Com es pot reformular l'estratègia de comerç electrònic d'una organització mitjançant l'anàlisi de dades?

Arturo Palomino Gayete

PID_00242821

Universitat Oberta de Catalunya

H2PAC

Director de la col·lecció: Lluís Pastor

L'encàrrec i la creació d'aquest material han estat coordinats
pel professor: Josep Cobarsí (2017)

Primera edició: setembre 2017

Av. Tibidabo, 39-43, 08035 Barcelona

Realització editorial: Oberta UOC Publishing, SL

Dipòsit legal: B-21.146-2017

Índex

4 **Com utilitzar un model H2PAC**

7 **El repte**

15 **El coneixement imprescindible**
17 1. Canal de distribució, el sector del comerç al detall en l'àmbit nacional
31 2. Anàlisi estratègica d'informació en el comerç al detall

45 **Les solucions**
89 Bibliografia

Com utilitzar un model H2PAC

Aquest model planteja resoldre propostes clau a partir d'ACTIVITATS concretes. A continuació us expliquem com treure'n profit mitjançant tres fases.

1 El repte

A les primeres pàgines apareix el repte que se us planteja en aquest material.

2 El coneixement imprescindible

A les pàgines centrals trobareu la **teoria imprescindible** que us ajudarà a entendre els conceptes clau i poder obtenir les respostes al repte.

3 Les solucions

A les pàgines finals trobareu el **solucionari** perquè pugueu resoldre correctament el repte.

El repte

Amb aquest material es pretén posar l'estudiant en la pell d'un científic de dades que vol potenciar les accions i optimitzar les decisions d'una empresa de comerç electrònic, semblant a Amazon, amb l'objectiu d'incrementar la rendibilitat de la inversió (ROI) i protegir-se d'informació desfavorable, emprant diferents tècniques estadístiques i de mineria de dades.

En l'apartat «Coneixement» definirem els conceptes bàsics del sector del comerç electrònic i l'àrea en què s'engloba, el sector del comerç al detall (*retail*): supermercats, hipermercats, establiments de descompte, especialistes, etc. Veurem els agents que són presents en el sector tradicional i les estratègies que han portat a l'èxit a alguns dels principals representants del sector –els *major players*–, com ara Mercadona. Continuarem amb la importància de triar el producte i el pla estratègic d'empresa. A continuació, analitzarem més detalladament el present i el futur del comerç electrònic, amb les noves tendències que estan revolucionant aquest sector tan incipient. Seguidament, definirem els conceptes del científic de dades, els **sistemes d'informació**, la **intel·ligència de negoci**, el ***Big Data***. En darrer lloc, per a completar l'àrea de coneixement, ampliarem la informació amb un petit recopilatori dels mètodes més utilitzats que estan a l'abast de l'investigador en el camp de l'**estadística** i l'**aprenentatge automàtic** per a extreure coneixement de les dades. Aquestes tècniques les utilitzen Amazon i altres empreses en línia: les **empreses «.com»**.

Al llarg del material, farem un resum del contingut teòric i, en l'apartat «Les solucions» exposarem dos casos pràctics, el primer dels quals es basa en un sistema de **recomanació de productes** amb el programari lliure R, amb una font de dades de productes simulada equiparable a un fitxer *log* de transaccions de compres reals en línia. Per mitjà d'aquest cas, l'alumne experimentarà amb els algorismes que utilitzen les empreses de comerç electrònic per a recomanar productes als usuaris. Aquests processos analitzen els hàbits i els patrons de compra dels usuaris en un rang temporal ampli, amb la finalitat d'extreure coneixement i predir accions del consumidor. Definirem l'algorisme *k-nearest neighbor* i l'algorisme *Apriori* i detallarem els pas-

sos que cal seguir per a complir l'objectiu de posar en marxa un sistema de recomanació de productes en temps real. Finalment, farem un exercici de **mineria d'opinió** (*sentiment analysis*), mitjançant el qual analitzarem l'opinió dels usuaris respecte a la nostra marca i els nostres productes, perquè l'alumne vegi la importància i el potencial de monitorar tot el que és a la ment del consumidor pel que fa al posicionament de la nostra oferta i la nostra imatge corporativa. L'objectiu d'aquest tipus d'anàlisi està lligat a la comprensió de les decisions dels usuaris amb la finalitat de dissenyar i millorar un **pla de màrqueting** i generar mecanismes i plans de contingència davant de situacions de crisis, per una falta de credibilitat o una opinió desfavorable.

Per a comprendre la importància d'aquest tipus d'anàlisi hem d'entendre el potencial en termes de volum i valor en vendes que es preveu que generaran en un futur proper les empreses de distribució i venda de productes per internet. Com veurem més endavant, el sector de gran consum engloba grans categories de béns que el consumidor compra habitualment en botigues i supermercats. Podem trobar-ne la classificació al Reial decret 367/2005, de 8 d'abril, pel qual es desenvolupa l'article 17.3 de la Llei 7/1996, de 15 de gener, d'ordenació del comerç al detall, i es defineixen els productes d'alimentació frescos i peribles i els productes de gran consum.[1] Aquest mercat mou anualment 72.000 milions d'euros a Espanya i aglutina productes d'alimentació, begudes, perfumeria i productes per a mascotes. El canal d'internet genera més de 500 milions d'euros, té un creixement entre el 2015 i el 2016 del 26% i representa un 1% de la despesa total realitzada en comerç al detall. El sector del comerç electrònic encara és molt petit, però els experts assenyalen que, amb aquest creixement, és evident el seu enorme potencial per a distribuïdors i fabricadors.

El sector del comerç al detall fa anys que té plataformes de venda en línia, però els distribuïdors que solament venen en línia tenen un

1 <https://www.boe.es/diario_boe/txt.php?id=BOE-A-2005-6795>

paper clau per a aquest creixement a Espanya. A escala internacional (als Estats Units, per exemple), les vendes del canal en línia acaparen una quota de mercat superior al 15%. A Europa els països que lideren el rànquing són el Regne Unit i França, tots dos amb quotes per sobre del 5%.

Les plataformes de venda en línia exclusiva són un sector amb molta evolució i millores contínues. Les empreses diversifiquen els seus productes amb la finalitat d'atreure nous compradors i reduir el risc. Recentment, Amazon ha anunciat el llançament del servei Prime Now. Amb aquest servei l'empresa passa a engrossir la seva oferta, perquè ofereix productes envasats de gran consum que subministra mitjançant enviament a domicili en una o dues hores, després de fer una comanda al seu web, com si es tractés d'una cistella de la compra. Paral·lelament proliferen altres webs com Ulabox, Tudespensa.com, Deliberry i Comprea, que generen friccions en el sector i empenyen els grups tradicionals (Carrefour, El Corte Inglés, Eroski, Dia, Lidl) a renovar i rellançar les seves plataformes, i també a millorar els seus processos operatius i logístics.

La globalització i sobretot internet han potenciat en les últimes dècades, gràcies a la difusió de les tecnologies de la informació, la proliferació d'aquests webs i l'acceptació per part del consumidor de l'ús d'aquest canal de compra, malgrat el recel inicial. Recentment, i com es pot observar en les accions d'aquests grans grups, el canal està en ple creixement i és evident l'auge que experimenta el comerç electrònic. Aquest canvi en els hàbits del consumidor no ha estat sobtat, sinó tot el contrari: el creixement ha estat gradual, lent i molt dependent de les millores en el sector tecnològic i, més concretament, de l'evolució que ha experimentat internet en els últims anys.

Internet facilita un altre tipus de relacions, no solament comercials sinó també de fidelització, i els sistemes d'informació, basats en la gestió de servei al client (*Consumer Relationship Management*, CRM), reben també un nou impuls amb la creació de noves eines adaptades a internet i l'anàlisi d'informació provinent dels transaccions que es

duen a terme en els servidors que allotgen els webs. En aquest camp, apareixen sistemes d'informació com Google Analytics, Tova Omniture o IBM Digital Analytics (Coremetrics), que ajuden a entendre les preferències, les decisions i les motivacions dels internautes en la seva navegació. En aquest sentit, és especialment interessant entendre el procés de compra dels visitants en les plataformes de comerç electrònic.

Un sector tan dinàmic com el comerç electrònic ha experimentat canvis estructurals en l'àmbit empresarial que gairebé no s'havien vist en el comerç al detall tradicional. Amazon, per exemple, s'ha anat transformant amb els anys, ja que va començar com un canal de venda de llibres en línia (llibres electrònics), passant per la venda de CD de música i de productes informàtics, i recentment venda de productes envasats de gran consum o fins i tot proveïdor de música reproduïda en temps real (*streaming*). L'abast del negoci d'Amazon va ser de 100.000 milions de dòlars el 2015, pràcticament el doble de la facturació de gran consum a Espanya. Aquesta facturació solament és comparable amb la d'empreses com Facebook, Google o Apple, i totes formen el tàndem que rep l'acrònim de **GAFA (Google, Amazon, Facebook i Apple)**.

Amb el pas dels anys, la tecnologia ha estat un dels puntals d'èxit de l'empresa de Jeff Bezos (creador d'Amazon), que ha innovat amb diferents tècniques de mineria de dades, aprenentatge automatitzat i algorismes d'optimització, ha creat sistemes de recomanació per a suggerir productes i ha utilitzat mineria de dades i reconeixement de patrons, modelització i predicció per a anticipar compres i optimitzar l'enviament de productes als seus magatzems locals –*anticipatory shipping*– amb la finalitat de minimitzar-ne els temps d'espera. També com la resta d'empreses del sector tecnològic, analitza les opinions i monitora les xarxes socials i els fòrums per a detectar tendències en la venda de productes i l'eficiència de la seva oferta actual.

Al llarg d'aquest material farem una introducció al món del comerç electrònic, fent èmfasi en la incursió recent d'Amazon en el mercat

espanyol de gran consum, i també en la creació a Madrid del primer centre d'operacions europeu (*hub*) del gegant americà. Veurem quines tècniques poden emprar aquest tipus d'empreses per a fer-se un lloc important en la distribució de productes de gran consum a Espanya i analitzarem algunes d'aquestes metodologies.

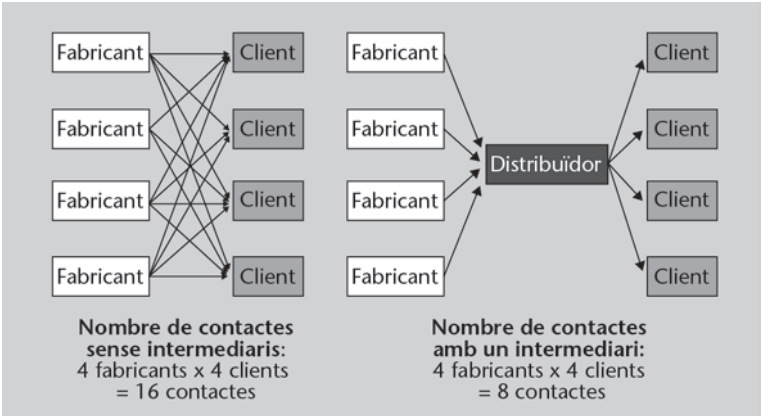
El coneixement imprescindible

1. Canal de distribució, el sector del comerç al detall en l'àmbit nacional

En aquest apartat veurem la utilitat de la figura del distribuïdor en el procés de venda de productes. Ja hem definit el sector de gran consum i els productes que el componen. Analitzarem els canals de distribució i en concret el sector del comerç al detall, que és el que s'encarrega de vendre els productes de gran consum.

Els canals de distribució sorgeixen com a conseqüència de la necessitat de separar els processos merament productius dels processos d'intercanvi comercial entre client i fabricant. La importància d'aquesta separació consisteix en la reducció del nombre d'intermediaris, de manera que diversos fabricants poden vendre els seus productes a un únic distribuïdor, el qual els posa a disposició del consumidor final a canvi d'un marge de benefici en la transacció.

Figura 1. Els intermediaris minimitzen els contactes



Font: Inma Rodríguez Ardura, Guillermo Maraver Tarifa i Francisco J. Martínez López (2005). *Canales de distribución* (pàg. 15). Barcelona: UOC.

Per al productor, el flux comercial es redueix a un procés entre dos únics agents; a més, l'especialització en les funcions productives li permet optimitzar els seus productes i, en molts casos, evita incórrer en la despesa addicional, que implicaria la necessitat de disposar d'establiments propis per a la venda directa. Com es pot observar en la figura 1, la figura del distribuïdor evita la necessitat que el fabricant estableixi contactes amb els clients finals.

Per **longitud del canal** entenem el nombre d'intermediaris existents entre el fabricant i el consumidor.

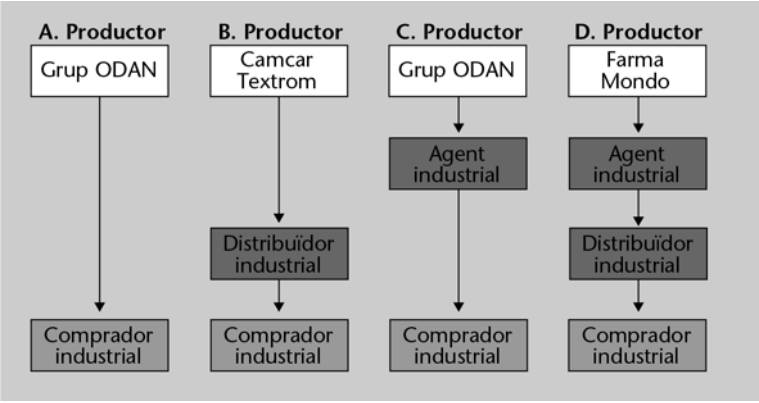
Segons la longitud del canal, els canals de distribució es poden subdividir en:

- **Canals de distribució directes**: el flux és directe entre fabricant i consumidor.
- **Canals de distribució indirectes curts**: el flux és indirecte entre el fabricant i el consumidor; apareix la figura del detallista i en alguns casos s'hi afegeix la del majorista.
- **Canals de distribució indirectes llargs**: el flux és indirecte entre fabricant i consumidor, i a més del detallista i el majorista apareix la figura de l'agent comercial.

En la figura 2 veiem les diferents estratègies d'estructura de distribució per les quals pot optar el fabricant.

De vegades, un fabricant opta inicialment per una estratègia indirecta llarga, i amb el temps decideix integrar, en les seves funcions, les dels agents intermediaris amb qui interactua, en el que coneixem com a **integració vertical**. Això sol produir-se quan un mercat passa a ser madur i l'empresa ha superat l'etapa d'internacionalització o d'implantació amb èxit. Arribat aquest moment, l'empresa coneix millor els diferents factors regionals i assumeix les funcions de l'agent comercial. En alguns casos, pot arribar a aconseguir integrar les funcions del distribuïdor industrial modificant d'aquesta manera els acords del canal; i en d'altres, senzillament s'eliminen les funcions.

Figura 2. Tipus de canals de distribució



Font: Inma Rodríguez Ardura, Guillermo Maraver Tarifa i Francisco J. Martínez López (2005). *Canales de distribución* (pàg. 38). Barcelona: UOC.

D'altra banda, també és habitual el pas contrari, que consisteix a afegir un canal de distribució. Això sol passar quan alguna innovació tecnològica afavoreix la **creació d'un nou canal**, que és acceptat massivament pel consumidor. En seria un exemple la venda per telèfon de productes com ara: serveis de telefonia, assegurances, cursos.

L'ús d'una nova tecnologia per part del consumidor és contemplat pel distribuïdor com una oportunitat per generar nous ingressos. Si és capaç d'adaptar-se als nous hàbits de consum abans que els seus competidors, aconseguirà nous compradors i evitarà la pèrdua de clients. El primer distribuïdor que incorpora el nou canal i genera els processos que faciliten la transacció amb èxit genera un avantatge competitiu, que pot arribar a representar un canvi disruptiu en l'entorn competitiu.

Un altre exemple clar de l'aparició d'un nou canal adoptat massivament per les empreses és internet. Gràcies a internet apareix el **comerç electrònic**, que permet a les empreses crear aparadors virtuals dels seus productes, publicar fotos al seu web, donar tota la informació possible mitjançant correu electrònic o descàrrega de documents PDF i fer la transacció comercial amb terminals de

punt de venda en línia que permeten realitzar pagaments amb targeta o contra reemborsament. És a dir, el fabricant pot fer fàcilment les funcions de venda directa a escala internacional sense tenir grans costos. També el distribuïdor pot estalviar costos de distribució fent-se més competitiu en preus i, per tant, atraient més clients que els canals tradicionals. Sorgeixen llavors empreses especialitzades en la venda per internet com Amazon, eBay, Jet i moltes d'altres.

En el procés de decisió de la tria del millor sistema de distribució per a un fabricant, és important plantejar un **enfocament analític per etapes**, en què, en primer lloc, s'ha de valorar la documentació existent sobre els possibles canals de distribució. Paral·lelament, cal analitzar i comprendre el sistema actual. En aquesta primera etapa, és important organitzar tallers i entrevistes sobre el canal de distribució existent, amb la finalitat de valorar les opcions que millor s'hi adaptin. Igualment, s'han d'analitzar els canals de distribució de la competència. Mentre s'està pendent de noves oportunitats a curt termini, cal desenvolupar un pla d'atac a curt termini, sempre que sigui necessari. De la mateixa manera que és important conèixer les limitacions i les possibilitats pròpies, no s'ha de perdre la perspectiva del que demana el client. En aquest sentit és important dur a terme una **anàlisi qualitativa** de les necessitats de l'usuari final per mitjà de reunions de grup i entrevistes personals, o bé fer anàlisis quantitatives mitjançant enquestes i altres tècniques. Més enllà del perímetre de negoci, també cal analitzar i comprendre les solucions que ofereixen empreses d'altres indústries properes. Després d'analitzar aquesta informació, els directius hauran de desenvolupar un **canal de màrqueting ideal**. Es compararà el canal ideal amb el resultat de l'examen de la solució actual i els canals de la competència amb la finalitat d'analitzar les divergències i sobre la base del resultat identificar i desenvolupar les opcions estratègiques que ens permetin dissenyar el canal òptim.

Si una empresa té pensat ocupar un lloc en un sector tan madur ha de tenir molt ben definida la seva estratègia, ja que, sens dubte, les altres empreses ja establertes no estaran disposades a posar-li-ho fàcil.

1.1. Agents tradicionals del sector del comerç al detall i estratègies d'èxit

En aquest apartat veurem l'esquema actual de la distribució a Espanya i les accions que han portat a l'èxit als *major players*. El sector del comerç al detall està en evolució contínua. El rànquing de vendes el controlen un nombre reduït d'empreses que es reparteixen el mercat de la distribució.

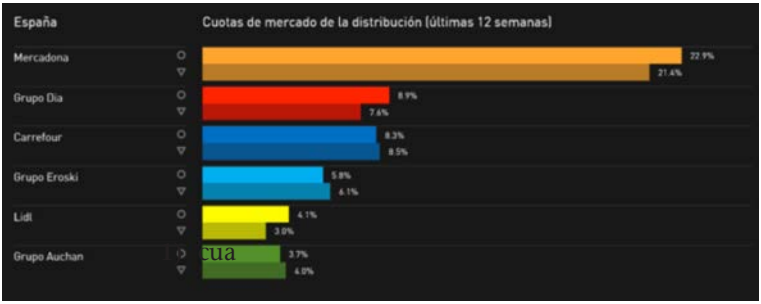
En general, en els àmbits d'alimentació, begudes, drogueria i perfumeria, les empreses solen classificar-se segons diferents criteris relatius a l'estratègia de preus, la grandària i la ubicació, que és el que entenem per classificació del canal de distribució. En el llenguatge d'estudis de mercat la classificació més habitual és la següent:

- **Híper**: en aquest grup hi trobem grans distribuïdors com Eroski, Carrefour, Alcampo o Caprabo. Es caracteritzen per tenir una presència limitada i estan situats habitualment en àrees de la perifèria o zones molt concorregudes de la ciutat, amb un gran espai de venda, un bon assortiment i diversitat de productes.
- **Súper**: hi trobem distribuïdors que habitualment tenen més presència als barris, un espai reduït, nivell d'assortiment mitjà i diversitat suficient de productes per a fer compres de proveïment, d'estoc i de primera necessitat.
- **Establiments de descompte (*discount*)**: és el grup de distribució amb una presència elevada als barris de ciutat, espais reduïts i nivell d'assortiment molt enfocat a productes bàsics, de primera necessitat i de preus molt baixos.
- **Especialistes**: engloben botigues de barri especialitzades en tipus de productes molt concrets, entre les quals hi ha xarcuteries, fruiteries, verduleries.
- **En línia**: són botigues tradicionals que ofereixen els seus productes a través d'aquest mitjà i empreses «.com», dedicades a la venda exclusiva per aquest canal.
- **Altres**: farmàcies, parafarmàcies, etc.

A Espanya el rànquing de vendes per distribuïdor del sector del comerç al detall el lidera Mercadona, empresa dirigida per Joan Roig,

fundada en 1977,[1] seguida pel grup Dia, Carrefour, Eroski, Lidl i el grup Auchan. En la figura 3, veiem un *benchmark* del sector a Espanya del primer semestre del 2016.

Figura 3. *Benchmark* del mercat de distribució, primer semestre del 2016



Font: aplicació de difusió oberta de Kantar Worldpanel a http://www.kantarworldpanel.com/es/grocery-market-share/spain

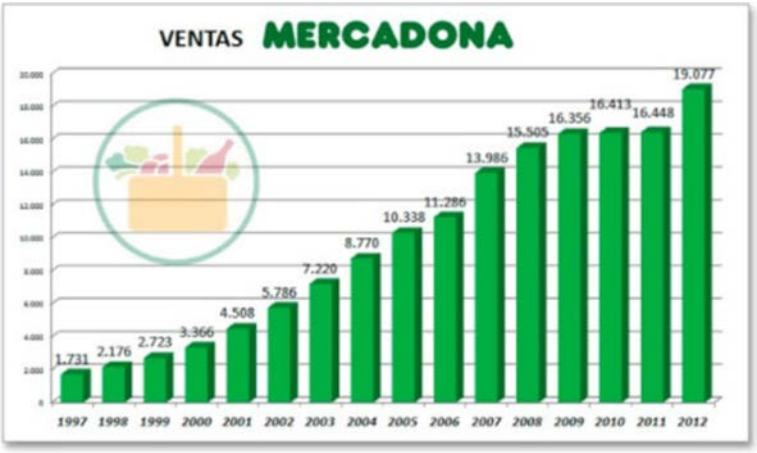
Aquest escenari no sempre havia estat així. En el passat n'hi havia un de molt diferent, en què Mercadona se situava a la cua del rànquing.[2] Tanmateix, un gran encert en les diferents estratègies va consolidar l'empresa de Joan Roig com el nou líder del sector a Espanya. En la figura 4 s'observa aquesta evolució.

Abans de l'arribada del distribuïdor valencià, dominaven el sector empreses com El Corte Inglés, Carrefour, Dia i Eroski. Les estratègies seguides per Joan Roig, que han influït en gran mesura en aquest lideratge, es poden sintetitzar en la figura 5.

1 <https://es.wikipedia.org/wiki/Juan_Roig_Alfonso>

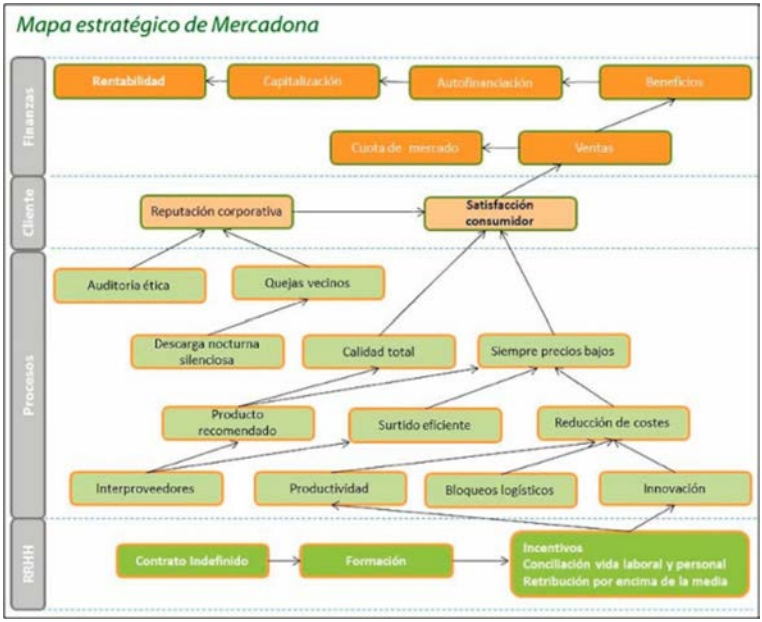
2 <http://mundodelaempresa.blogspot.com.es/2013/12/empresas-el-modelo-de-exito-de-mercadona.html>

Figura 4. Evolució de vendes de Mercadona fins al 2012



Font: «El mundodelaempresa» a http://mundodelaempresa.blogspot.com.es/2013/12/empresas-el-modelo-de-exito-de-mercadona.html

Figura 5. Mapa estratègic de Mercadona



Font: «El mundodelaempresa» a http://mundodelaempresa.blogspot.com.es/2013/12/empresas-el-modelo-de-exito-de-mercadona.html

Com veiem, a la base se situa el personal, que actua com a motor de la productivitat i la innovació. La innovació és, per tant, un dels punts diferencials en comparació amb esquemes més tradicionals. La innovació repercuteix en la reducció de costos. Paral·lelament, la gestió eficient de l'assortiment, provinent dels interproveïdors (proveïdors de marca blanca), acciona la palanca, que és el principal valor estratègic del pla d'acció de Mercadona, juntament amb l'eslògan «Sempre preus baixos», puntal fonamental de l'èxit de la cadena.

Per tant, podem dir que, com és lògic, la productivitat és un factor important per a l'èxit, però si una organització vol destacar ha de tenir molt en compte l'**assortiment**, la **innovació** i els **preus**.

Aquests són els punts clau diferenciadors que es troben a la base dels processos de l'estratègia seguida per l'executiva de Mercadona.

Figura 6. «Sempre preus baixos», eslògan de la cadena valenciana

Font: «El mundodelaempresa» a http://mundodelaempresa.blogspot.com.es/2013/12/empresas-el-modelo-de-exito-de-mercadona.html

En el factor de l'assortiment de productes, la relació de Mercadona amb els seus interproveïdors és un cas d'integració vertical, en què la cadena recorre als fabricants de productes de gran qualitat, que s'integren en el paraigua de marques blanques de la cadena (Hacendado, Deliplus, Compys, Bosque Verde, Como Tú, 9.60, Dermik i Solcare), per a atendre les categories d'alimentació, cosmètica, alimentació de mascotes, drogueria, perfumeria, dermoestètica i protecció solar.

Un nou competidor que vulgui entrar en el mercat de la distribució a Espanya s'haurà de fixar en les estratègies que han funcionat millor a escala regional per a intentar replicar-les o millorar-les.

1.2. Els plans d'empresa i la importància de triar el producte

A continuació veurem com es classifiquen els productes i la importància de conèixer les categories en les quals volem competir per tal de delimitar i identificar millor els nostres rivals. En el procés de gestació d'una empresa, la decisió d'elecció de combinació de factors productius no és una tasca senzilla. Sovint, el procés que hem definit no és el que obtindrà els millors resultats, bé perquè no és el que busca el consumidor o bé perquè la fórmula triada ja existeix amb una combinació de costos i qualitat més competitius que la nostra. En aquest sentit, és important definir un **pla d'empresa**, que habitualment es plasma en un document on es recull la idea de projecte que es pretén engegar i es concreta la idea de negoci amb el detall més gran possible.

Aquesta eina no és un document estàndard, sinó que sol seguir un esquema amb uns conceptes comuns a la majoria de projectes de negoci. Se solen tenir en compte tres grans conceptes: la **idea de negoci**, el **producte** i la **comercialització**.

En general, és un document dinàmic que sol variar en el temps a mesura que es millora el plantejament. Cal destacar que tenir un pla d'empresa ben definit sol servir per a obtenir finançament, socis o col·laboradors.

Qualsevol dels punts del document són essencials –finançament, selecció de personal, forma jurídica, etc.–, però, sens dubte, el producte mereix ser analitzat detalladament.

> Els productes són essencials en el flux circular de la renda del mercat de béns i serveis.

Les empreses generen ingressos oferint un producte atractiu als consumidors; els consumidors, al seu torn, decideixen dedicar una part de la seva renda a la compra de béns i serveis.

La primera gran divisió d'activitat de l'empresa es basa en el tipus d'oferta, que pot ser de dos tipus. El primer grup és el de les empreses de béns i el segon és el de les empreses que ofereixen serveis, és a dir, el tipus d'oferta segons la seva tangència. Dins de l'**oferta de béns** podem classificar els productes, segons la durabilitat, en **béns peribles** i **béns imperibles**. En el primer grup tindríem els productes de consum relativament ràpid, com el cafè, el sabó, les fruites; i en el segon grup, els productes que poden usar-se diverses vegades, com els electrodomèstics, els cotxes o els ordinadors.

Segons els principals instituts d'estudis de mercat, la classificació de béns peribles inclou productes d'alimentació i begudes, drogueria, perfumeria familiar, alimentació per a mascotes i productes per al bebè, és a dir, els productes que habitualment trobem en els distribuïdors de gran consum.[3]

Una empresa de comerç electrònic que vengui productes d'informàtica o electrodomèstics òbviament no necessitarà fixar-se en Mercadona, Eroski o Lidl, ja que el seu entorn competitiu és molt diferent.

---

3 <http://www.promonegocios.net/producto/tipos-productos.html, http://www.kantarworldpanel.com/es/Noticias/El-Gran-Consumo-se-estabiliza-en-el-segundo-trimestre-cuotas-distribucion-junio>

En aquest sentit serà molt important tenir molt clara la categoria de productes que volem oferir.

1.3. El comerç electrònic i les noves oportunitats

En aquest apartat veurem el sector del comerç electrònic amb més detall, i posarem l'accent en les noves oportunitats que van acompanyades de noves millores tecnològiques i en els hàbits dels consumidors. En els últims anys hi ha hagut una evolució tecnològica evident, que ha estat motor d'impuls de creixement del sector, que ha derivat en un nou esquema de distribució que coneixem com a **comerç electrònic (*e-commerce*)**.[4] La seva característica principal és el mitjà a través del qual s'estableix el contacte amb el client, l'oferta, la comanda i el mitjà de pagament: internet en qualsevol de les seves formes, mitjançant ordinador, tauleta o telèfon intel·ligent.

Entre les empreses que utilitzen aquest canal, hi ha distribuïdors tradicionals que s'adapten a les noves tecnologies i creen el seu propi web de comerç electrònic (Carrefour en línia, Eroski en línia, DIA en línia…), però, d'altra banda, també hi ha empreses «.com» de distribució en línia, originades a internet, per a la venda de productes a baix preu, que readapten els seus assortiments i comencen a oferir productes de gran consum mitjançant estratègies de diversificació.

> N'és un clar exemple Amazon, que es va gestar originalment com una empresa de venda de llibres en línia i que en els últims temps està oferint tot tipus de productes, des d'electrodomèstics i productes electrònics, serveis d'informàtica en núvol i, més recentment, fins i tot béns peribles.

En l'altre extrem, hi trobem grans superfícies, com ara el gegant nord-americà Walmart, que adopten el nou canal amb la compra de webs de comerç electrònic i practiquen l'estratègia d'integració vertical, com la recent adquisició de Jet.com per a fer front a Amazon.

---

4 Worldpanel, K. (2014). *Accelerating the Growth of E-commerce in FMCG*. Kantar Worldpanel.

Amb la finalitat de conèixer les amenaces i les oportunitats, els agents han d'analitzar l'entorn per a planificar l'entrada al mercat, marcar-se uns objectius de creixement o identificar riscos i amenaces. Un primer enfocament implica fer un esquema del **grau d'estabilitat de l'entorn** per a valorar la conveniència d'adoptar les diferents estratègies o fins i tot considerar la possibilitat d'abandonar el projecte.

Les oportunitats per a aquest tipus d'empreses són evidents per les xifres que hem esmentat en l'apartat 1. Amb aquestes xifres de negoci és obvi que val la pena fer-se un lloc i entrar a competir. La gran fortalesa que té aquest tipus d'empresa és el fet de ser una marca ja establerta i coneguda amb un aparell logístic optimitzat i amb l'enorme poder de negociació d'una companyia que domina el comerç en línia a escala internacional. Sens dubte, aquests dos elements del DAFO superen amb escreix els elements en contra de feblesa i amenaces, ja que entren a competir en un tipus de canal molt incipient en el qual la pitjor amenaça són ells mateixos i on no tenen les febleses d'altres negocis menys diversificats i regionals com els ja establerts.

Com hem vist, una empresa com Amazon té grans arguments a favor i a més disposa també de la inèrcia de les noves tecnologies, de les quals és el màxim exponent, un camp en el qual els altres competidors ho tenen realment complicat per a aconseguir el seu nivell.

1.4. Noves tendències i evolució del sector

Hem parlat d'exemples d'empreses que adopten estratègies de diversificació de productes, estratègies d'integració horitzontal o vertical. D'altra banda, també és habitual el pas transversal, que consisteix a afegir un nou canal de distribució. Això sol passar quan alguna innovació tecnològica afavoreix la creació d'un nou canal, amb acceptació massiva per part del consumidor. En seria un exemple la venda per telèfon de productes com ara serveis de telefonia, assegurances, cursos. Un altre exemple més recent és la venda per internet, que rep

el nom de comerç electrònic. Una variant d'aquest últim canal és el de **comerç electrònic mòbil** o ***m-commerce***.[5]

El comerç electrònic mòbil té especial interès en els darrers anys. Actualment els usuaris tendeixen, cada vegada més, a fer funcions amb el telèfon intel·ligent que abans solien fer amb l'ordinador personal. Els telèfons intel·ligents i les tauletes incorporen cada cop més funcions. Al principi, la seva gestació va ser la fusió de dues tecnologies amb funcions separades, els organitzadors personals –Personal Data Assistant (PDA) o PC de butxaca– i els telèfons mòbils. Els primers permetien gestionar dades, com ara agendes, contactes, petites bases de dades i fins i tot dibuixar notes a mà alçada amb llapis òptics. A poc a poc els segons van anar incorporant les funcions dels primers fins al punt que no es concebia un telèfon mòbil sense funcions de PDA ni un PDA sense funcions de telefonia mòbil. La telefonia mòbil, al seu torn, va dotar els dispositius de la possibilitat d'enviar i rebre dades per internet, la qual cosa va obrir un ventall enorme de possibilitats: oci, feina i comerç, entre d'altres.

El comerç electrònic mòbil va créixer un 69% el 2015 a Espanya respecte a l'any anterior, un creixement només superat pel Brasil segons el Zanox Mobile Performance Barometer 2015 (primer semestre), que mesura l'evolució d'aquest canal basant-se en l'anàlisi de més de 4.300 anunciants en 11 territoris. A escala global aquest creixement està per sobre d'un 140% i la despesa total en ordinadors i mòbils ha crescut un 9% respecte a l'any anterior. Els telèfons intel·ligents són els dispositius amb més creixement en nombre de transaccions, en comparació amb les tauletes i els ordinadors. Però els ordinadors continuen tenint una quota superior en el global de transaccions, encara que van cedint terreny a les noves tecnologies. Les despeses amb un valor més elevat es fan mitjançant tauleta.

5 <http://www.xataka.com/moviles/htc-una-historia-de-poco-ruido-y-muchas-nueces, http://www.distribucionactualidad.com/el-m-commerce-crece-un-60-con-un-gasto-por-carrito-de-95-euros/, https://www.emarketer.com/public_media/docs/eMarketer_Mobile_Commerce_Roundup_2016.pdf, http://blog.zanox.com/en/zanox/2016/03/22/zanox-mobile-performance-barometer-2016/>

Possiblement aquests aparells donen més confiança al consumidor en aquelles transaccions que comporten una despesa més elevada.

Les noves tecnologies impliquen també un canvi en les estratègies de màrqueting. Les empreses de màrqueting tradicional, acostumades a fer campanyes en els mitjans de comunicació de massa tradicionals, han hagut d'adaptar els seus processos i productes a les noves tecnologies. Per la seva banda, l'entorn digital ha afavorit que l'oferta, les campanyes i les anàlisis del retorn d'inversió s'hagin adaptat i hagin incorporat tècniques més complexes per tal de:

- Millorar l'experiència de compra.
- Analitzar els efectes de les campanyes.
- Optimitzar els processos de subhasta d'espais publicitaris.
- Generar sistemes de recomanació per a fidelitzar clients.
- Optimitzar els temps d'enviament, estoc.

A més, hi ha molts altres processos que anys enrere se sustentaven en el saber fer dels experts en màrqueting i que actualment són articulats i reinventats pels departaments d'intel·ligència de negoci, tecnologies de la informació i ciència de dades. Entre les tècniques utilitzades per aquests departaments tenen un interès especial les de gestió de dades massives, intel·ligència de negoci, aprenentatge automàtic, estadística i analítica web.

2. Anàlisi estratègica d'informació en el comerç al detall

2.1. Sistema d'informació, intel·ligència de negoci i dades massives

En aquest apartat veurem les eines de què disposa un científic de dades per a traduir les dades a coneixement amb el qual dotar d'elements clau la presa de decisions de l'organització. Entre els recursos que afecten una empresa podem trobar factors establerts tradicionalment com els **humans**, els **financers** i els **materials**. Recentment gràcies als avenços tecnològics també trobaríem factors intangibles, com ara el **valor de la marca**, la **recerca** i la **informació**. Aquesta última és especialment important per a la supervivència d'una empresa. Gràcies a la informació l'empresa és capaç d'establir un rumb que li permeti optimitzar les vendes i superar els obstacles que apareguin en l'entorn competitiu.

L'empresa rep informació de l'entorn per mitjà d'agents externs, però també emet cap a l'exterior i genera informació interna mitjançant mètriques derivades de les operacions diàries, tant de producció com dels recursos propis. Així, l'empresa pot recollir informació de fonts de dades obertes o d'instituts d'estudis de mercat si parlem de fonts externes; paral·lelament, pot generar informació comptable, segons el marc legal, i fer-la pública, o bé recollir informació de transaccions, despeses i mesurament de productivitat, per a treballar-la internament, amb la finalitat de detectar punts febles i forts.

La informació es genera a partir de les dades. Les dades es recullen per mitjà de diverses fonts (per exemple, a través de sensors en les màquines productives) o bé mitjançant personal d'administració amb introducció manual en sistemes, entre molts altres. Aquestes

dades són processades i convertides en informació. Algunes es descarten o bé es desen, però no s'utilitzen de forma immediata (*Data Exhaust*). Perquè una dada es consideri 'informació', ha de tenir un valor informatiu. Per exemple, el nombre d'habitants per país o el nombre de vendes d'una marca són informació. En canvi, una sèrie de dades aleatòries no té valor informatiu. D'aquesta forma, dades inintel·ligibles es processen i es converteixen en informació intel·ligible, comprensible i útil. Finalment, d'aquesta informació intel·ligible i ordenada, disponible per a ser analitzada, s'obté coneixement en forma de conclusions, previsions i recomanacions útils per a la presa de decisions.

Així, els **sistemes d'informació** emmagatzemen i permeten extreure dades rellevants per a la presa de decisions.

Per tant, es tracta de repositoris i eines en els quals abans de la gestió s'han dut a terme processaments i preprocessaments de dades mitjançant processos per lots (*batch*) o càrregues (*bulk*), que han dut a terme, al seu torn, un sedàs d'informació rellevant de totes les possibles fonts de dades existents en les organitzacions (*Data Lake*). Aquests sistemes d'informació els podem classificar de la manera següent:

- ***Transactions Processing Systems*, TPS**. Són els sistemes encarregats de gestionar l'execució de processos indivisibles en les organitzacions, coneguts com a *transaccions*.

- ***Management Information Systems*, MIS**. Aquests sistemes permeten la presa de decisions pel que fa a la gestió empresarial.

- ***Decision Support System*, DSS**. Els DSS són sistemes experts que permeten a l'alta direcció obtenir un suport en el procés de presa de decisions.

- ***Executive Information Systems*, EIS**. Són un tipus específic de DSS, però amb la finalitat d'ajudar en la presa de decisions en l'àmbit executiu sènior, que permet obtenir informació sobre

vendes, despeses i evolució de l'empresa i dels seus departaments.

- ***Office Automatization Systems*, OAS**. Són aquells sistemes englobats tant en el programari com en el maquinari que permeten la gestió automatitzada d'informació de l'empresa en l'àmbit administratiu, en tasques de comunicació, digitalització, emmagatzematge i processament de tasques rutinàries i estandarditzades.

- ***Expert Systems*, ES**. És un programari o la combinació de programari més maquinari capaç d'emular un comportament humà que permet executar tasques amb un cert grau de complexitat de manera similar a com ho faria un ésser humà.

- ***Entreprise Resource Planning*, ERP**. Són sistemes capaços d'aglutinar en un únic sistema d'informació tots els processos de funcionament d'una empresa en un únic component, recopilant informació en tots els àmbits en l'organització: logística, compres, comptabilitat, finances, estoc, comandes, recursos humans, producció, projectes i molts d'altres.

- ***Customer Relationship Management*, CRM**. Els sistemes de gestió de relació amb el client són un tipus de programari amb orientació específica a l'atenció i el seguiment d'accions amb el client, el seu objecte és gestionar els contactes, les vendes i l'oferta/demanda amb els clients.

- ***Supply Chain Management*, SCM**. Els sistemes de gestió de cadenes de subministrament recopilen informació i processos en un sistema d'informació i integren dades relatives a operacions, logística, aprovisionament, emmagatzematge, distribució, postvenda i compres.

Avui dia el concepte de sistemes d'informació està fortament lligat al de tecnologies de la informació i les comunicacions (TIC).[6] Aquestes permeten automatitzar tot el cicle pel qual circula la informació a

6 <https://www.techopedia.com/definition/770/decision-support-system-dss>

les empreses, des de la recollida de dades mitjançant sensors o introducció manual de dades, passant pel modelatge de dades, fins a la presa de decisions. Gràcies a aquesta automatització, les empreses són capaces de donar resposta gairebé immediata a qualsevol operació i alhora optimitzar els processos interns i els serveis als clients. Les empreses s'han adonat que si apliquen aquestes tecnologies es poden posicionar enfront de la competència com una empresa més eficient, alhora que estalvien costos i milloren els beneficis.

Les TIC tenen l'avantatge d'oferir una alta disponibilitat de la informació gràcies a l'emmagatzematge. D'altra banda, la innovació contínua en aquestes tecnologies ha facilitat la implementació de metodologies d'estadística i intel·ligència de negoci de gran importància per a la presa de decisions i de gran valor per a les empreses.

Les eines d'intel·ligència de negoci (***Business Intelligence*, BI**) s'han implantat i establert en la majoria d'empreses. Les grans corporacions disposen d'un departament en el qual s'inverteixen recursos de BI, per la importància del valor afegit que impliquen aquestes eines per a fer front als competidors i conèixer l'entorn de mercat. Aquests departaments s'han convertit en el centre neuràlgic de les organitzacions, on s'analitzen i modelen les dades, es fan prediccions o es contrasten teories, entre moltes altres funcions.

Aquestes eines d'intel·ligència de negoci aporten valor a tots els processos de les organitzacions i permeten l'accés a tots els àmbits d'informació per a analitzar-la i donar resposta les diferents problemàtiques de les empreses.

D'aquesta forma, els departaments de BI responen a preguntes sobre el passat, mitjançant estadístiques descriptives amb dades històriques. Sobre aquest passat permeten saber per què s'han produït determinats fets gràcies a les anàlisis de determinació de causalitat amb sèries temporals. D'altra banda, permeten generar escenaris

amb models de multivariants, per a esbrinar quines serien les conseqüències de determinades decisions estratègiques. Aquests mateixos models també es poden utilitzar per a predir resultats en el futur i anticipar els efectes posteriors a una acció.

És molt important tenir una idea de quina serà la millor distribució d'informació. Perquè la informació sigui útil per a la presa de decisions són necessàries eines per a distribuir el coneixement generat en el moment adequat i als usuaris que puguin treure'n profit.

Els principals sistemes de presentació d'informació dels quals podem disposar són els informes estàndard i els sistemes de consulta. Els **informes estàndard** o *reports* són el sistema més tradicional de representar la informació. Són una presentació en dues dimensions on cadascuna pot tenir una estructura jeràrquica. Presenten resultats d'una forma predefinida i, per tant, no es poden utilitzar en una anàlisi dinàmica d'aquests resultats. És habitual l'automatització d'aquest tipus de presentació i són nombroses les eines que se solen emprar, des de Qlik o Tableau, passant per SAS, fins a eines més especialitzades com Report Builder.

Els **sistemes de consulta** o eines d'anàlisi *ad-hoc* són sistemes que permeten aconseguir un gran nivell de detall (*drill down*) o modificar els eixos d'anàlisi. El nivell més bàsic d'aquestes eines implica treballar amb dades extretes en una taula dinàmica (*pivot table*) d'un full de càlcul. Les eines més sofisticades permeten gestionar dinàmicament la consulta en el magatzem de dades abans que siguin presentades.

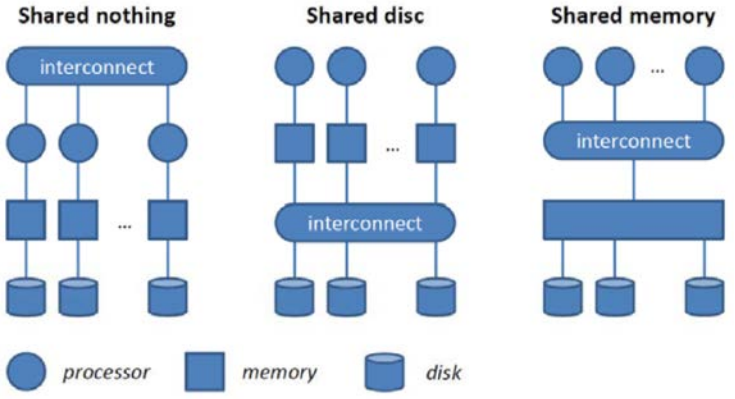
Aquests sistemes requereixen que els seus usuaris, a més del coneixement tècnic de l'eina, puguin utilitzar ajudes a la consulta que accedeixen a les definicions de la semàntica de les dades i, per tant, al contingut del magatzem de dades.

No podem entendre el concepte de dades massives (*Big Data*) sense entendre abans el de base de dades distribuïda. Aquest tipus de base de dades és aquella en què la gestió de dades es distribueix en diversos nodes d'una xarxa. Cada node és una base de dades per ella

mateixa, amb certa heterogeneïtat, i en tot moment els nodes es poden comunicar mitjançant la xarxa. Aquest esquema permet la computació paral·lelitzada, que dota l'arquitectura d'una gran potència de càlcul –en paral·lel– amb tanta velocitat i amb tanta capacitat de processament de dades com vulguem o siguem capaços de disposar (**escalabilitat**).

Aquest tipus d'arquitectures (*parallel distributed databases*) poden optar per una sèrie de combinacions de recursos en diversos àmbits –processador, memòria o disc– i l'objectiu de qualsevol d'aquestes permet dotar la infraestructura d'una millora substancial en tots o alguns dels seus àmbits, segons l'esquema triat. En la figura 7 veiem els diferents esquemes pels quals podem optar.

Figura 7. Esquemes de distribució segons elements d'infraestructura



Font: D. DeWitt, J. Gray (1992). *Parallel Database Systems: The future of High Performance Database Processing.*

Les **dades massives**[7] són un canvi de paradigma en el processament i el modelatge de dades, i una evolució i confluència de diverses tècniques complementàries i tecnologies que permeten abastar un **volum**, **varietat** i **velocitat** de processament de dades poc habituals, la qual cosa habitualment es coneix com les *3 V* del *Big Data*, encara que alguns autors n'inclouen d'altres, com la **veracitat** i el **valor**.

7 www.gartner.com/it-glossary/big-data/

Dins del *Big Data*, s'entén que, per a ser considerat com a tal, un procés i un conjunt de dades s'engloba dins d'aquest paradigma si compleix una sèrie de condicions de volum, velocitat, varietat, veracitat i valor. Les àrees relacionades amb aquests conceptes engloben diferents disciplines i funcionalitats.

En el concepte de volum i velocitat trobem les àrees de consultes declaratives i l'optimització de consultes. En el concepte de varietat i variabilitat hi ha les àrees de qualitat de dades, integració de dades, mineria web, mineria de textos i recuperació de dades desestructurades. En el de veracitat, s'hi inclouen la consistència de dades, el raonament estadístic, la incertesa, la connexió i la fusió de dades. En el concepte de valor, hi ha l'analítica de dades, i dins de la mineria de dades, la simulació, l'algorísmia i l'aprenentatge automàtic.

> En l'apartat «Anàlisi d'opinions per a la gestió del coneixement», veurem un exemple d'anàlisi de text (*text mining*) a partir de dades desestructurades provinents de Twitter, que ens permetrà comprovar la importància d'aquest tipus de dades. Aquest tipus d'anàlisi té sentit si es fa en temps real (*real time streaming*), el component velocitat de les 3 V del *Big Data*. D'altra banda, cal disposar de tècniques per a processar dades desestructurades, la varietat; i finalment, és evident la importància de disposar d'una estructura de base de dades distribuïdes si volem abastar el processament de piulades d'una xarxa de més de 500 milions d'usuaris que generen 65 milions de piulades diàries, el volum.

2.2. Estadística[8]

A continuació veurem les eines estadístiques que té a la seva disposició el científic de dades per a extreure coneixement de les dades, interpretar-ne els resultats i traduir-los a accions per a la presa de decisions en l'organització.

8 A la bibliografia hi trobareu referències específiques d'aquest apartat.

L'estadística té el seu origen en el segle XVIII. L'arrel de la paraula *estadística* és Estat. Això no és casualitat, ja que originalment la principal ocupació o font d'anàlisi de l'estadística eren les dades dels estats, i tots els esforços anaven encaminats a resoldre grans assumptes dels governs, però posteriorment el perímetre d'estudi es va anar ampliant i va arribar a abastar dominis tan dispars com l'economia, la biologia o el màrqueting.

La matèria primera de l'estadística són les dades. Podríem definir-la com una ciència, l'objecte de la qual és l'anàlisi, la interpretació, la representació i l'organització de dades.

Les dades que es recullen són de diversos tipus. El primer gran grup és el de les **dades quantitatives**, com per exemple la recollida d'alçades d'una classe d'alumnes. Un altre grup important és el de les **dades qualitatives**, com per exemple les localitats on resideixen els alumnes. Les dades qualitatives poden ser **ordinals**, com les que recollim amb el concepte de grandària (petit, mitjà, gran), o **nominals**, com els colors (blau, verd, groc…).

Amb les dades, l'estadística és capaç d'extreure coneixement, sintetitzar la informació, representar-la o fins i tot fer prediccions. Entre les diferents tècniques clàssiques destaquem les següents:

- **Anàlisi de la variància (ANOVA).** L'anàlisi de la variància (ANOVA) s'utilitza per a examinar diferències entre diversos grups. Si bé per a contrastar diferències entre dos grups podríem aplicar el contrast d'hipòtesi de la diferència de mitjanes, en el cas que vulguem veure si hi ha diferències entre més de dos grups podem considerar l'ANOVA. Aquesta tècnica s'emmarca dins dels models bivariants, atès que es treballa sobre mostres d'una variable i sobre diverses poblacions.

- **Contrast d'hipòtesis.** Els contrastos d'hipòtesis es fan a partir de la construcció d'intervals de confiança amb les dades d'una

mostra determinada. En els contrastos partim d'una hipòtesi que hem de confirmar o rebutjar partint dels valors d'un paquet estadístic i la seva funció de probabilitat. Una hipòtesi serà rebutjada o acceptada sempre amb un grau de significació estadística.

- **Regressió lineal múltiple**. Una regressió lineal simple ens permet relacionar dues variables, una d'explicativa i una altra d'explicada. Què passa quan una variable és explicada per més d'una variable independent? En general, veurem que la majoria de fenòmens que s'estudien dependran de més d'una variable. Per a fer aquests estudis, usarem la regressió lineal múltiple. La relació que s'estableix entre les variables independents (o explicatives) i la variable depenent (o explicada) és una relació lineal, que habitualment expressarem de la manera següent:

$$I = \beta 0 + \beta 1 \cdot X1 + \beta 2 \cdot X2 + \ldots + \beta k \cdot Xk$$

  En aquesta fórmula els diferents βi representen els diversos paràmetres del model. Conèixer la relació que hi ha entre les variables és conèixer quin és el valor d'aquests paràmetres. L'estimació mínima quadràtica dels paràmetres servirà per a saber quins són aquests valors. La significació individual i global dels paràmetres i el coeficient de determinació ens ajudaran a treure conclusions sobre la relació final que s'estableix per mitjà del model.

- **Anàlisi de conglomerats jeràrquica**. L'anàlisi de conglomerats jeràrquica és una de les tècniques més usades en modelitzacions multivariants per a agrupar elements que tenen propietats similars. Un element central d'aquesta anàlisi és el concepte de distància, que s'empra per a saber com és de proper un element respecte d'un altre mitjançant la comparació de les seves característiques.

  Per exemple, si volem saber si dos països, *X* i *I*, són similars quant al seu desenvolupament en telecomunicacions, haurem de considerar diferents indicadors en telecomunicacions (per exemple, línies de telèfon, nombre d'usuaris d'internet, etc.) i comparar-los amb una distància (possi-

blement la distància euclidiana, «de», és la més utilitzada a l'hora de fer aquesta anàlisi de classificació):

$$d_e = \sqrt{\sum_i (Xi - Yi)^2}$$

- **Anàlisi de conglomerats no jeràrquica**. L'anàlisi de conglomerats té com a objectiu l'agrupació de les variables analitzades en grups que tenen característiques similars i, alhora, permet reduir el nombre de casuístiques. En el cas de l'anàlisi de conglomerats no jeràrquic, els grups es defineixen prèviament ja sigui a partir dels criteris considerats a l'hora de definir les distàncies, o bé perquè cadascuna de les variables que considerem s'agrupa amb el veí més proper. En aquest cas, les necessitats de dades són menors.

- **Anàlisi de correspondències múltiple**. En l'anàlisi de correspondències simple es veu com podem analitzar relacions entre variables, que poden ser no mètriques. En concret, es veu com podem relacionar una llista d'atributs amb diferents individus (empreses, en aquest cas) a partir de la valoració dels usuaris. Ara, en l'anàlisi de correspondències múltiple no tindrem una llista única d'atributs, sinó que habitualment hi haurà diferents variables qualitatives amb diferents categories, entre les quals vulguem estudiar relacions de dependència i independència. La base d'aquesta anàlisi és disposar d'una taula de contingència *Z* en la qual es relacionin individus amb categories. Si fem una llista de totes les categories de manera consecutiva, podrem escriure:

$$z_{ij} = \begin{cases} 1, & \text{si l'individu } i \text{ escull la categoria } j \\ 0, & \text{si l'individu } i \text{ no escull la categoria } j \end{cases}$$

- **Anàlisi de sèries temporals**. Suposem que volem estudiar l'evolució d'una variable (per exemple, del valor d'un determinat fons d'inversió) al llarg del temps. Aquest valor, que anem observant a mesura que passen els dies, constitueix un exemple del que es de-

nomina *sèrie temporal*. L'estudi de les sèries temporals és bàsic en l'àmbit econòmic i empresarial, atès que és el mètode quantitatiu més utilitzat a l'hora de fer previsions o de crear expectatives futures. Una sèrie temporal està formada per quatre components: la tendència, el cicle, el component estacional i el component erràtic. L'anàlisi clàssica de sèries temporals estudia el valor que pren la sèrie temporal a cada moment del temps en funció de tots o d'alguns d'aquests quatre components segons un esquema additiu o multiplicatiu. Per tant, segons el nombre de components que figuri en una sèrie temporal s'utilitzaran uns mètodes d'estimació o uns altres. Per a detectar quins d'aquests components conté una sèrie temporal podem utilitzar els mètodes gràfics o recórrer a contrastos estadístics; el contrast de Daniel ens ajudarà a esbrinar si una sèrie té tendència i el de Kruskal-Wallis ens determinarà si la sèrie té component estacional.

El coeficient de correlació entre dues variables no contemporànies (Xt i $Yt - k$) es denomina *correlació serial*, i la correlació d'una variable amb si mateixa diferida k períodes (Yt i $Yt - k$), *autocorrelació*. Calcular aquestes correlacions mitjançant la fórmula de Pearson és com si calculéssim un coeficient de correlació.

El que es busca trobar amb la correlació serial és, per exemple, quina relació hi ha entre el valor d'una variable en un determinat període amb el valor d'una altra variable en el període immediatament següent. En canvi, amb l'autocorrelació es pretén trobar la relació entre el valor d'una variable en un període concret i el valor de la mateixa variable en el període anterior o posterior.

- **Anàlisi factorial**. En l'anàlisi factorial s'engloben totes aquelles tècniques que tenen com a objectiu reduir el nombre de variables amb les quals treballem i, per tant, la dimensió de la matriu de dades inicial. A partir d'una base de dades s'obté un nombre reduït de factors o noves variables que sintetitzen la informació de partida i que permeten estudiar les relacions existents entre les variables inicials. En aquestes tècniques és de gran ajuda la representació gràfica simplificada de les files i columnes, ja que moltes vegades resulta clau per a la interpretació dels resultats.

  Encara que totes les tècniques d'anàlisi factorial tenen una estructura comuna, segons com sigui la naturalesa de les dades

s'originen diferents mètodes. L'anàlisi factorial de components principals analitza taules de variables quantitatives o mètriques (individus x variables mètriques). L'anàlisi factorial de correspondències simples analitza taules de contingència i, en general, qualsevol taula de nombres positius, sempre que la suma d'una fila i una columna tinguin sentit i es puguin interpretar. L'anàlisi factorial de correspondències múltiples analitza taules de variables qualitatives (individus x variables qualitatives).

2.3. Aprenentatge automàtic

L'**aprenentatge automàtic** o *machine learning* (ML) és el camp de la informàtica que estudia mètodes automàtics per tal de fer prediccions basades en experiències anteriors d'un sistema.

Una de les finalitats d'aquest camp és produir «bons» models o introduir millores en els models tradicionals.

Els models són descripcions compactes d'una mostra de dades que permet predir escenaris i generar simulacions. L'ML s'ocupa de diferents àrees però les més habituals són els mètodes de classificació i els models de predicció. L'ML és una de les branques de la intel·ligència artificial i, per tant, una de les seves finalitats és dotar les computadores de la capacitat d'aprendre.

És un camp que beu de diverses fonts, la més destacable, sens dubte, és l'estadística, amb la qual comparteix algunes àrees d'estudi. De vegades trobem metodologies estadístiques englobades dins del ventall de tècniques que s'ensenyen habitualment en els cursos d'aprenentatge automàtic i analítica. Entre els mètodes més coneguts trobem els següents:

- **Xarxes neuronals**. Una xarxa neural està formada per "neurones", en forma de nodes d'un graf i unes connexions o capes que els comuniquen que reben una ponderació. Les xarxes formades representen allò que volem analitzar, que queden reflectides de forma analítica en forma de grafs. S'utilitzen símbols per a representar-les i resulta difícil extreure parts del conjunt ja que la seva forma queda subjecta a una visió global del plantejament. Amb les Neural networks es realitza la inferència robusta del problema que s'analitza i mitjançant un procés d'aprenentatge i adaptació són capaços de fer una representació del subjecte d'anàlisi i de fer predicció de comportament i d'estructures o classes a partir de modificacions adaptatives en els seus pesos.

- **Mètodes de vectors de suport (*Support Vector Machines*, SVM)**. Es basen en el principi de Minimització del Risc Estructural de la teoria de l'aprenentatge computacional. En els problemes de classificació intenten trobar la solució al problema de minimització del risc estructural, buscant una manera o hiperplà que millor separa les classes. Els SVM són molt multidisciplinaris. Podem utilitzar funcions de llindar lineals per trobar la solució al problema de classificació, però sens dubte el seu potencial rau a millorar aquesta linealitat mitjançant altres funcions més adequades (nucli) que poden ser polinomials, radials (RBF), sinusoïdals i de moltes altres formes. A més de fer-se servir en problemes de classificació també s'utilitzen per a la predicció i per a problemes de segmentació.

- **Xarxes bayesianes**. Un tipus de sistemes intencionals són les anomenades xarxes causals probabilístiques. Una xarxa causal probabilística, o xarxa bayesiana, es defineix sobre un graf dirigit acíclic (GDA), on els nodes representen variables i els arcs del graf descriuen relacions de dependència, tipus causa-efecte, entre les variables. Si un node amb una variable A és pare d'un node (predecessor immediat) amb una variable B, llavors es considera que A és una causa directa de B, o bé que B és un efecte directe d'A. Aquestes relacions de dependència es quantifiquen en cada node amb la distribució de probabilitats condicionada de la variable associada a aquest node respecte a les variables en els

nodes pare. Per tant, en una xarxa causal tenim alhora un component qualitatiu i un component numèric de representació del coneixement: el qualitatiu descriu les relacions de dependència o independència entre les variables que intervenen en la descripció del problema, i el quantitatiu o numèric quantifica aquestes relacions mitjançant probabilitats (o més en general podria ser mitjançant mesures d'incertesa).

- **Arbres de regressió i classificació (C&RT).** En termes generals, el propòsit de les anàlisis per mitjà d'algorismes de construcció d'arbres és determinar un conjunt de condicions lògiques (divisòries) que permeten una predicció o classificació precisa dels casos. Els problemes de tipus de regressió són generalment aquells en els quals s'intenta predir els valors d'una variable contínua a partir d'una o més variables predictores contínues i/o categòriques. Els problemes de tipus de classificació són generalment aquells en els quals intentem predir els valors d'una variable depenent categòrica (classe, pertinença a un grup, etc.) a partir d'una o més variables predictores contínues i/o categòriques.

- **Algoritmes genètics**. Els algorismes genètics són algorismes de cerca estocàstics inspirats en els fenòmens naturals d'herència genètica i que el millor és el que sobreviu (supervivència de les espècies). En una població d'individus, les noves generacions estan més adaptades al mitjà que les precedents i, de mitjana, les noves poblacions seran més ràpides, amb un grau de mimetisme més gran, etc. que les anteriors perquè això és el que els permet sobreviure. L'analogia amb el procés biològic de l'evolució dirigeix tots els passos dels algorismes genètics. Així, considerarem poblacions d'individus representats pels seus cromosomes (cromosomes formats per gens), i davant d'un cromosoma considerarem el seu encreuament amb un altre cromosoma o la seva mutació. Els creuaments i les mutacions portaran a noves generacions de poblacions. Com tots els mètodes de cerca, els algorismes genètics permeten trobar una solució a un problema determinat. No obstant això, com que són algorismes estocàstics, normalment no troben la millor solució al problema sinó una que s'aproxima a l'òptima.

Les solucions

1. Sistemes de recomanació de productes

A continuació veurem com un científic de dades es pot enfrontar al disseny i la implementació d'un sistema de recomanació de productes amb l'objecte de maximitzar les vendes en el canal en línia de la seva empresa.

Els sistemes de recomanació de productes tenen una finalitat intrínseca més enllà de millorar l'experiència de compra per als usuaris, que no és altra que incrementar les vendes per al distribuïdor. Mitjançant algorismes que analitzen els patrons de compra, s'extreuen comportaments i regles que es repeteixen al llarg de l'historial de dades i que suggereixen relacions entre productes i/o compradors que van més enllà de la casualitat.

Tenim tres grans grups de sistemes de recomanació:

1) **Basats en els compradors que compren productes similars**. Aquests busquen compradors similars en els seus actes de compra. Un sistema d'aquest tipus analitza les compres dels consumidors, busquen evidències de similitud entre parelles i dedueixen quins són els productes en els quals no hi ha hagut coincidència. Finalment, recomanen aquests productes no coincidents als compradors semblants, basant-se en la similitud en les seves preferències i sota la hipòtesi que el producte no coincideix a causa que el comprador en desconeix l'existència, però podria agradar-li, atès que altres compradors similars han decidit comprar-lo.

2) **Basats en els productes que compren compradors similars**. Aquests analitzen els patrons des de la perspectiva del producte, busquen conjunts de productes molt similars segons els compradors que els compren. Quan un producte ha estat comprat per tipus de compradors molt similars es pot recomanar un altre pro-

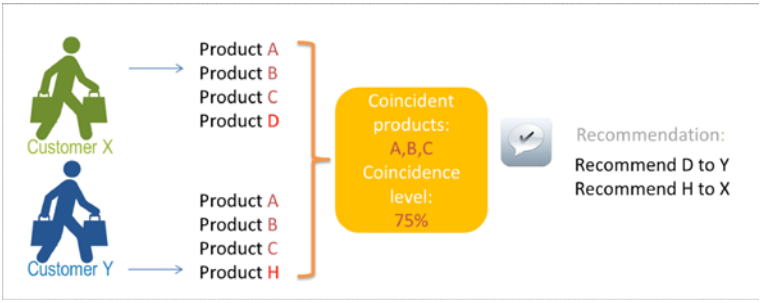
ducte complementari que també compren els consumidors que compren el primer.

3) **Basats en productes que es compren habitualment amb altres productes**. Aquest tipus de sistema recorre a l'historial de compres buscant patrons que es repeteixen de manera freqüent. Busca productes que habitualment han estat comprats conjuntament en un gran nombre de cistelles o d'ocasions de compra. Dedueix quin producte es compra habitualment juntament amb un altre o altres productes.

Entre els tipus d'algorismes basats en la similitud de compradors i la similitud de productes hi hauria els que estan basats en distàncies com el ***k-nearest neighbor***, que analitza els hàbits de compra d'una base de dades de clients, busca les semblances entre tots els clients a partir del seu historial de compres i localitza aquells clients que més s'assemblen, de tal manera que és capaç de recomanar productes que el client objectiu encara no ha comprat però que sí que ha comprat un client que ha fet un tipus de compres molt semblants a les seves.

Per exemple, imaginem un cas com el de la figura 8, en el qual tenim un client «X» d'Amazon que ha comprat els productes A, B, C i D. D'altra banda, trobem un altre client «I», que ha comprat A, B, C i H. L'algorisme trobaria similituds entre ambdós consumidors, ja que tots dos han comprat els productes A, B i C, i al seu torn, seria capaç de recomanar a «X» la compra de H, atès que «I» és un tipus de client molt semblant que, a més de tenir gustos molt semblants a «X», també ha comprat el producte H, que «X» encara no ha provat.

Figura 8. Exemple de recomanació per compradors similars (*alikes*)



Font: elaboració pròpia

Un altre tipus d'algorismes amb finalitats molt semblants serien els englobats dins del camp de regles d'associació, com **Apriori**. Aquest algorisme busca entre la base de dades de compres, patrons i regles d'associació. D'aquesta manera, és capaç de detectar quins productes té més propensió a comprar el client si també ha comprat un altre producte o una combinació de productes en una cistella determinada.

La forma en què l'algorisme genera recomanacions és el següent: imaginem un cas com el de la figura 9, en el qual tenim una base de dades on es repeteix la cistella amb els productes A, B, C un nombre elevat de vegades (per exemple, 90 de cada 100 vegades). En aquest cas l'algorisme trobarà que 90 de cada 100 vegades que un client ha comprat A i B també ha comprat C. De la mateixa manera indicarà que 90 de cada 100 vegades que un client ha comprat A i C també ha comprat B i, finalment, també detectarà que 90 de cada 100 vegades que un client ha comprat B i C també ha comprat A. Per tant, si l'algorisme es proposa recomanar un producte a un comprador que ha fet una compra d'A i C, podria recomanar el producte B per a la compra següent o en el mateix acte de compra.

Figura 9. Exemple de recomanació per patrons de compra

Font: elaboració pròpia

En aquest cas pràctic veurem com una empresa com Amazon es podria beneficiar d'una base de dades amb historial de productes peribles d'un panell de consum d'una font provinent d'una empresa privada d'estudis de mercat. En l'exemple usarem dades simulades de compra en línia. L'objectiu és que l'estudiant sàpiga comprendre la utilitat del mètode i pugui interpretar els resultats amb la finalitat d'obtenir coneixement d'aquest algorisme del camp de la mineria de dades.

Amb el material dels punts 1, 2 i 3 l'estudiant ha vist conceptes rellevants sobre el sector del comerç al detall del mercat espanyol, també ha entès la importància de triar el producte i el canal, i finalment ha vist com esprémer al màxim els mètodes que ofereixen els camps de la intel·ligència de negoci, la mineria de dades i l'estadística.

El cas pràctic posa l'estudiant en la pell d'un científic de dades, que té com a objectiu generar un sistema de recomanació en temps real de productes sobre la base de les compres realitzades en un establiment de venda en línia que ofereix productes d'alimentació, similar a Amazon Prime Now. Sens dubte, una empresa com Amazon parteix d'una situació de líder de mercat en el comerç electrònic i d'un potencial en operacions i logística dins d'aquest àmbit que difícilment les altres empreses poden aconseguir. El punt en què Amazon té menys capacitat és en el coneixement de mercat del sector de gran consum. Si una empresa de venda solament a través d'internet vol fer-se un lloc i entrar a competir en aquest sector tan madur, haurà de recórrer a fonts de dades de tercers i al saber fer en metodologies aplicades al comerç al detall espanyol. D'altra banda, haurà d'adaptar els seus processos i metodologies a les de la nova categoria, i utilitzar eines de mineria de dades com les que es veuran en aquesta solució.

Anteriorment en aquest apartat, hem definit dos mètodes que poden ser útils per a crear un sistema de recomanació de productes. Ens centrarem en el segon, l'Apriori, que pertany al camp d'algorismes de **regles d'associació** (*association rules field*). El problema que analitzen es coneix com **anàlisi de cistelles de compra (ACC)**. L'ACC assumeix que el mercat està compost per un conjunt elevat d'elements (ítems), com el pa, l'aigua, els iogurts, etc., que es poden combinar en les cistelles dels compradors. L'objectiu que es persegueix és conèixer quins elements apareixen en la mateixa cistella un nombre més gran de vegades i alhora usar aquesta informació per a obtenir algun tipus de benefici, habitualment econòmic. Les conclusions que s'extreuen d'aquests patrons poden servir per a saber com combinar promocions aprofitant les sinergies entre productes, o com situar en els lineals virtuals de compra aquests productes que habitualment es compren alhora.

En el context de gran consum, tenim cistelles i tenim productes, encara que podríem utilitzar aquesta estratègia en altres contextos molt diferents amb finalitats similars. Per això, definim el concepte de **transaccions**, que són tot el conjunt d'elements que tenen lloc en un moment determinat, en el nostre cas productes de cada cistella, d'un en un. I per **elements** entenem esdeveniments que es combinen per a formar una transacció; en el nostre cas, productes que formen cada cistella, d'un en un.

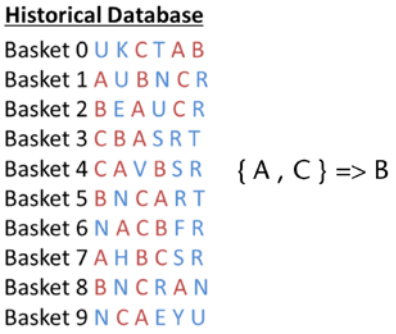
De totes les possibles combinacions d'elements, estem interessats en particular en les **combinacions d'elements més freqüents (CEF)**, és a dir, aquelles combinacions d'elements que es repeteixen més sovint en les cistelles.

En un segon estadi analitzarem les regles d'associació (AR), és a dir, una vegada coneguem les combinacions més freqüents d'elements podrem generar recomanacions segons com s'associen els diferents elements.

> Entenem per **regla** un conjunt d'elements que generen un patró o una recomanació.

Per exemple, en aquesta figura 10 s'observa la regla següent:

Figura 10. Exemple amb 10 cistelles



Font: elaboració pròpia

Per a entendre tot el procés necessitem també definir prèviament el concepte de suport, freqüència, confiança i elevació. El **suport** és el nombre de vegades que es repeteix una combinació de productes. El suport mínim serà un nombre de suport que desitgem considerar com a mínim. És a dir, un element serà freqüent sempre que el seu suport sigui superior al valor de suport mínim. Aquest suport s'expressa com el coeficient entre el nombre de transaccions en què apareix una certa combinació d'elements respecte al total de transaccions possibles, això és, com un percentatge respecte al nombre total de transaccions.

> En l'exemple de la figura 10 tenim que el suport de { A, B, C } és 9/10, ja que trobem aquesta combinació 9 vegades i tenim en total 10 cistelles.

La **freqüència** és el nombre de vegades que un conjunt d'elements expressats com una regla es repeteix al llarg de totes les cistelles possibles, o el que és el mateix, el nombre de cistelles en les quals apareix aquesta combinació d'elements en tot el conjunt de transaccions.

> Per exemple, el suport de { A, C } és 1.0, ja que és present en totes les cistelles. La regla { A, C } => B té una freqüència que correspon al suport de { {A, C} $\cup$ B }, és a dir, 0.9, perquè els dos conjunts (els tres elements) apareixen combinats 9 vegades de 10.

La **confiança** és el quocient entre el suport de la regla i els elements de l'esquerra de la regla.

> En el nostre cas el suport dels elements de l'esquerra de la regla { A, C } és 1.0 i la freqüència de la regla { A, C } => B és de 0.9, i per tant la confiança és de 0.9.
>
> Cal observar que aquesta confiança coincideix amb la de la regla { A, B } => C formada pels mateixos elements, ja que el suport de la part esquerra { A, B } és de 0.9 i la freqüència de la regla és de 0.9, per tant, la confiança és d'1.0 (de 0.9/0.9).

Finalment, l'**elevació** és un concepte que intenta extreure regles rellevants i que es faci una cerca intel·ligent de les regles rellevants.

Parteix del supòsit que si l'element de la dreta de la regla és molt freqüent, llavors és molt probable que qualsevol regla aparegui amb una confiança i un suport elevats solament pel fet d'incloure aquest element. Amb aquest nou valor eliminem aquest biaix. Per a calcular-ho farem el quocient entre la confiança de la regla i el suport de l'element de la dreta.

> En l'exemple { A, C } => B tenim una confiança de 0.9 i un suport per a B de 0.9, per tant, una elevació d'1.

Per a treballar amb aquesta resolució l'alumne s'haurà d'instal·lar el programari gratuït R.[1]

A continuació, se subministra, mitjançant l'enllaç següent, una base de dades de compra en línia amb dades simulades:

Enllaç de la base de dades de cistelles simulades[2]

Amb aquesta base de dades, aplicarem la tècnica Apriori i se'ns demanarà jugar amb la taula de dades intentant trobar un conjunt de deu regles d'associació rellevants. Però abans haurem de decidir si volem conservar o eliminar els valors NA (valors no disponibles) i altres processos relacionats amb l'assegurament de la qualitat de dades. Aquesta fase prèvia és la que es coneix com a *preprocessament* (*pre-processing*).

Partim d'un *dataframe*, al qual hem anomenat **df**, amb l'estructura que mostra la figura 11. Tenim tantes columnes com el nombre màxim de productes oposats en una cistella i tantes files com nombres de cistelles que han emplenat els consumidors.

---

1 És recomanable llegir prèviament els primers tres capítols de Castillo, A. J. S. «Métodos Estadísticos con R y R Commander». [Recurso electrónico gratuito] <http://cran.r-project.org/doc/contrib/Saez-Castillo-RRCmdrv21.pdf>

2 Per a l'anàlisi descrit a continuació, el consultor proporcionarà enllaç a un fitxer mitjançant l'aula.

Figura 11. Taula encreuada de cistella per producte

BVE	Element1	Element2	Element3	Element4	Element5	Element6	...	Element37
Basket1	ACEITUNAS	CERVEZA CON ALCOHOL	FRUTA	JUDIAS VERDES	LECHE CONDENSADA	POLLO ASADO	...	...
Basket2	ACEITE	ACEITUNAS	PAN MOLDE	PATATAS FRITAS	PAVO	VERDURA	...	...
Basket3	ALCACHOFAS	CERVEZA CON ALCOHOL	FLAN	FRUTA	HUEVOS	PAN TOSTADO	...	...
Basket4	ACEITUNAS	CERVEZA CON ALCOHOL	FRUTA	PAN	PATATAS FRITAS	PLATO PREPARADO	...	...
Basket5	COLORACION	FRUTA	MEMBRILLO	PAN	VERDURA	YOGUR	...	...
Basket6	LECHE CONDENSADA	LEGUMBRES	VERDURA				...	...
Basket7	CERVEZA CON ALCOHOL	CONSERVAS DE ATUN	CONSERVAS MEJILLONES	CONSERVAS SARDINAS	ESPECIAS	FRUTA	...	...
Basket8	CERVEZA CON ALCOHOL	FRUTA	LECHE CONDENSADA	POLLO ASADO	VERDURA		...	...
Basket9	ACEITUNAS	CERVEZA SIN ALCOHOL	FRUTA	PAN MOLDE	SOPAS	VERDURA	...	...
Basket10	ACEITUNAS	CACAO SOLUBLE	CAFE	CERVEZA CON ALCOHOL	CONSERVA DE PESCADO	CONSERVAS MEJILLONES	...	...
Basket11	CALDO	FLAN	FRUTA	PAN	POLLO CONGELADO	VERDURA	...	...
Basket12	ACEITUNAS	CERVEZA SIN ALCOHOL	HUEVOS	LEJIA	QUESO	SOPAS	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
Basket 6472	CAFE	CERVEZA CON ALCOHOL	CHORIZO	FRUTA	FUET	LECHE	...	...

Font: elaboració pròpia

Com es pot observar, el nombre màxim d'elements que hem trobat en una cistella és de 37. És a dir, la cistella més gran oposada té 37 productes diferents. D'altra banda, tenim un total de 6.472 cistelles en la base de dades simulada.

El fet de tenir cistelles més petites fa que apareguin valors NA en algunes cel·les. Hem de plantejar-nos si té sentit en aquesta etapa del procés. L'etapa següent consistirà a passar cada fila del nostre *dataframe* a una llista. Entendre què fem en aquesta etapa ens permetrà veure com plantejar aquest problema.

Per a passar la taula a una llista tenim diverses opcions. Com que el nostre conjunt de dades és petit, podem fer-ho amb un simple bucle de la manera següent, executant aquestes línies de *script*.

```
a_list<-list()

a_list<-list()
for (i in 1:dim(df)[1]){
  if( length(df[i,which(df[i,]!="")]) >= 2 )
    a_list<-append(a_list, list( as.character(df[i,which(d
f[i,]!="")]) ))
}
```

El preprocessament que hem necessitat es concentra en la condició següent, la qual filtra cistelles d'un sol element.

```
length(df[i,which(df[i,]!="")]) >= 2
```

Això té sentit, i es basa en el fet que volem obtenir recomanacions de productes sobre la base de la compra d'altres productes, i una cistella amb un únic element no ens proporciona la informació que necessitem.

A continuació, haurem de posar un nom als elements de les nostres llistes de cistelles per tal d'identificar-los. Utilitzarem la sentència següent:

```
names(a_list) <- paste("Basket",c(1:dim(df)[2]), sep = "")
```

Vegem quin aspecte té la nostra llista per a entendre bé quina acció hem dut a terme. Cal observar que per a cada cistella tenim el sufix «Basket» seguit d'un valor o seqüència autonumèrica (Basket1, Basket2, Basket3…). Usarem la funció «head()» per a veure'n el resultat:

```
head(a_list)
$Basket1
[1] "ACEITUNAS"           "CERVEZA CON ALCOHOL" "FRUTA"
[4] "JUDIAS VERDES"       "LECHE CONDENSADA"    "POLLO ASA-
DO"
[7] "SOPAS"               "VERDURA"

$Basket2
[1] "ACEITE"         "ACEITUNAS"      "PAN MOLDE"      "PA-
TATAS FRITAS"
[5] "PAVO"           "VERDURA"

$Basket3
[1] "ALCACHOFAS"          "CERVEZA CON ALCOHOL" "FLAN"
[4] "FRUTA"               "HUEVOS"              "PAN TOS-
TADO"
[7] "QUESO"               "VERDURA"             "ZUMO"

$Basket4
 [1] "ACEITUNAS"           "CERVEZA CON ALCOHOL" "FRUTA"
 [4] "PAN"                 "PATATAS FRITAS"      "PLATO
PREPARADO"
 [7] "SALCHICHAS"          "VERDURA"             "VINAGRE"
[10] "VINO"

$Basket5
[1] "COLORACION" "FRUTA"      "MEMBRILLO"  "PAN"
"VERDURA"
[6] "YOGUR"      "ZUMO"

$Basket6
[1] "LECHE CONDENSADA" "LEGUMBRES"        "VERDURA"
```

D'aquesta forma, veiem la composició de productes de les tres primeres cistelles.

El pas més important per a usar l'algorisme *Apriori* implica instal·lar i carregar la llibreria «arules», per la qual cosa utilitzarem les funcions **install.packages** i **library**, com es mostra a continuació.

```
install.packages("arules")
library(arules)
```

El pas següent és passar la nostra llista al format de transaccions necessari perquè la funció entengui les dades que li proporcionem.

```
trans <- as(a_list, "transactions")
```

Podem veure si tot ha funcionat extraient un resum del resultat amb la funció **summary()**.

```
summary(trans)
transactions as itemMatrix in sparse format with
 5802 rows (elements/itemsets/transactions) and
 197 columns (items) and a density of 0.03582346

most frequent items:
VERDURA   FRUTA   LECHE     PAN   QUESO (Other)
   2273    2077    1491    1398    1357   32350

element (itemset/transaction) length distribution:
sizes
  2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
17  18  19  20
666 830 740 666 554 421 353 276 228 170 150 123 114 103  63
53  43  44  32
 21  22  23  24  25  26  27  28  29  30  31  32  33  34  35
36  37
 30  29  20  11  19   8   9   7   4   7   5   7   5   3   5
3   1
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   5.000   7.057   9.000  37.000

includes extended item information - examples:
          labels
1         ACEITE
2      ACEITUNAS
3 ACONDICIONADOR

includes extended transaction information - examples:
  transactionID
1       Basket1
2       Basket2
3       Basket3
```

L'*output* ens informa primer sobre les dimensions del nostre conjunt de transaccions. Veiem que tenim 5.802 cistelles, que hem reduït en comparació amb les 6.472 transaccions originals després d'eliminar les cistelles d'un únic producte (compra esporàdica). A continuació, veiem que els ítems que apareixen en més cistelles són **verdura**, **fruita**, **llet**, **formatge** i **pa**, i el nombre de cistelles en les quals apareix cada producte. El resultat també ens mostra la distribució de cistelles per nombre d'elements. Així, per exemple, podem veure que tenim 666 cistelles amb dos productes, 830 amb tres productes i en l'extrem superior una única cistella amb 37 elements (cistella de càrrega o estoc). També veiem algunes dades estadístiques descriptives per a aquesta distribució, amb el mínim, la mediana, la mitjana, el rang interquartílic i el màxim. En les dues últimes sortides de resultats, la funció ens mostra un exemple d'elements (productes) i un exemple de transaccions (cistelles).

El pas següent és llançar l'algorisme. Ja hem definit el concepte de suport i de suport mínim. Recordeu que aquest és el nombre mínim de suport que admetem per a una regla, és a dir, no inclourem cap regla que tingui un suport per sota d'aquest valor. També és necessari usar el valor mínim de confiança, que ens permetrà filtrar les regles més rellevants. Utilitzarem, per a això, la funció **Apriori()**.

```
rules<-apriori(trans,parameter=list(supp=.01, conf=.1,
target="rules"))
```

Per a inspeccionar les regles tenim precisament la funció **Inspect**, que ens mostrarà un resum de les principals regles. El nombre de regles que hem de visualitzar les especificarem amb el paràmetre «n».

```
inspect(head(sort(rules,by="lift"),n=5))
```

Podem extreure ja algunes conclusions. Per exemple, a partir dels conjunts de dos elements de les nostres dades simulades, podem veure que els consumidors que compren **flam** també solen incloure **natilles** a la mateixa cistella. Per a tres elements, veiem que els consumidors que inclouen **aigües** i **verdures** també solen incloure **pernil curat**. Els coeficients d'elevació d'aquestes regles són molt elevats, i per tant podem dir que aquestes regles encara que tenen suport i confiança baixos tenen molta rellevància.

A continuació, ens pot interessar conèixer com són les cistelles que inclouen **cues**. Per a aquesta finalitat, la funció ens deixa escollir filtrar el conjunt de regles per a aquelles que tenen l'element **cues** en el costat dret amb el paràmetre «rhs». Vegem-ne un exemple:

```
> rulesRCola <- subset(rules, subset = rhs %in% "COLA" &
lift > 1.5)
> inspect(rulesRCola)
   lhs                          rhs       support confiden-
ce     lift
1  {BEBIDA NARANJA CON GAS} => {COLA} 0.01206481  0.3465347
3.367829
2  {CERVEZA SIN ALCOHOL}    => {COLA} 0.01016891  0.2543103
2.471539
3  {BEBIDA SIN GAS}         => {COLA} 0.01172010  0.2605364
2.532047
4  {SALCHICHAS}             => {COLA} 0.01223716  0.2500000
2.429648
5  {CERDO}                  => {COLA} 0.01034126  0.1617251
1.571740
```

```
6  {EMBUTIDO}              => {COLA} 0.01465012  0.2492669
2.422523
7  {LIMPIADOR HOGAR}       => {COLA} 0.01361599  0.2393939
2.326572
8  {TOMATE FRITO}          => {COLA} 0.01051362  0.1626667
1.580891
9  {POLLO CONGELADO}       => {COLA} 0.01275422  0.1859296
1.806975
10 {CERVEZA CON ALCOHOL}   => {COLA} 0.01895898  0.2135922
2.075816
11 {PASTAS BOLLERIA}       => {COLA} 0.01430541  0.1736402
1.687538
12 {FRUTOS SECOS}          => {COLA} 0.01620131  0.1649123
1.602715
13 {AGUAS}                 => {COLA} 0.02016546  0.1700581
1.652726
14 {CAFE}                  => {COLA} 0.01723544  0.1639344
1.593212
15 {PATATAS FRITAS}        => {COLA} 0.02223371  0.1804196
1.753424
16 {AGUAS,
    VERDURA}               => {COLA} 0.01068597  0.2357414
2.291075
17 {LECHE,
    PATATAS FRITAS}        => {COLA} 0.01051362  0.2618026
2.544353
18 {FRUTA,
    PATATAS FRITAS}        => {COLA} 0.01034126  0.2222222
2.159687
19 {PATATAS FRITAS,
    VERDURA}               => {COLA} 0.01103068  0.2064516
2.006419
20 {FRUTA,
    PASTEL BOLLERIA}       => {COLA} 0.01051362  0.1799410
1.748773
21 {PASTEL BOLLERIA,
    VERDURA}               => {COLA} 0.01275422  0.2078652
2.020157
```

```
22 {DULCES,
    VERDURA}                => {COLA} 0.01034126  0.1666667
1.619765
23 {LECHE,
    QUESO}                  => {COLA} 0.01085832  0.1567164
1.523063
24 {LECHE,
    VERDURA}                => {COLA} 0.01723544  0.1618123
1.572588
```

Es pot observar que en les cistelles en què el consumidor compra **beguda de taronja amb gas** sol aparèixer també el producte **cola**. A partir de la regla 16 veiem regles amb tres elements i podem destacar que en les cistelles en què apareixen els productes **aigües** i **verdura** o en les cistelles que apareixen **llet** i **patates fregides** també hi apareix **cola**.

Per a veure les regles en les quals apareix **cola** en la part esquerra, fem el mateix però utilitzant el paràmetre «lhs». Vegem-ne un altre exemple:

```
> rulesInLCola <- subset(rules, subset = lhs %in% "COLA" &
lift > 1.5)
> inspect(rulesInLCola)
   lhs                rhs                           support
confidence     lift
1  {COLA}          => {BEBIDA NARANJA CON GAS} 0.01206481
0.1172529 3.367829
2  {COLA}          => {BEBIDA SIN GAS}         0.01172010
0.1139028 2.532047
3  {COLA}          => {SALCHICHAS}             0.01223716
0.1189280 2.429648
4  {COLA}          => {CERDO}                  0.01034126
0.1005025 1.571740
5  {COLA}          => {EMBUTIDO}               0.01465012
0.1423786 2.422523
6  {COLA}          => {LIMPIADOR HOGAR}        0.01361599
0.1323283 2.326572
```

```
7  {COLA}           => {TOMATE FRITO}           0.01051362
0.1021776 1.580891
8  {COLA}           => {POLLO CONGELADO}        0.01275422
0.1239531 1.806975
9  {COLA}           => {CERVEZA CON ALCOHOL}    0.01895898
0.1842546 2.075816
10 {COLA}           => {PASTAS BOLLERIA}        0.01430541
0.1390285 1.687538
11 {COLA}           => {FRUTOS SECOS}           0.01620131
0.1574539 1.602715
12 {COLA}           => {AGUAS}                  0.02016546
0.1959799 1.652726
13 {COLA}           => {CAFE}                   0.01723544
0.1675042 1.593212
14 {COLA}           => {PATATAS FRITAS}         0.02223371
0.2160804 1.753424
15 {COLA,
    VERDURA}        => {AGUAS}                  0.01068597
0.2540984 2.142847
16 {COLA,
    PATATAS FRITAS} => {LECHE}                  0.01051362
0.4728682 1.840095
17 {COLA,
    LECHE}          => {PATATAS FRITAS}         0.01051362
0.2946860 2.391284
18 {COLA,
    FRUTA}          => {PATATAS FRITAS}         0.01034126
0.2857143 2.318482
19 {COLA,
    VERDURA}        => {PATATAS FRITAS}         0.01103068
0.2622951 2.128442
20 {COLA,
    FRUTA}          => {PASTEL BOLLERIA}        0.01051362
0.2904762 1.961982
21 {COLA,
    VERDURA}        => {PASTEL BOLLERIA}        0.01275422
0.3032787 2.048455
22 {COLA,
```

```
    VERDURA}        => {DULCES}                0.01034126
0.2459016 1.668680
23 {COLA,
    PAN}            => {VERDURA}               0.01223716
0.5916667 1.510273
24 {COLA,
    YOGUR}          => {QUESO}                 0.01034126
0.4225352 1.806595
25 {COLA,
    QUESO}          => {YOGUR}                 0.01034126
0.3680982 1.778273
26 {COLA,
    QUESO}          => {LECHE}                 0.01085832
0.3865031 1.504018
27 {COLA,
    FRUTA}          => {LECHE}                 0.01413306
0.3904762 1.519479
28 {COLA,
    VERDURA}        => {LECHE}                 0.01723544
0.4098361 1.594815
```

Veiem que els resultats són molt semblants. Hi ha una simetria pràcticament exacta entre les regles segons la posició en la regla del producte triat. No obstant això, cal observar que no és ben bé així. En l'anterior resultat vèiem la regla {CERVESA SENSE ALCOHOL} => COLA, però si ens fixem en aquesta sortida de resultats no trobem la regla simètrica {COLA} => CERVESA SENSE ALCOHOL. Això indica que gairebé sempre que es compra **cervesa sense alcohol** el client també s'emporta **cola**, però no és habitual que quan un client compri **cola** també inclogui **cervesa sense alcohol** a la cistella.

Vegem-ne un altre exemple. Ara filtrarem pel producte **dolços**, que inclou el concepte de **galetes dolces**. En aquesta ocasió, utilitzarem la funció «Label()», que dóna com a resultat un *output* més resumit. En el mateix *output* hem inclòs les regles amb el producte a banda i banda de la regla per a més comoditat.

```
> rulesInNATA <- subset(rules, subset = lhs %in% "DULCES" &
lift > 2)
> labels(rulesInNATA)
 [1] "{DULCES} => {AZUCAR}"
 [2] "{DULCES} => {TABLETA CHOCOLATE}"
 [3] "{AZUCAR,DULCES} => {LECHE}"
 [4] "{DULCES,LECHE} => {AZUCAR}"
 [5] "{DULCES,VERDURA} => {JAMON CURADO}"
 [6] "{DULCES,LECHE} => {TABLETA CHOCOLATE}"
 [7] "{DULCES,FRUTA} => {TABLETA CHOCOLATE}"
 [8] "{DULCES,VERDURA} => {TABLETA CHOCOLATE}"
 [9] "{DULCES,JAMON YORK} => {QUESO}"
[10] "{DULCES,QUESO} => {JAMON YORK}"
[11] "{DULCES,VERDURA} => {JAMON YORK}"
[12] "{DULCES,PASTAS BOLLERIA} => {LECHE}"
[13] "{DULCES,LECHE} => {PASTAS BOLLERIA}"
[14] "{DULCES,LECHE} => {CAFE}"
[15] "{DULCES,QUESO} => {CAFE}"
[16] "{DULCES,FRUTA} => {CAFE}"
[17] "{DULCES,VERDURA} => {CAFE}"
[18] "{DULCES,LECHE} => {CONSERVAS DE ATUN}"
[19] "{DULCES,YOGUR} => {PASTEL BOLLERIA}"
[20] "{DULCES,VERDURA,YOGUR} => {QUESO}"
[21] "{DULCES,QUESO,VERDURA} => {YOGUR}"
> rulesInNATA <- subset(rules, subset = rhs %in% "DULCES" &
lift > 2)
> labels(rulesInNATA)
 [1] "{AZUCAR} => {DULCES}"
 [2] "{TABLETA CHOCOLATE} => {DULCES}"
 [3] "{AZUCAR,LECHE} => {DULCES}"
 [4] "{LECHE,TABLETA CHOCOLATE} => {DULCES}"
 [5] "{FRUTA,TABLETA CHOCOLATE} => {DULCES}"
 [6] "{TABLETA CHOCOLATE,VERDURA} => {DULCES}"
 [7] "{LECHE,PASTAS BOLLERIA} => {DULCES}"
 [8] "{CAFE,LECHE} => {DULCES}"
 [9] "{CAFE,QUESO} => {DULCES}"
[10] "{CAFE,VERDURA} => {DULCES}"
[11] "{CONSERVAS DE ATUN,LECHE} => {DULCES}"
```

```
[12] "{PASTEL BOLLERIA,YOGUR} => {DULCES}"
[13] "{QUESO,VERDURA,YOGUR} => {DULCES}"
```

Veiem un patró curiós en aquestes regles pel que fa a la simetria que hem esmentat, ja que en les regles que inclouen **dolços** en la part esquerra hi apareixen dues referències amb **pernil dolç** i una amb **pernil curat**. Per tant, sembla que hi ha una pauta entre els compradors de galetes **dolces**, atès que també solen comprar **pernil dolç** i **curat**. No obstant això, no passa tan sovint que quan algú compra aquests dos tipus de productes també hi inclogui **galetes dolces**, ja que entre les regles rellevants amb **dolços**, en el costat dret de la regla no observem en cap cas **pernil dolç** ni **pernil curat**.

Amb aquestes regles podríem elaborar, per tant, un sistema de recomanació, en què a mesura que l'usuari anés afegint productes a la cistella anessin apareixent en la part inferior o lateral de la pantalla productes de la categoria que apareix a la dreta de les nostres regles d'associació.

> Per exemple, si un usuari del nostre web de comerç electrònic afegís un flam a la cistella, podríem recomanar-li natilles; i si usuari afegís una beguda de taronja amb gas, podríem recomanar-li que comprés a continuació una beguda de cola; i finalment, si un comprador hagués afegit dolços i verdura, podríem recomanar-li pernil dolç. D'aquesta forma, facilitaríem a l'usuari la localització dels productes que probablement estan en la seva llista de compra i l'experiència de compra en línia resultaria més agradable i àgil per al consumidor.

2. Anàlisi d'opinions per a gestió del coneixement

A continuació veurem com el científic de dades d'aquest cas pot analitzar les opinions que els clients aboquen a les xarxes socials sobre la marca i els productes que el distribuïdor inclourà en l'assortiment en línia.

La mineria de textos és una de les branques de la lingüística computacional que tracta d'obtenir informació i coneixement a partir de conjunts de dades que en principi no tenen un ordre o no estan disposades en origen per a transmetre aquesta informació.

La mineria de textos comprèn tres activitats fonamentals:

1) Recuperació d'informació, és a dir, seleccionar els textos pertinents.
2) Extracció de la informació inclosa en aquests textos: fets, esdeveniments, dades clau, relacions entre aquestes, etc.
3) Realització de la mineria de dades per a trobar associacions entre les dades clau que s'han obtingut prèviament en els textos.

La mineria de textos s'ajuda d'altres tècniques com la categorització de text, la recuperació d'informació i el processament de llenguatge natural i l'aprenentatge automatitzat. És l'eina que ens faltava per a poder enllaçar el coneixement quantitatiu amb el qualitatiu. L'objectiu, per exemple, és poder desglossar què ha volgut dir un client quan ens ha deixat unes observacions anotades en un full de reclamacions, i com afecta aquest fet el negoci que genera aquest client i altres de similars.

En l'àmbit comercial, resulta interessant trobar patrons ocults de consum dels clients per a poder explorar nous horitzons. Així ma-

teix, predir el comportament d'un futur client, basant-se en les dades històriques de clients que han presentat el mateix perfil, ajuda a poder retenir-lo durant el màxim temps possible.

Aquest sistema de recuperació d'informació és capaç no solament de retornar objectes (com paraules clau, imatges, etc.) rellevants per a una consulta, sinó també d'inferir les actituds dels emissors en esmentar els objectes en els resultats de cerca. Aquestes actituds valoratives ens donen una informació molt rellevant del sentiment de l'emissor.

En empreses orientades al client, aquesta tècnica permet obtenir una informació basada en dades implícites després de la recol·lecció per mitjà del retorn (*feedback*) del client, avançar-nos a les necessitats dels clients, tenir una informació més completa sobre la valoració del servei i millorar la fidelització.

Atès que hi ha una gran quantitat d'informació textual, podem trobar coneixement a partir de dades textuals sense estructurar.

La mineria de textos constitueix una eina de gran utilitat, ja que al voltant d'un 80% de la informació de les organitzacions està emmagatzemada en forma de text no estructurat.

La mineria de dades desestructurades es nodreix de la informació continguda en fitxers de text, a internet, ja que gran part de la informació està desestructurada.

Les xarxes socials són, per descomptat, una font inesgotable d'obtenció de dades de diferents fonts, com ara llibres, piulades, opinions sobre articles, etc.

La neteja de dades també s'anomena *fase de preprocessament*. El preprocessament és una tasca molt important en la mineria de textos, el processament de llenguatge natural i la recuperació de la informació. Bàsicament consisteix a netejar o filtrar la informació, eliminant els elements que no ens aporten informació d'interès i deixant només els que seran valuosos en l'etapa posterior, perquè aquesta fase sigui al més productiva, eficaç i eficient possible.

En el cas de les xarxes socials, les fonts de text del qual són comentaris, poden provenir de Twitter. Aquests són els passos que se segueixen per a «netejar» la informació:

1) Les etiquetes (#) són útils en la cerca de les piulades que es volen analitzar, per la qual cosa el primer filtre es limitaria només a les etiquetes definides en les categories d'informació que s'ha d'analitzar.
2) La **tokenització** i detecció de paraules derivades, és a dir, la separació del text per paraules o *tokens*.
3) S'eliminen signes de puntuació que no aporten informació extra.
4) S'eliminen les *stop-words*, paraules que tampoc aporten informació rellevant: articles, preposicions, pronoms. Això contribuirà a un millor rendiment dels algorismes de processament.
5) Detecció de paraules derivades (*stemming*): s'analitza l'arrel de les paraules i si hi ha paraules repetides amb la mateixa arrel s'eliminen i se substitueixen per la paraula que millor s'adapta al significat de l'arrel o lexema.
6) Transformació de majúscules a minúscules.
7) Totes les piulades preprocessades d'aquesta manera es converteixen en un únic corpus; un conjunt gran i estructurat de textos, que serà objecte de processament en les anàlisis d'opinió.

L'anàlisi d'opinió consisteix a classificar els comentaris dels clients en tres grups en funció del seu significat i de la implicació que tenen en l'anàlisi de resultats. Les polaritats de les opinions s'agrupen en tres categories: comentaris positius, negatius i neutres. Les **opinions positives** són les que usem per a descriure alguna cosa que desitgem de bon grat, un estat de felicitat o alegria, mentre que les **negatives** les utilitzem en àmbits o escenaris circumstancials que no són desitjats per nosaltres.

Segons si una paraula determinada evoca un sentiment positiu, negatiu o neutre, es construeix un diccionari lexicogràfic o lèxic. Aquest **lèxic** pot contenir totes les paraules que se'ns ocorrin (ja hi ha lèxics construïts), classificades segons la polaritat que els assignem. Quan el corpus ja preprocessat és analitzat i «comparat» amb

el lèxic que usem com a base, mitjançant diferents tècniques com la puntuació positiva o negativa i diferents algorismes, podem obtenir un valor que ens indica si el comentari expressa una opinió positiva, negativa o neutra. El resultat de l'anàlisi ens donarà la informació necessària perquè les empreses prenguin les decisions més encertades per a aconseguir la millor relació i valoració dels seus clients.

Quan l'algorisme compara dades amb bases de dades de paraules amb polaritat, utilitza el que coneixem com a *ontologies*. Hi ha dos grans grups d'ontologies, les genèriques i les específiques de cada indústria.

- **Ontologia genèrica**: representa conceptes generals que no són específics d'un domini. Per exemple, ontologies sobre el temps, ontologies de conducta, de causalitat, etc. Es poden reutilitzar en diferents dominis. Aquestes ontologies recullen paraules de tot el vocabulari sense diferenciar per categoria i s'agrupen en paraules positives i paraules negatives.

- **Ontologia específica de cada indústria**: les indústries que tenen com a base l'explotació de la informació i del coneixement. Aquest tipus d'ontologia explicita els conceptes, les propietats i les relacions existents pròpies del domini industrial.

  Per exemple, una ontologia sobre telèfons mòbils tindrà una recopilació de termes en què el concepte *reduït* pot tenir un sentit positiu si el terme es refereix a atributs del disseny, ja que en la indústria dels telèfons mòbils un disseny compacte és una cosa positiva. No obstant això, el mateix terme pot tenir una connotació negativa si es refereix als atributs d'un diamant, ja que un diamant de mida reduïda té una connotació negativa.

L'aplicació sobre el màrqueting de productes i serveis és evident. Podem conèixer el grau de satisfacció dels nostres clients (i així saber si hem de millorar o mantenir-nos en la mateixa línia), i també què és el que a aquests els agradaria que els oferís una organització concreta. Això implica fer anàlisis predictives sobre tendències i consums. Els avantatges, per tant, són obvis.

Abans de definir el terme *crawling*, hem de definir el concepte d'API.

> Un **API** és una funcionalitat d'una aplicació que permet que altres desenvolupadors utilitzin algunes de les seves característiques mitjançant llibreries, biblioteques o paquets de funcions desenvolupades amb aquesta finalitat.

Per a fer una anàlisi d'opinions, cal utilitzar alguna eina de *crawling* per a recollir les piulades dels usuaris de forma ràpida i eficient. Twitter posa a disposició dels usuaris el servei **Twitter Search** i també un **API** que envia les piulades que després poden ser recollides per llibreries de Java, R i altres llenguatges de programació i *scripting*.

A continuació tractarem el cas de **TALC (Talcum Powder de Johnson & Johnson)**, un producte cosmètic amb efectes molt perjudicials per a la salut del consumidor que provocava, fins que es va retirar, des d'irritacions de pell fins a càncer d'ovari (figura 12).

El producte va generar una alarma social a causa d'un judici que va enfrontar la companyia i la família d'una jove que va morir, segons l'acusació, arran de l'ús del producte. Per aquest motiu, J&J va haver de pagar una compensació econòmica de 72 milions de dòlars. Setze estudis fets el 2003 van demostrar que l'ús del producte incrementava el risc de patir càncer d'ovari tres vegades més que l'habitual.

Figura 12. Talcum Powder

En l'anàlisi recollirem dades mitjançant R i l'API de Twitter, i analitzarem el sentiment general sobre el producte. Aquesta anàlisi es pot automatitzar i dur-se a terme per a tot el dossier de referències d'una empresa de distribució de comerç electrònic, semblant a Amazon, amb la finalitat de detectar productes impopulars que els directors de producte haurien d'eliminar ràpidament de l'assortiment per a desvincular la imatge de l'empresa de productes perjudicials per a la salut i amb mala acceptació.

Per a fer ús de l'API de Twitter seguirem els passos que indiquem a continuació. L'objectiu d'aquesta primera fase és configurar un compte per a establir les contrasenyes que permetran la comunicació entre l'API i les nostres llibreries de R.

Twitter Search té certes limitacions si el que volem és recollir piulades de forma dinàmica. Per a importar piulades relacionades amb una paraula d'interès, primer s'ha de crear aplicació. Cliqueu l'enllaç següent, TwitterApps, o bé escriviu directament «apps.twitter.com» a l'explorador.

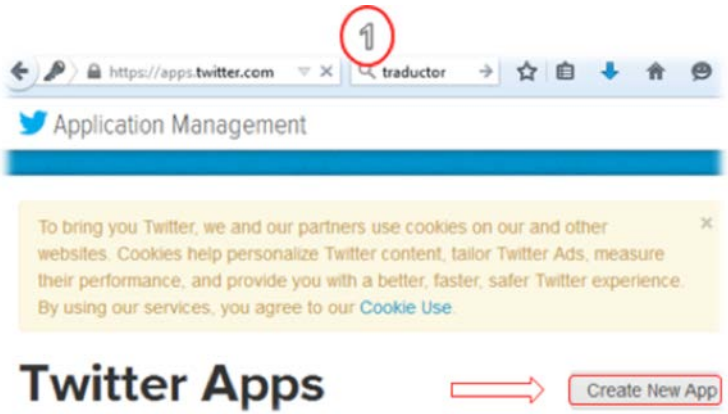
Cal disposar d'un usuari de Twitter per a crear una aplicació (figura 13).

1) Des de Twitter Apps, feu clic a «sign in».
2) A continuació, introduïu el vostre usuari i la vostra contrasenya.

Figura 13. Registre

Una vegada oberta la sessió feu clic a «Crear nova aplicació» (figura 14).

Figura 14. Creació d'app

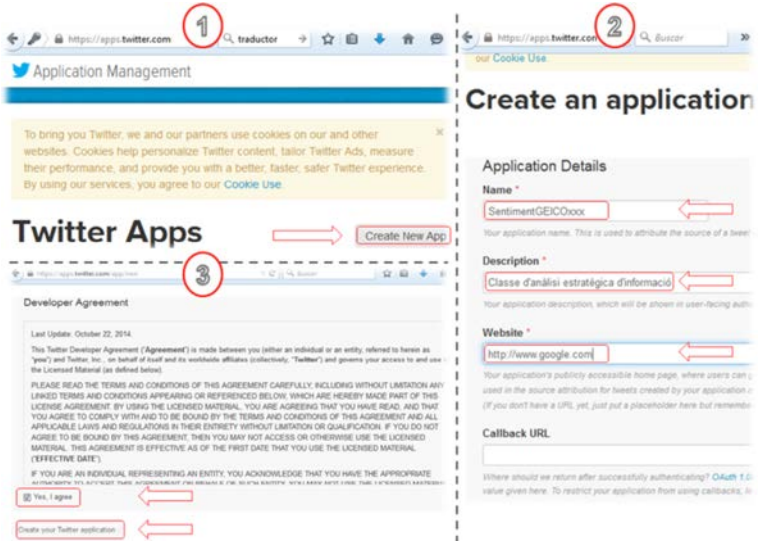


1) A continuació, introduïu la informació següent (figura 15):

- **Nom de l'aplicació**. No ha d'existir prèviament. Per tant, si introduïu un nom i Twitter us adverteix que ja existeix, n'introduirem un altre, per exemple: app001, app002, app003...
- **Descripció**. Qualsevol descripció, per exemple: «Classe d'AEI».
- **Lloc web**. Afegiu-hi un domini vàlid, per exemple: http://www.google.com. És important no oblidar afegir-hi «http://».
- **Callback URL**. No cal que ho empleneu.

2) Feu clic a «Yes, I agree».
3) Per acabar feu clic a «Create your Twitter application». La vostra aplicació ja s'ha creat.

Figura 15. Detalls de l'aplicació



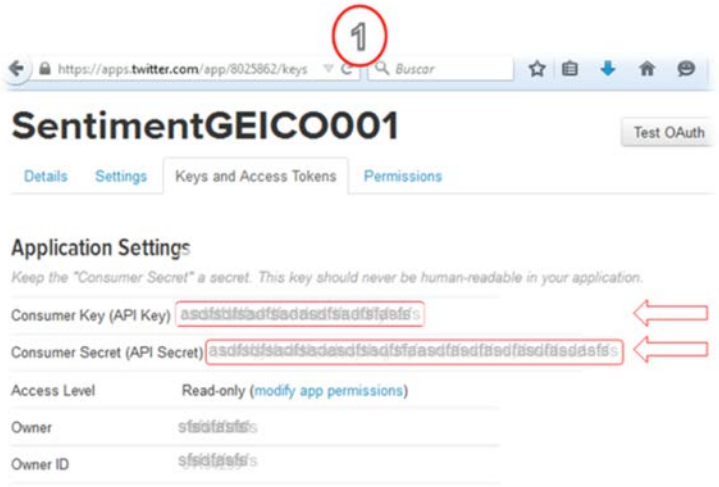
És possible que Twitter us demani que registreu el vostre telèfon. Seguiu les instruccions de l'enllaç següent i el podreu afegir molt ràpidament: Telèfon. Sempre el podreu desvincular després de llegir aquesta guia.

És possible que Twitter no us enviï el codi si el vostre operador és estranger. Cal tenir un operador admès per Twitter.

Seguidament deseu la informació següent de l'apartat «Key and Access Tokens», que us serà útil més endavant (figura 16):

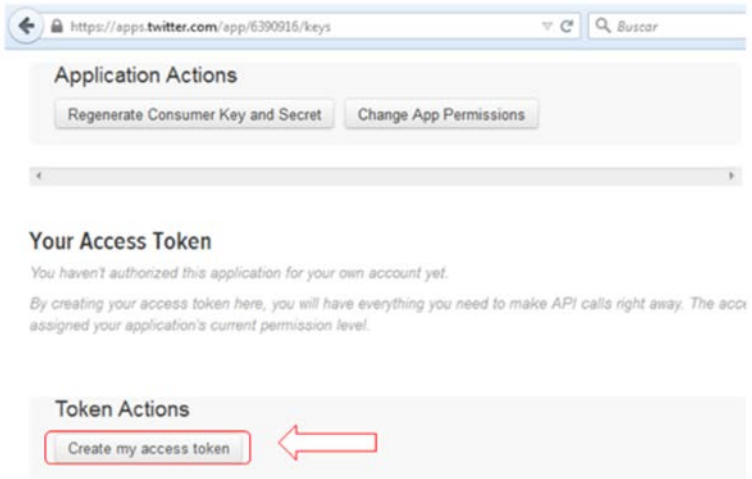
- Consumer Key
- Consumer Secret

Figura 16. Configuració



Més avall veurem l'opció «Token Actions» i «Create my access token» (figura 17).

Figura 17. *Tokens*

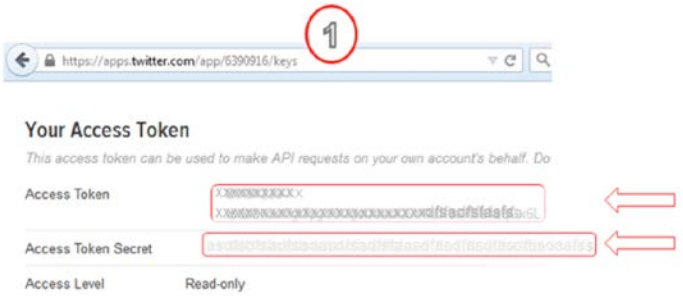


Feu clic a «Create my access token». Això generarà dos codis addicionals que trobarem a la part inferior, amb els noms següents (figura 18):

- Access Token
- Access Token Secret

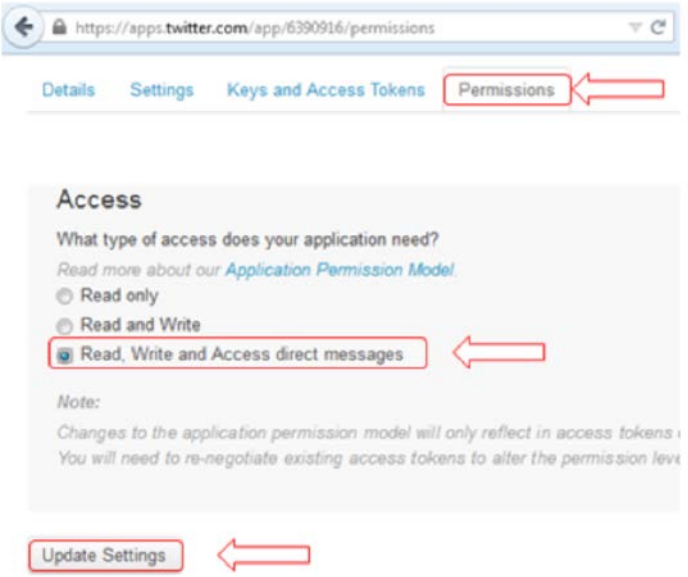
Preneu nota també d'aquests codis que utilitzareu més endavant.

Figura 18. *Tokens* d'accés



Finalment, doneu accés de lectura i escriptura a l'aplicació, anant en l'apartat «Permisions», on heu d'escollir «Read, Write Access Direct Messages» i per acabar feu clic a «Update Settings» (figura 19).

Figura 19. Actualitzar



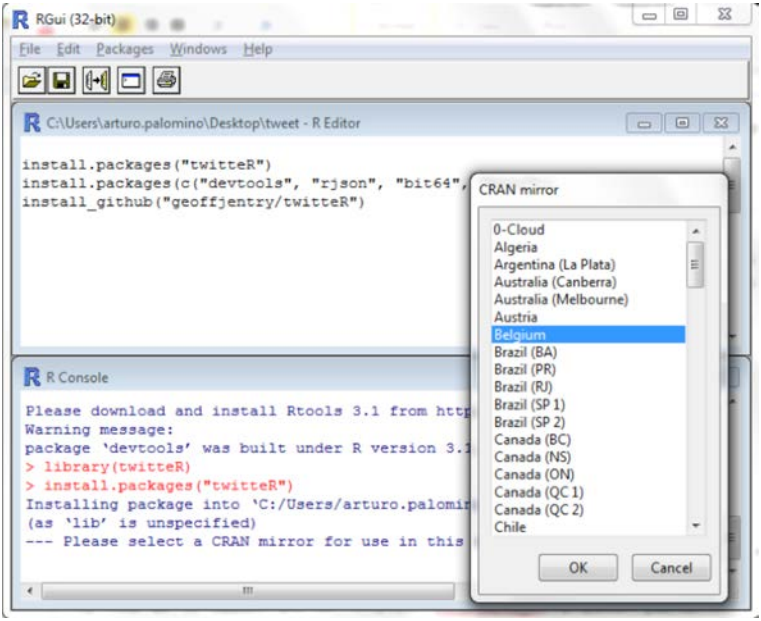
Per a recollir piulades amb R introduïu les ordres següents a l'editor.

```
install.packages("devtools")
install.packages("twitteR")
```

Executeu el codi seleccionant les línies i fent «Control + R». Assegureu-vos que teniu connexió a internet abans d'executar.

L'ordre «install.packages» busca les llibreries en un servidor extern. Trieu un servidor proper i feu clic a «OK». Això pot trigar uns minuts. Haurà acabat quan veiem a la consola el caràcter «>» (figura 20).

Figura 20. Instal·lació



Una vegada instal·lat el paquet «twitteR», reinicieu R. Podeu desar els *scripts* i en reiniciar, obrir-los amb l'opció de menú: File > «open script». No cal guardar sessions, només els *scripts*.

Després d'executar les ordres «install.packages», els paquets queden guardats en R i no cal tornar-los a executar en cada sessió. Per a no

tornar a llançar-les per error, deixarem les línies «comentades» posant el caràcter «#» al davant:

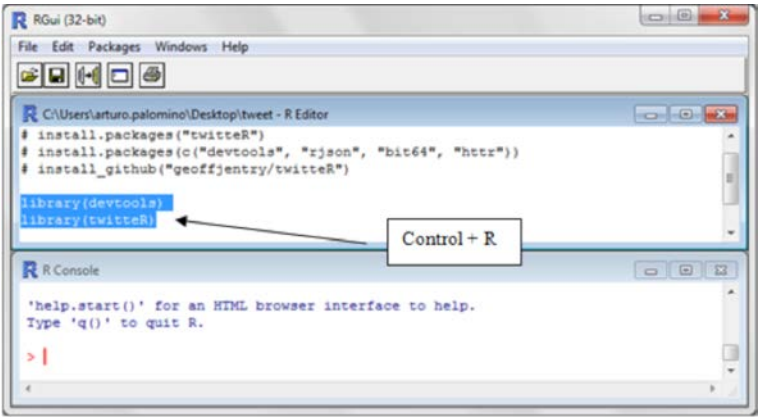
```
# install.packages("devtools")
# install.packages("twitteR")
```

A continuació, després de reiniciar R i tornar a l'editor, escriviu les ordres següents:

```
library(devtools)
library(twitteR)
```

Executeu seleccionant les dues línies alhora i fent «Control + R» (figura 21).

Figura 21. Llibreries



A continuació, introduïu entre les cometes («xxx») els codis que hem guardat de TwitterApps (Consumer Key, Consumer Secret, Access Token i Access Token Secret). Seleccioneu les línies i feu «Control + R».

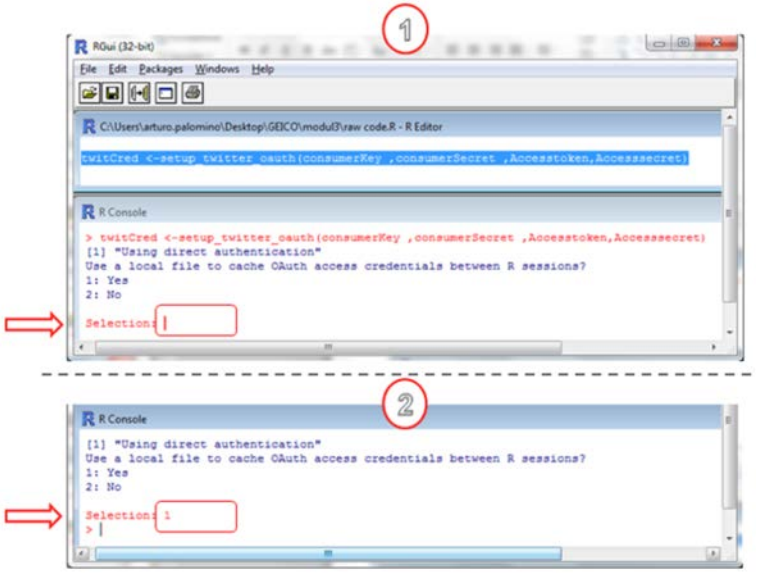
```
consumerKey <- "XXX"
consumerSecret <- "XXX"
Accesstoken <- "XXX"
Accesssecret <- "XXX"
```

Afegiu i executeu la funció següent, que farà la connexió amb la nostra aplicació. Feu «Control + R».

```
twitCred <-setup_twitter_oauth(consumerKey ,consumerSecret
,Accesstoken,Accesssecret)
```

La consola us demanarà que introduïu una opció; trieu 1. Escriviu 1 i premeu «Enter» (figura 22).

Figura 22. Autenticació



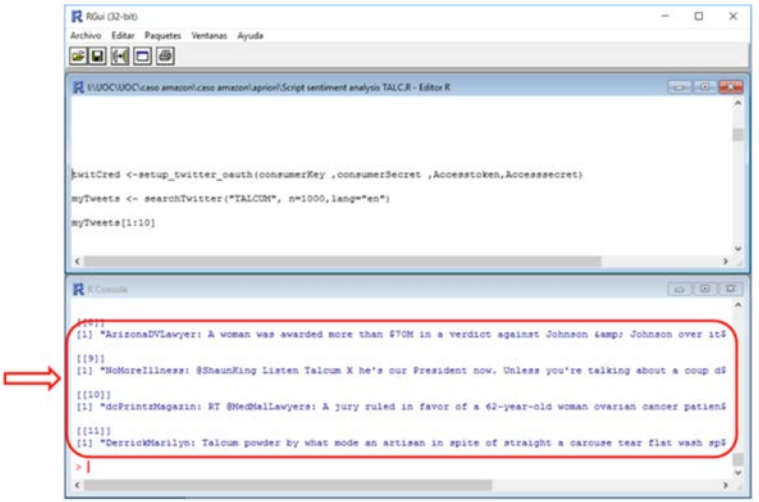
Per a recollir tots les piulades que incloguin la paraula «TALC», per exemple, utilitzarem l'ordre següent. Amb n=1000 limitem el nombre de piulades a 1.000. D'altra banda, triem l'idioma anglès amb la variable lang=«en». L'ordre exacta que s'ha d'executar és la següent:

```
myTweets <- searchTwitter("TALCUM", n=1000,lang="en")
```

Per a veure les primeres deu piulades farem servir la línia de codi següent, que haurem d'executar (figura 23).

```
myTweets[1:10]
```

Figura 23. Revisió

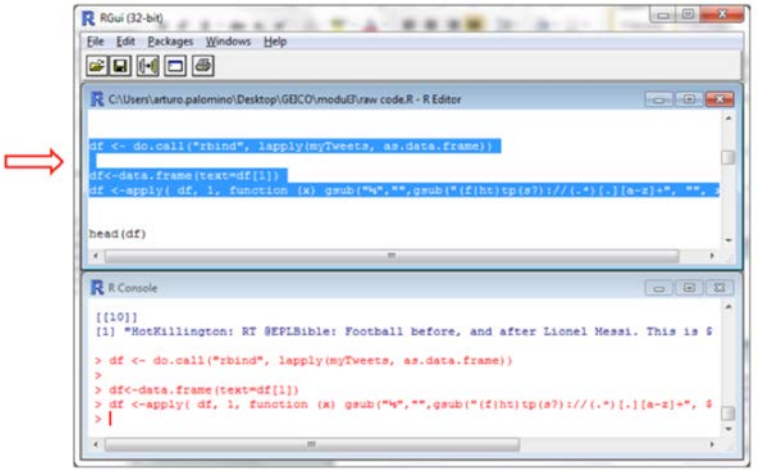


A continuació, ens quedarem amb les piulades del *crawling* desant-les en un *dataframe* denominat «df» (figura 24):[3]

```
df <- do.call("rbind", lapply(myTweets, as.data.frame))
df<-data.frame(text=df[1])
df <-apply( df, 1, function (x) iconv(gsub("(f|ht)tp(s?)://
(.*)[.][a-z]+", "", x), "latin1", "ASCII", sub="") )
```

3 Per a més informació sobre *dataframes*, reviseu el capítol 2.2.4 de Castillo, A. J. S. «Métodos Estadísticos con R y R Commander». [Recurso electrónico gratuito] <http://cran.r-project.org/doc/contrib/Saez-Castillo-RRCmdrv21.pdf>

Figura 24. *Dataframe*



Ara baixeu una llibreria de mineria de textos per a la part del preprocessament de les piulades.

```
install.packages("tm")
library(tm)
```

A continuació, per a fer el preprocessament, crearem el corpus.[4]

Executeu la línia següent:

```
myCorpus <- Corpus(VectorSource(df))
```

4 Per a més informació, llegiu:
Feinerer, I. (2015). «Introduction to the tm Package Text Mining in R». 2013-12-01.
Liau, Bee Yee; Tan, Pei Pei (2014). «Gaining customer knowledge in low cost airlines through textmining». *Industrial Management & Data Systems* (vol. 114, núm. 9, pàg. 1344-1359).
Nguyen, Tung Thanh; Quan, Tho Thanh; Phan, Tuoi Thi (2014). «Sentiment search: an emerging trend on social media monitoring Systems». *Aslib Journal of Information Management* (vol. 66, núm. 5, pàg. 553-580).

Passeu a minúscula el text, elimineu-hi la puntuació, la numeració, les *stop-words*, els espais en blanc, etc.

```
myCorpus<- tm_map(myCorpus, content_transformer(tolower))
myCorpus <- tm_map(myCorpus , PlainTextDocument)
myCorpus <- tm_map(myCorpus, removePunctuation)
myCorpus <- tm_map(myCorpus, removeNumbers)
myStopwords <- c(stopwords('english'), "available", "via",
"http")
myCorpus <- tm_map(myCorpus, removeWords, c(myStopwords))
myCorpus <- tm_map(myCorpus , stripWhitespace)
```

Deseu una còpia del corpus.

```
dictCorpus <- myCorpus
```

Baixeu llibreries addicionals.

```
install.packages("SnowballC")
install.packages("RWeka")
install.packages("rJava")
install.packages("RWekajars")

library("SnowballC")
library("RWeka")
library("rJava")
library("RWekajars")
```

Conserveu l'arrel lèxica de les paraules fent *stemming*.

```
myCorpus <- tm_map(myCorpus, stemDocument)
```

Completeu els lexemes (pot trigar uns minuts).

```
myCorpus <- tm_map(myCorpus, stemCompletion,
dictionary=dictCorpus)
```

Netegeu el corpus.

```
dataframe<-data.frame(text=unlist(sapply(dictCorpus[1:1000]
[1:1000], `[`, "content")), stringsAsFactors=F)
myCorpus<-Corpus(VectorSource(dataframe$text))
```

Filtreu les paraules amb una freqüència mínima; en aquest cas freqüència 4 mínima.

```
myDtm <- TermDocumentMatrix(myCorpus, control =
list(minWordLength = 4))
```

Visualitzeu les paraules més freqüents per sobre d'un cert llindar i les més associades a «TALC».

```
findFreqTerms(myDtm, lowfreq=5)
findAssocs(myDtm, 'TALCUM', 0.10)
```

Canvieu el format de les dades a matriu.

```
m <- as.matrix(myDtm)
```

Per a visualitzar la informació instal·larem un paquet addicional.[5]

```
install.packages("wordcloud")
library(wordcloud)
```

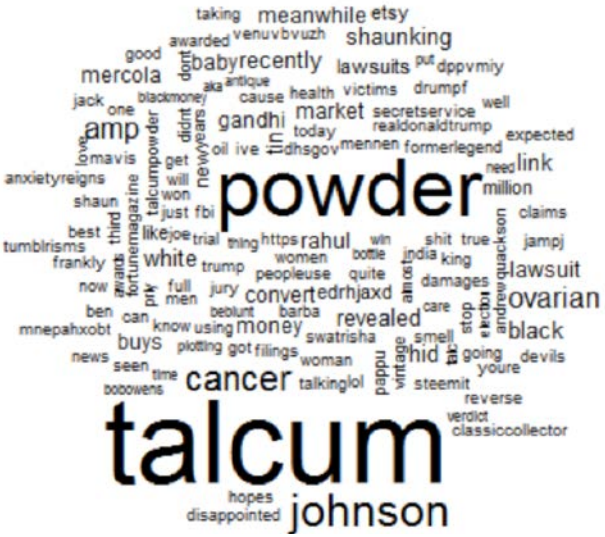
Farem un WordCloud per a veure el núvol de paraules que envolten «TALC» més habitualment a les piulades.

```
v <- sort(rowSums(m), decreasing=TRUE)
myNames <- names(v)
d <- data.frame(word=myNames, freq=v)
wordcloud(d$word, d$freq, min.freq=10)
```

5 Per a més informació, llegiu Fellows, I.; Fellows, M. I.; Rcpp, L. (2012). «Package "wordcloud'"». <http://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf>

El núvol de paraules ens permet observar d'una forma molt visual quines són les paraules que estan més associades amb el terme «TALC» en les 1.000 piulades (figura 25).

Figura 25. Núvol de paraules



Font: elaboració pròpia

Podem veure paraules com *Johnson*, *veredicte*, *legalitat*, *danys*, *càncer*, *decepció*, *dona*, *càncer d'ovari*, que fan al·lusió a l'alarma social que va generar el producte i la seva relació amb aquesta malaltia mortal.

Per a tenir un indicador de sentiment o opinió general de les piulades recollides, carregarem unes taules de paraules subjectives de diccionari.[6]

6 Per a més informació, llegiu:
Liau, Bee Yee; Tan, Pei Pei (2014). «Gaining customer knowledge in low cost airlines through textmining». *Industrial Management & Data Systems* (vol. 114, núm. 9, pàg. 1344-1359).
Nguyen, Tung Thanh; Quan, Tho Thanh; Phan, Tuoi Thi (2014). «Sentiment search: an emerging trend on social media monitoring Systems». *Aslib Journal of Information Management* (vol. 66, núm. 5, pàg. 553-580).

Copieu els dos fitxers d'aquest document, «**positive-words.txt**» i «**negative-words.txt**», en una ruta que recordeu fàcilment.

A continuació, copieu en els *scripts* les línies següents, que haureu de modificar, substituint les rutes per aquella en què hem desat els dos fitxers adjunts.

```
pw<-read.table("C:/RUTA DEL FITXER/positive-words.
txt",sep=";", stringsAsFactors = F)
nw<- read.table("C:/ RUTA DEL FITXER/negative-words.txt",
stringsAsFactors = F)
```

Haureu de substituir la ruta «C:/RUTA DEL FITXER/» per la ruta on es troben els dos fitxers que heu copiat. Seguidament executeu les dues línies. Vigileu de no confondre «/» amb «\».

A continuació, copieu la funció següent a l'script i executeu-la.

```
sentimen<- function(x){
   sentiment = 0
   palabras = 0
   NEGATIVAS=0
   POSITIVAS=0
   ratio1=0
   ratio0=0
   for (i in 1:length(myCorpus)) {#i<-1
       doc <- myCorpus[[i]]
       vw <- strsplit(doc[[1]],' ')
       s = 0
       P=0
       N=0
       for (w in vw[[1]]) { #w<-vw[[1]][6]  w<-"unbeatable"
         palabras<-palabras+1
           if (length(which(w%in%pw[[1]])) > 0) {
```

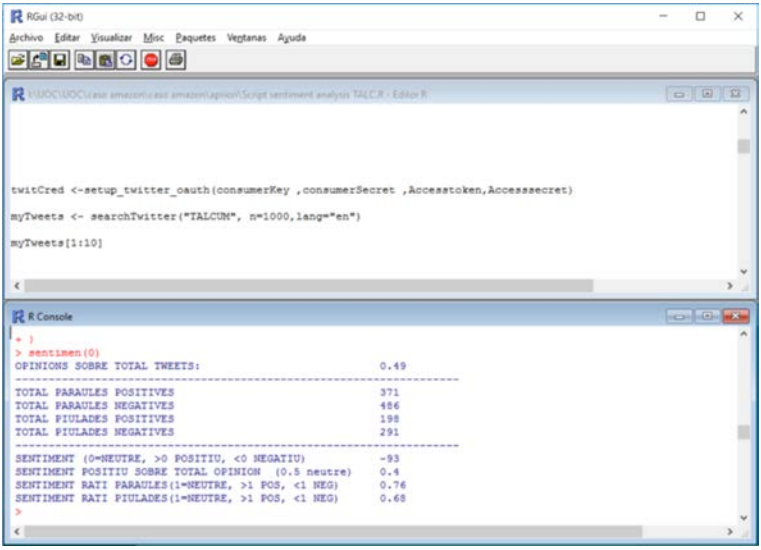
Dinsoreanu, Mihaela; Potolea, Rodica (2014). «Opinion driven communities' detection». *International Journal of Web Information Systems* (vol. 10, núm. 4, pàg. 324-342).

```
                s = s + 1
            P=P+1
        }
        if (length(which(w%in%nw[[1]])) > 0) {
            s = s -1
            N=N+1
        }
    }
    sentiment = sentiment + min(max(-1,s),1)
    if (s>0) {ratio1=ratio1+1}
    if (s<0) {ratio0=ratio0+1}
    NEGATIVAS=NEGATIVAS+N
    POSITIVAS=POSITIVAS+P
}
cat(paste("OPINIONS SOBRE TOTAL TWEETS:
", round((ratio0+ratio1)/length(myCorpus),2)))
cat("\n")
cat(paste ("-----------------------------------------
----------------------"))
cat("\n")
cat(paste("TOTAL PARAULES POSITIVES
", POSITIVAS))
cat("\n")
cat(paste("TOTAL PARAULES NEGATIVES
", NEGATIVAS))
cat("\n")
cat(paste("TOTAL PIULADES POSITIVES
", ratio1))
cat("\n")
cat(paste("TOTAL PIULADES NEGATIVES
", ratio0))
cat("\n")
cat(paste ("-----------------------------------------
----------------------"))
cat("\n")
cat(paste("SENTIMENT (0=NEUTRE, >0 POSITIU, <0 NEGATIU)
", sentiment))
cat("\n")
```

```
   cat(paste("SENTIMENT POSITIU SOBRE TOTAL OPINION  (0.5
neutre)   ", round(ratio1/(ratio0+ratio1),2) ))
   cat("\n")
   cat(paste("SENTIMENT RATI PARAULES(1=NEUTRE, >1 POS, <1
NEG)     ", round(POSITIVAS/ ifelse(NEGATIVAS==0,0.001,NEGA
TIVAS),2 ) ))
   cat("\n")
   cat(paste("SENTIMENT RATI PIULADES(1=NEUTRE, >1 POS, <1
NEG)     ", round(ratio1/ ifelse(ratio0==0,0.001,ratio0),2
) ))
   cat("\n")
}
sentimen(0)
```

El resultat serà similar al següent (figura 26). No coincidirà exactament, atès que depèn del moment d'execució:

Figura 26. Resultat



El que indica aquest resultat (figura 26) és el següent:

- **OPINIONS SOBRE EL TOTAL DE PIULADES**: indica el nombre de piulades que expressen una opinió no neutra (positiva o negativa). En aquest cas el 49% de les piulades expressaven una opinió en un o altre sentit (polaritat).
- **TOTAL PARAULES POSITIVES**: Indica el nombre de paraules que expressen una opinió positiva. En aquest cas s'han trobat 371 paraules positives del total.
- **TOTAL PARAULES NEGATIVES**: Indica el nombre de paraules que expressen una opinió negativa. En aquest cas s'han trobat 486 paraules negatives del total.
- **TOTAL PIULADES POSITIVES**: Indica el nombre de piulades en què el nombre de paraules positives per piulada superava el nombre de paraules negatives. En aquest cas s'han trobat 198 piulades amb sentit positiu.
- **TOTAL PIULADES NEGATIVES**: Indica el nombre de piulades en què el nombre de paraules negatives per piulada superava el nombre de paraules positives. En aquest cas s'han trobat 291 amb sentit negatiu.
- **SENTIMENT**: Indica el nombre de piulades positives, menys el nombre de piulades negatives. Així, si el resultat és igual a 0, el sentiment general és neutre; si el resultat és superior a 0, llavors el sentiment general és positiu; i negatiu en l'altre cas.
- **SENTIMENT SOBRE EL TOTAL D'OPINIONS**: Expressa la ràtio de nombre de piulades positives entre el total de piulades que expressen opinió positiva o negativa. Els valors inferiors a 0,5 indiquen opinió negativa general; els valors superiors a 0,5 indiquen opinió positiva general.
- **SENTIMENT, RÀTIO DE PARAULES**: Expressa la ràtio de nombre de paraules positives entre el total de paraules negatives. Una ràtio superior a 1 indica opinió general positiva, mentre que una ràtio inferior a 1 indica opinió general negativa.
- **SENTIMENT, RÀTIO DE PIULADES**: Expressa la ràtio de nombre de piulades positives entre el total de piulades negatives. Una ràtio superior a 1 indica opinió general positiva, mentre que un ràtio inferior a 1 indica opinió general negativa.

L'anàlisi ens revela clarament que aquest producte no està ben vist pel consumidor. Davant d'aquest resultat la decisió del distribuïdor

ha de ser evitar-ne la venda, tenint en compte que no solament perjudicarà la seva imatge i comportarà menys vendes, sinó que, en aquest cas tan extrem, també posarà en risc la salut dels clients.

Bibliografia

Castillo, A. J. S. «Métodos Estadísticos con R y R Commander». [Recurso electrónico gratuito] <http://cran.r-project.org/doc/contrib/Saez-Castillo-RRCmdrv21.pdf>

Dinsoreanu, M.; Potolea, R. (2014). «Opinion driven communities' detection». *International Journal of Web Information Systems* (vol. 10, núm. 4, pàg. 324-342).

Feinerer, I. (2015). «Introduction to the tm Package Text Mining in R». 2013-12-01.

Fellows, I.; Fellows, M. I.; Rcpp, L. (2012). «Package "wordcloud'"». <http://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf>

Liau, B. Y.; Tan, P. P. (2014). «Gaining customer knowledge in low cost airlines through text mining». *Industrial Management & Data Systems* (vol. 114, núm. 9, pàg. 1344-1359).

Nguyen, T. T.; Quan, T. T.; Phan, T. T. (2014). «Sentiment search: an emerging trend on social media monitoring Systems». *Aslib Journal of Information Management* (vol. 66, núm. 5, pàg. 553-580).

Santana, J. S.; Farfán, E. M. «El arte de programar en R: un lenguaje para la estadística». <http://cran.r-project.org/doc/contrib/Santana_El_arte_de_programar_en_R.pdf>

Apartat Estadística

Baró Llinàs, J.; Alemany Leira, R. (2002). «Anàlisi cluster». *Estadística II*. Barcelona: Universitat Oberta de Catalunya.

Baró Llinàs, J.; Alemany Leira, R. (2002). «Sèries temporals». *Estadística II*. Barcelona: Universitat Oberta de Catalunya.

Baró Llinàs, J.; Alemany Leira, R. (2008). «El model de regressió múltiple». *Estadística II* (4a. ed.). Barcelona: Universitat Oberta de Catalunya.

Gibergans Bàguena, J.; Gil Estallo, A. J.; Rovira Escofet, C. (2009). «Anàlisi de la variància». *Estadística* (4a. ed.). Barcelona: Universitat Oberta de Catalunya.

Gibergans Bàguena, J.; Gil Estallo, A. J.; Rovira Escofet, C. (2009). «Combinatòria i tècniques de recompte». *Estadística* (4a. ed.). Barcelona: Universitat Oberta de Catalunya.

Greenacre, M. M. (2008). «Inferència estadística: part 4 i 5». *Estadística I* (3a. ed.). Barcelona: Universitat Oberta de Catalunya.

Salvador Figueras, M. (2003). «Análisis de correspondencias».

Què és H2PAC?

El model H2PAC resol propostes clau a partir d'ACTIVITATS.

Aquesta forma d'aprenentatge parteix d'un REPTE: l'activitat que hauràs de resoldre. Per això et facilitem un contingut teòric, EL CONEIXEMENT IMPRESCINDIBLE, que t'ajudarà a entendre els conceptes essencials per a poder afrontar el desafiament plantejat inicialment.

A més del contingut teòric, el model també et facilita LES SOLUCIONS, una proposta de resolució del repte exposat.

El repte d'aquesta obra és l'aplicació de l'anàlisi d'informació com a eina estratègica en un context empresarial. L'obra fa un recull de tècniques d'anàlisi, i en planteja exemples contextualitzats i aplicats.

H2PAC

Una iniciativa universitària d'**Oberta Publishing** per a gaudir i aprendre

Universitat Oberta de Catalunya

PID_00242821