guanyador del partit, i a continuació agafa el element web *market-inline-block-table-wrapper* que conté tots els mercats restants. Aquí per a cada un dels mercats obté el nom i si aquest conté el nom d'algun dels participants els canvia per *local* o *visitor*. Per tal d'estalviar temps i tenir les dades més netes tenim una

taula amb els mercats que interessen analitzar, i aquí és quan es fa la comprovació de si el nom del mercat actual està dins d'aquesta. Si és així procedeix a extreure totes les dades de les apostes amb les seves quotes associades buscant tots els elements amb una classe anomenada *selection-link* i agafant els atributs *data-selection-key* per les apostes i *data-selection-price* per les quotes.

Finalment es neteja una mica la informació obtinguda traient l'inici del string fins al caràcter @ obtingut com a aposta i en el cas que sigui un handicap substituint la lletra final *H* o *A* com a *local* o *visitor*.

Per generar i exportar el *dataset* s'ha utilitzat la llibreria **pandas**, que permet generar taules mitjançant els objectes *dataframe*. S'ha generat el mètode *generate_event_dataframe*, que permet introduir com a paràmetres les columnes del conjunt de dades. Aquests poden ser o bé valors constants, o bé llistes amb el mateix nombre d'elements. En el cas d'indicar un valor constant en un dels camps, el mètode de **pandas** que genera el *dataframe* a partir d'un diccionari de les llistes i valors constants que li hem indicat (utiltizant el nom de la columna com a clau) posarà aquest valor a totes les files. Es generaran tantes files com elements tinguin les llistes.

Aquest mètode es crida quan ja s'han llegit les dades d'una pàgina (esdeveniment esportiu) i el *dataframe* resultant s'annexa al *dataframe* principal que contindrà totes les dades. Posteriorment, s'utilitza el mètode *dataframe.to_csv* (paràmetre index=False per tal de no exportar l'identificador numèric que **pandas** li afegeix automàticament a les files) sobre aquest *dataframe* principal que serveix per exportar-lo a un fitxer en format CSV.

10. Dataset

L'enllaç del DOI del dataset publicat a Zenodo és el següent:

https://doi.org/10.5281/zenodo.7348647

11. Vídeo

L'enllaç del vídeo demostració és el següent:

https://drive.google.com/file/d/1LQIoZRFcQlFh_dlOgQaakLWm5Tac1X8-/view?usp=share_link

12. Taula de contribucions

Contribucions	Signatura
Investigació prèvia	Adrià Setó Balcells, Miquel Arisa Fuente
Redacció de les respostes	Adrià Setó Balcells, Miquel Arisa Fuente
Desenvolupament del codi	Adrià Setó Balcells, Miquel Arisa Fuente
Participació al vídeo	Adrià Setó Balcells, Miquel Arisa Fuente