TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Anàlisi de correspondències pel tractament de dades textuals: aplicació sobre les descripcions dels projectes presentats a EIT Health

Autor: Arnau Reynals Rodríguez

Director: Dr. Belchin Adriyanov Kostov

Departament: Departament d'Estadística i Investigació Operativa

Convocatòria: Q2 2020/2021

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

RESUM

El ritme accelerat de la informàtica en els últims vint anys ha permès treballar amb grans volums de dades que anteriorment no era possible. Una de les àrees que en genera més és la de dades textuals. En els darrers anys, aquest camp ha guanyat un pes molt important. Malgrat això, la falta de coneixement sobre les metodologies ha limitat la seva utilització i extracció d'informació. En aquest sentit, aquest treball final de grau pretén demostrar que mètodes d'anàlisi factorial com Anàlisi de Correspondències i les seves extensions poden ser de gran utilitat a l'hora de tractar amb dades textuals determinant temaris, classificant els objectes i descrivint les seves característiques. Per portar a terme aquest objectiu, es treballarà sobre una base de dades de l'institut EIT Health, una institució de la Unió Europea que dóna suport i finançament a diferents projectes i startups en l'entorn de la salut.

PARAULES CLAUS: ANÀLISI DE DADES TEXTUALS, TAULES DE CONTINGÈNCIA, ANÀLISI DE CORRESPONDÈNCIES, CLÚSTER JERÀRQUIC, TEXT MINING

CLASSIFIACIÓ AMS

68T50 Natural language processing

62H17 Contingency tables

62H25 Factor analysis and principal components; correspondence analysis

62H30 Classification and discrimination; cluster analysis

ABSTRACT

The rapid pace of IT over the last twenty years has made it possible to work with large volumes of data that were previously impossible. One of the areas that generates the most is textual data. In recent years, this field has gained significant weight. Nevertheless, the lack of knowledge about methodologies has limited its use and information extraction. In this sense, this final thesis aims to demonstrate that factor analysis methods such as Correspondence Analysis and its extensions can be very useful when dealing with textual data determining themes, classifying objects and describing their characteristics. To achieve this goal, we will work on a database of the EIT Health institute, a European Union institution that supports and finances different projects and startups in the field of health.

KEY WORDS: TEXTUAL DATA ANALYSIS, CONTINGENCY TABLE, CORRESPONDENCE ANALYSIS, CLUSTER ANALYSIS, TEXT MINING

AMS CLASSIFICATION

68T50 Natural language processing

62H17 Contingency tables

62H25 Factor analysis and principal components; correspondence analysis

62H30 Classification and discrimination; cluster analysis

ÍNDEX

1 INTRODUCCIÓ

El ritme accelerat de la informàtica en els darrers vint anys ha permès generar i treballar amb grans volums de dades que anteriorment no era possible. Una de les àrees on se'n generen més és en les de tipus textual, ja siguin preguntes obertes, discursos, llibres, xarxes socials (tweets, missatges, etc.).

Amb l'augment del nombre de dades textuals generades, aquest camp ha guanyat un pes molt important. La falta de coneixement sobre les metodologies i com tractar-les ha limitat l'extracció d'informació d'aquestes. Tot i això, en els últims anys s'han desenvolupat un seguit de tècniques, englobades sota el terme de *Text Mining*, que permet extreure informació a partir de textos. Alguns exemples poden ser: l'anàlisi de freqüència, l'agrupació de paraules, l'anàlisi de sentiments o el Processament del Llenguatge Natural (NPL en angles). Aquesta última tècnica utilitza Arbres de Decisió, Xarxes Neuronals i Regressions Logístiques entre altres metodologies per treure *insights* a partir de text.

L'objectiu d'aquest treball de final de grau és demostrar que l'Anàlisi de Correspondències (CA) combinat amb mètodes de classificació jeràrquics també poden ser de gran utilitat a l'hora de tractar amb dades textuals determinant temaris, classificant els objectes i descrivint les seves característiques.

Per portar a terme aquest, es treballarà sobre la base de dades de l'institut EIT Health (EIT Health), una institució de la Unió Europea que dóna suport i finançament a diferents projectes i empreses emergents en l'entorn de la salut.

Com a variable principal s'utilitzarà les descripcions de les propostes dels projectes presentats a diferents convocatòries d'EIT Health i com a variables secundàries es consideraran: el país on es desenvoluparà el projecte, el programa de EIT Health on han aplicat, el nivell tecnològic de la companyia (TRL), si van ser seleccionades o no, el pressupost del qual disposen i el gènere de la persona que va presentar la proposta.

Com a objectius secundaris es vol agrupar diferents projectes de EIT Health sota característiques comunes i tipologia similar, fet que pot ajudar a l'organització a impulsar diferents estratègies per tal de maximitzar l'eficiència dels recursos disponibles. Aquest estudi servirà per veure si certes regions tenen més interès per temes concrets i també com es poden agrupar els diferents projectes en funció de la temàtica que tractin.

El treball es divideix en tres blocs: un primer on s'exposen els materials i la metodologia utilitzada, on es parla sobre l'origen de la base de dades, com es construeix la matriu i els càlculs matemàtics necessaris per a l'anàlisi. A continuació, s'explica el desenvolupament de l'estudi amb l'obtenció del glossari, *metakeys*, clústers i la seva descripció detallada. Finalment, un últim apartat on es presenten les conclusions.

Aquest treball m'ha permès conèixer en profunditat un món nou, el de les dades textuals, que durant aquests quatre anys de carrera no se n'ha parlat massa, en un futur pot ser un dels més interessants i d'on es pugui extreure més informació. A més, el fet de poder realitzar l'estudi amb la base de dades de EIT Health, empresa en la qual he estat un any treballant, em suposa un repte afegit.

2 MATERIALS I MÈTODES

En aquest capítol s'explicarà quin és l'origen de la base de dades que ha permès obtenir el Corpus Lèxic per la posterior anàlisi de correspondències, s'entén com a corpus el conjunt d'ocurrències (paraules) en un determinat text. Per començar s'explicarà com funciona EIT Health i la seva manera de treballar, tot seguit el pre-procés de neteja de les dades i el filtratge final per obtenir el corpus lèxic organitzat de la manera següent: *projectes x paraules.*

2.1 EIT HEALTH

EIT (*European Institute of Innovation and Technology*), creat el 2008, és una iniciativa de la Unió Europea, emmirallada en el MIT de Boston, que vol fomentar la innovació i l'emprenedoria arreu d'Europa amb una simple idea "*through diversity there is strength*". EIT es divideix en sis branques diferents: Climate-KIC, Digital, InnoEnergy, Health, Raw Materials i Food.

EIT Health neix a partir que científics d'arreu d'Europa s'adonessin que des de fa cent cinquanta anys cada generació augmenta l'esperança de vida uns cinc anys, però per contra, això fa que la despesa en cures hagi augmentat fins a nivells molt elevats, els últims estudis calculen que es gasten uns 115 bilions d'euros a l'any per aquests motius[1]. Per això es vol invertir esforços, diners i coneixement en intentar prevenir aquestes malalties i alhora treballar per descobrir i perfeccionar tècniques de tractament. El seu objectiu final és augmentar al màxim la qualitat de vida de les persones.

Tal com s'ha comentat en la introducció EIT Health dona finançament, acompanyament i assessorament a diferents projectes de salut, alguns dels més destacats són: Biel Glasses, un projecte català d'ulleres intel·ligent per ajudar a les persones amb problemes de visió. DYNSEO - Brain games apps for all, una empresa francesa que ha desenvolupat una sèrie d'aplicacions mòbils per estimular l'activitat cognitiva de la gent gran i prevenir trastorns cognitius.

EIT Health té la seu central a Munic, Alemanya però amb representació a altres països a través de sis centres d'innovació regionals anomenats CLC (*co-location centres*): *EIT Belgium-Netherlands*, *EIT France, EIT Germany, EIT Scandinavia, EIT Spain* i *EIT Ireland-Uk* a més a més tenen un centre anomenat *Innostar* que engloba les regions no cobertes amb els sis centres principals.

Per tal de donar finançament i assessorament als diferents projectes cada CLC ofereix una sèrie de programes (*EIT Health Catapult, Brigehead, GoldTrack...*) amb diferents característiques, fent així que els impulsors de cada candidatura triïn quin programa s'ajusta més a les necessitats de la seva proposta i omplin la sol·licitud d'inscripció.

La suma de les diferents sol·licituds a cada programa de EIT Health és el que forma la base de dades inicial. Aquestes sol·licituds contenen informació relativa al projecte presentat: el nom del projecte, el país, el nivell tecnològic, el programa, la possible inversió que han rebut fins al moment i un llarg de camps, entre tots ells el que més interessa per l'anàlisi és el de la descripció textual del projecte.

2.2 CORPUS

La base de dades inicial conté un total de 5.030 registres, totes les aplicacions a EIT Health des del 2016 fins a 2020. Durant els primers anys ni la informació que es demanava ni la manera com s'obtenien les dades era la mateixa que s'utilitza avui en dia, per tant hi ha camps que no s'han recollits en tots els registres.

La base de dades inicial te un total de 50 variables, però totes aquestes no són d'interès per la realització del treball. Després de valorar les més interessants s'escullen aquests quinze variables.

- *Project ID*: Identificador del projecte presentat
- *Project Description*: Descripció textual del projecte presentat
- *Year*: Any de participació
- *Programm:* Programa en el qual es va inscriure
- *CLC*: CLC responsable del Programa
- *Status*: Si el projecte va acabar seleccionat o no
- *Country*: País d'origen del projecte
- *Company Name*: Nom de la companyia que presenta el projecte
- *Category*: Categoria on es classifica la proposta
- *TRL:* Nivell de maduresa tecnològica *(Technology Readiness Leavels)*
- *Gender:* Gènere de la persona encarregada de la proposta
- *Startup ID:* Identificador de la companyia
- *Funding:* Finançament que ha rebut la proposta fins al moment
- *Valuation:* Valor monetari de la companyia
- *Funding Stage:* Etapa en la qual es troba el projecte pel que fa a finançament.

Un primer filtratge realitzat és agafar únicament aquells projectes que han omplert el camp: *Project description*, es passa de 5.030 a 3.892 registres.

Un cop treballant amb les dades s'arriba a la conclusió que els camps *Company Name* i *Startup ID* són redundants i no aporten informació comparable entre els diferents projectes, per aquest motiu s'eliminen.

Finalment la base de dades té un total de 3.892 registres amb tretze variables, entre aquestes hi ha un identificador, la variable text, vuit variables factors i tres variables numèriques.

La Taula 2.1 mostra l'esquema visual de com està estructurada la base de dades

Taula 2.1: Estructura de la base de dades

PROJECT ID	*PROJECT DESCRIPTION*	STATUS	YEAR	PROGRAMME	CLC	COUNTRY	CATEGORY	TRL	GENDER	FUNDING	VALUATION	FUNDING STAGE
2017-LLBC-12	delivery demand custom skincar [...] product device	0	2017	Bootcamp	3	DE	3	9	1	-1	-1	0
2017-LLBC-13	digit inclus seniocarespottv [..] digit society	0	2017	Bootcamp	3	DE	3	9	1	-1	-1	0
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
12574	product nlift stand [...] regulatori submiss	0	2020	Start-up Rescue instrument	6	GB	1	4	1	-1	6	2

Les variables factors codificades són les següents:

- Status: 1 = Seleccionat, 0 = Rebutjat
- CLC: 1 = Belgium-Netherlands 2= France 3= Germany 4=Scandinavia 5=Spain 6= Ireland-UK 7= Innostars, 8 = *KIC LE,* 0 = Desconegut
- Category: 1 = *BioTech*, 2=*MedTech* ,3=*DigitalHealth*, 4=*Service/Care Model*, 5= *Other*, 0= Desconegut.
- Gender: 1=Home, 2 = Dona, 3= Altre 4= No ho vol dir, 0 = Desconegut
- Valuation: -1=Desconegut, 1=[1,250), 2=[250,500), 3=[500,1000), 4=[1.000,2.000), 5=[2.000,5.000), 6= ≥5.000.
- Funding Stage: 1=*Pre Seed*, 2= *Seed*, 3= *Series A*, 4= *Series B*, 5= *Series C*, 0=Desconegut.

A partir de les descripcions dels projectes s'ha construït el corpus. El conjunt dels 3.892 registres formen un corpus de 288.465 paraules amb 34.955 mots diferents.

Com a tota base de dades abans de procedir a l'anàlisi cal preparar i "netejar" les dades, en tractar-se de dades textuals els procediments a seguir són diferents, no és únicament mirar els valor *missings* i recodificar variables sinó que cal seguir una sèrie de procediments per tal de poder elaborar un bon anàlisi. [2]

- Eliminar les *stop words*:

La majoria de paraules més freqüents en els textos manquen de significat propi, ja que serveixen per unir paraules o frases, com pot ser el cas d'articles, preposicions, pronoms..., per aquest motiu i per tal de reduir la dimensionalitat es procedeix a eliminar-les[3].

- Eliminar números, signes de puntuació i passar totes les lletres a minúscules.

Per tal de facilitar el processament del text i no haver-hi confusions s'eliminen tots els signes de puntuació, números i es passen a minúscules totes les lletres.

- Lematització de totes les paraules amb una post revisió individual.

Lematitzar és el procés de relacionar una família de mots derivats a un de comú per exemple les paraules *"presentation", "presented", "presenting"* poden ser convertides a "*present*". En aquest cas s'ha utilitzat l'algoritme de Porter's[4] per dur a terme la lematització. Un cop aplicat aquest algoritme es va procedir a una revisió individualitzada d'aquelles paraules que podrien generar dubtes.

- Eliminar aquelles paraules que tinguin una freqüència inferior al 1% dels registres[5], en el cas d'aquesta base de dades 38.

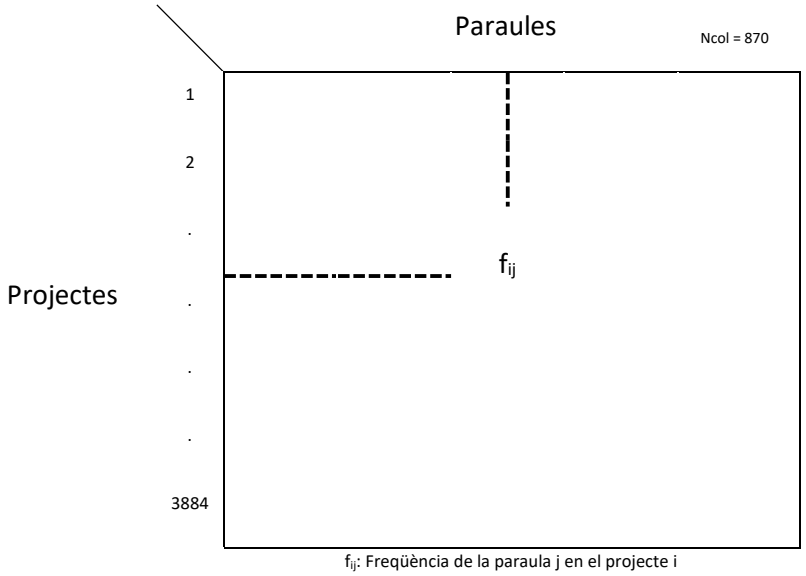
Aquelles paraules que apareixen menys de 38 cops són eliminades.

Un cop processat el text es procedeix a construir la matriu amb la qual es treballarà, on a les files hi ha els projectes i a les columnes les paraules, que en acabar el processat són 870, a més a més hi ha 8 registres que no tenen cap d'aquestes paraules finals i en conseqüència són eliminats.

En paral·lel a aquesta matriu hi ha la taula auxiliar amb la informació corresponent a cada projecte.

En la figura número 2.1 es mostra de manera esquemàtica.

Figura 2.1: Esquema de la matriu corpus



f_{ij}: Freqüència de la paraula j en el projecte i

STATUS	YEAR	...	VALUATION	FUNDING STAGE
0	2017	..	-1	0
0	2017	...	-1	0
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
0	2020	..	6	2

3 METODOLOGIA ESTADÍSTICA

En aquest apartat es presenta la metodologia utilitzada en l'estudi de les propostes de EIT Health: Anàlisi de Dades Textual (ADT), Anàlisi de Correspondències (AC) i Mètode de Classificació.

3.1 CONTEXT

L'anàlisi textual és relativament nou, els primers *papers* on es parla sobre estadístiques lèxiques daten de 1944[6] on es volia realitzar un estudi comparatiu del vocabulari de grans autors. Tot i això aquests primers mètodes estadístics no foren tan importants com els impulsats des de l'escola francesa amb Jean – Paul Benzécri com a màxim exponent[7] on introdueixen el concepte d'anàlisi de correspondències dins del "*text mining*".

3.2 ANÀLISI ESTADÍSTIC DE DADES TEXTUALS

L'estudi de l'anàlisi textual està basat en les tècniques estadístiques desenvolupades en part per l'Escola Francesa d'Anàlisi de Dades [8].

L'anàlisi de dades textuals (ADT) es refereix a procediments que impliquen enumerar les ocurrències de les unitats lingüístiques bàsiques (paraules) i realitzar algun tipus d'anàlisi estadístic a partir d'aquests recomptes.

El desenvolupament de les tècniques de l'anàlisi textual han fet que el processament de textos s'hagi consagrat com una eina interdisciplinària, integrada per l'estadística, l'anàlisi del discurs, la lingüística i la investigació documental entre altres. De la mateixa manera l'ADT és cada vegada més utilitzat en diversos camps de les ciències com en la Història, Política, Sociologia, etc. El ADT desenvolupat a partir de les aportacions de Jean – Paul Benzécri ha permès la construcció i estudi de grans matrius de dades mitjançant l'aplicació de l'Anàlisi Factorial en taules *n x p* (registres x paraules).

3.2.1 Codificació de la informació

Per tal de dur a terme l'estudi primer cal considerar una nova variable anomenada variable textual, codificada en una taula projectes x paraules (taula lèxica) construïda mitjançant el recompte de les diferents paraules en les descripcions del projectes.

Tal com s'ha vist en la creació de la taula lèxica per aquest problema, apartat anterior, identificar les paraules requereix un precís procés que consisteix a corregir faltes, lematitzar paraules, eliminar *stop words* i definir un llindar de freqüència mínima.

Per concloure, en aquest estudi cada fila de la taula representa un projecte i cada columna les diferents paraules conservades dels projectes, el sumatori de cada columna és el nombre total de cops que apareix la paraula en les diferents descripcions.

3.3 ANÀLISI DE CORRESPONDÈNCIES (AC)

El mètode AC, com s'utilitza en l'actualitat, s'atribueix a Benzécri (1973), Escofier (2003) i a Greenacre, (2008) tot i que el seu "pare" és Benzécri.

Benzécri (1981) y Lebart y cols. (1985) van iniciar l'aplicació d'aquest mètode al camp textual, temps després, Lebart va formalitzar mitjançant llenguatge matemàtic aquesta aplicació.

L'anàlisi de correspondències (AC) és una tècnica estadística que s'aplica a l'anàlisi de taules de contingències i construeix un diagrama cartesià basat en l'associació entre les variables analitzades. En aquest diagrama es representen conjuntament les diferents modalitats de la taula de contingència, de manera que la proximitat entre els punts representats està relacionada amb el nivell d'associació entre aquestes modalitats.

3.3.1 Formulació matemàtica

En una taula de contingència, la cel·la *ij* conte la freqüència X_{ij}, és a dir, el nombre de cops que ha ocorregut en l'observació *i* el fenomen *j*. A partir de la suma total de les freqüències s'obté el valor K ($K = \sum_{ij} k_{ij}$). La taula es transforma en una taula de proporcions $f_{ij} = \frac{k_{ij}}{K}$ i es crea una nova columna al final de tot que té com a valor el sumatori de les k_{ij} per cada fila, $f_{i\cdot} = \sum_j f_{ij}$ i es crea el mateix terme per les columnes on $f_{\cdot j} = \sum_i f_{ij}$ és el sumatori per columna de cada fila. Aquests valors són les proporcions marginals.

Figura 3.1: Taula de contingència

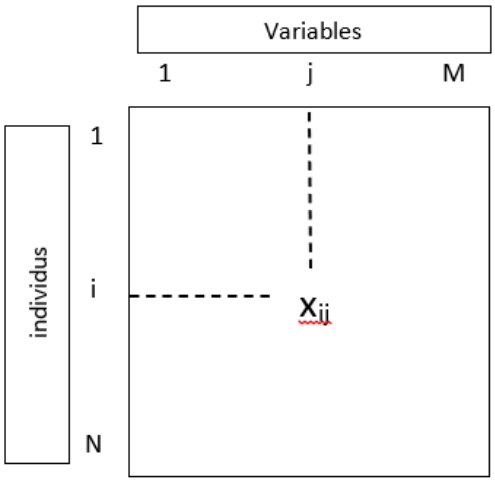
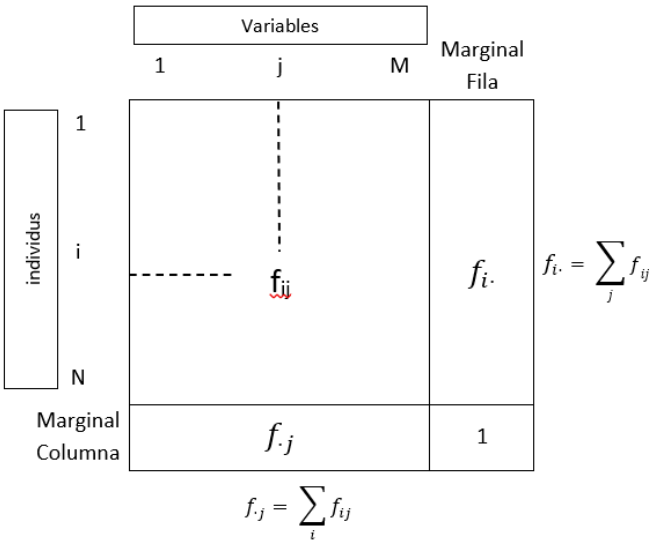


Figura 3.2: Taula de proporcions



En aquest procediment es busca trobar el grau de dependència entre les variables i els individus.

Cal recordar que si dos esdeveniments són Independents la Probabilitat que succeeixin els esdeveniments A i B és P(A∩B) = P(A)*P(B), per tant en el cas de la taula de proporcions es podria escriure com, $f_{i\cdot} * f_{\cdot j} = f_{ij}$, de manera anàloga $\frac{f_{ij}}{f_{i\cdot}} = f_{\cdot j}$ o bé $\frac{f_{ij}}{f_{\cdot j}} = f_{i\cdot}$

Per tal de detectar aquesta dependència es calculen les taules dels perfils fila i columna.

3.3.2 Taula de perfil fila i perfil columna

Figura 3.3: Taula perfil fila

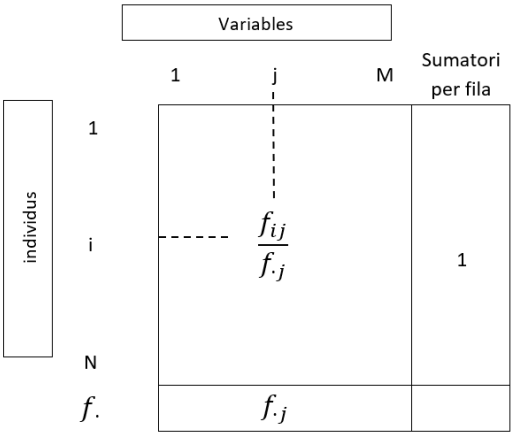
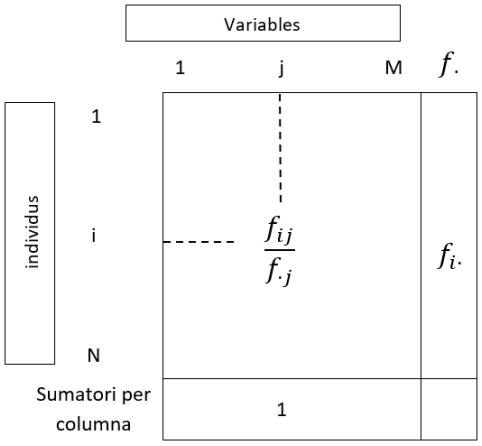
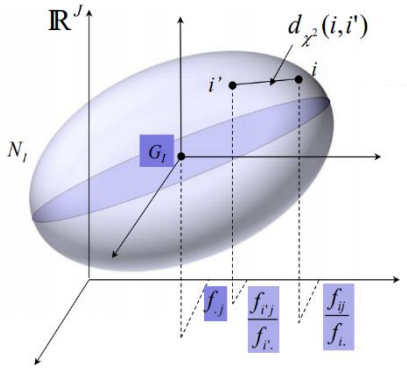


Figura 3.4: Taula perfil columna



A partir de la taula perfil fila es representen els individus en el pla factorial definits per les M variables .

Figura 3.5: Núvol de punts per individu



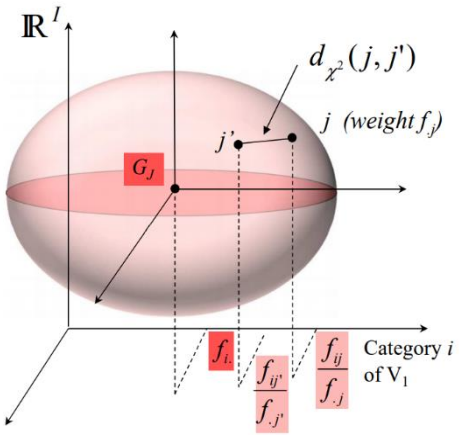
En la figura 3.5 es mostra com es defineix el núvol de punts on Gi és el centre de gravetat de les files.

Es defineix:

$d_{\chi^2}(i,i') = \sum_{j=1}^{J} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2$ com la distància entre dos individus.

De manera idèntica es construeix el núvol de punts per les columnes a partir de la taula perfil columna.

Figura 3.6: Núvol de punts per variable



En el cas d'independència total en la representació gràfica es trobarien totes les observacions en el centre dels perfils, ja que recuperant l'expressió de la probabilitat condicionada $\frac{f_{ij}}{f_{i\cdot}} = f_{\cdot j}$ o bé $\frac{f_{ij}}{f_{\cdot j}} = f_{i\cdot}$ es veu clarament com per cada fila o columna depenent de la taula, tots els perfils fila seria igual a la distribució marginal de les columnes i tots els perfils columna seria igual a la distribució marginal de les columnes. Per tant com a conseqüència es pot afirmar que com més dispersió hi hagi més dependència hi haurà.

3.3.3 Distància chi-quadrat i inèrcia.

L'estadístic per mesurar aquesta dispersió és el chi-quadrat que es calcula de la forma següent.

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(f_{ij} - f_{i\cdot} * f_{i\cdot})^2}{f_{i\cdot} * f_{i\cdot}}$$

Com més gran sigui el valor de l'estadístic menor és la independència entre les dades.

Sobre el núvol de punt original es calcula la inèrcia que quantifica el grau de desviació respecte de la independència de les dades.

Inertia = $\frac{\chi^2}{K}$ on K és el nombre total d'observacions. La Inertia és la mateixa si es calcula pel núvol de punts perfil fila com pel columna.

Es busca trobar un nou pla factorial que contingui la mateixa inèrcia total però que les primeres dimensions maximitzin la inèrcia explicada.

Un cop calculats els nous punts en el nou pla factorial, sigui el de les files o el de les columnes, com que la inèrcia total explicada per files i columnes és la mateixa, es pot calcular les coordenades de les files o les columnes en cada cas a partir de les formules de transició.

$$F_s(i) = \frac{1}{\lambda_s} * \sum_{j=1}^{J} \frac{f_{ij}}{f_{.j}} * G_s(j)$$

On $F_s(i)$ és la coordinada de la fila i en l'eix s i $G_s(j)$ la coordenada de la columna j en l'eix s i $\lambda_s$representa la inèrcia per l'eix s.

Per més detalls es pot consultar el llibre de Husson et al. (2017)

Totes les figures d'aquest apartat han sigut extretes de Husson F., Lê S. & Pagès J. (2017) Exploratory Multivariate Analysis by Example Using R 2nd edition, 230 p., CRC/Press

3.4 *EINES DE SUPORT A LA LECTURA DELS EIXOS PRINCIPALS.*

3.4.1 Metakeys

Les *metakeys* definides per Kerbaol [9] són grups de paraules les quals les seves contribucions a un eix són molt altes. D'aquesta manera tenim dues *metakeys* per eix, una de positiva i l'altre de negativa. Les *metakeys* formades defineixen els temes sobre el que parlen els registres.

Per tal de facilitar l'explicació dels passos a seguir per la construcció de metakeys s'utilitzarà l'esquema dissenyat per Morin [10]

La figura 3.7 mostra el que es pot obtenir en el primer espai factorial quan els documents són monotemàtics. Definint A,B,C i D com a grups de temes, segons les paraules o documents, en aquest cas cada tema té la seva pròpia projecció sobre l'eix i la interpretació a partir d'aquesta figura permet identificar de manera clara els diferents grups temàtics i els documents relacionats amb ells. De manera idíl·lica es vol construir un esquema com aquesta, ja que la interpretació és molt senzilla.

Figura 3.7: Esquema idíl·lic

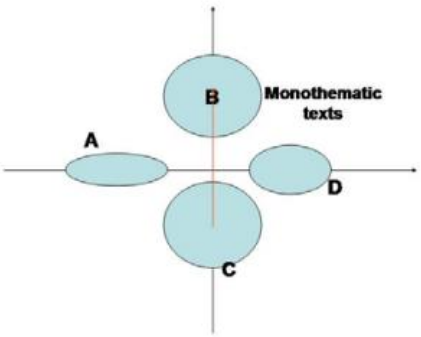
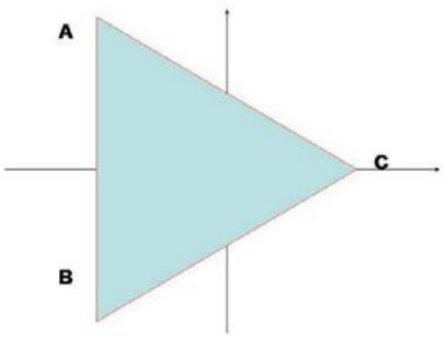


Figura 3.8: Esquema comú



Per altre banda la Figura 3.8 correspon a la situació més freqüent, algun dels temes estan ben representats i són clarament diferenciables (cas del C) en el primer eix i està en oposició amb altres temes A,B. Les projeccions dels temes A i B en la part esquerra del primer eix es barregen, és difícil de interpretar els resultats. Per poder-ho fer es selecciona aquelles paraules i documents amb una gran contribució a la inèrcia, en el cas d'aquest estudi 6 cops la contribució mitjana per paraula, es conservaran aquells eixos dels quals podem trobar *metakeys*.

Un cop identificades les metakeys, cal una revisió d'aquestes per part d'un expert en el camp de l'estudi. En aquest punts s'ha de mostrar els resultats de diferents maneres ja que s'està buscant una associació significativa de paraules que podria referir-se a un tema en particular.

3.5 *MÈTODE DE CLASSIFICACIÓ JERÀRQUIC*

El mètode de clúster o de classificació és un procés estadístic que consisteix en agrupar o desagrupar individus en funció de la seva homogeneïtat, cada una d'aquestes agrupacions s'anomena clúster. En aquest estudi el mètode permetrà agrupar les diferents propostes a partir dels eixos factorials del AC.

En aquest estudi es centra l'interès en els anomenats mètodes jeràrquics que tenen per objectiu agrupar clústers per formar un de nou o bé separar algun de ja existents per originar altres dos, de manera que si successivament es va efectuant aquest procés d'aglomeració o divisió es minimitzi alguna distància o bé es maximitzi alguna mesura de similitud. Com es pot deduir per l'explicació els mètodes jeràrquics es divideixen en ascendents i descendents.

En la realització d'aquest treball es farà ús del mètode ascendent. Primer de tot cal definir una distància entre individus o clústers i tot seguit aplicar la classificació jeràrquica mitjançant un algoritme general. Aquest procés iteratiu és el mateix per tots els mètodes de classificació jeràrquica ascendent, les diferències es basen en el mètode per calcular la distància entre clústers.

Sigui n el conjunt d'individus de la mostra, on resulta el nivell K=0 amb n grups. En el següent nivell s'agruparan aquells dos individus que tinguin una major similitud (menor distància), resultant així $n-1$ grups, a continuació i seguint amb la mateixa estratègia, s'agruparan en el nivell posterior aquells dos individus (o clústers ja formats) amb menor distància, d'aquesta manera en el nivell L tindrem $n-L$ grups formats. Si es continua amb el mateix procediment, arriba el moment en què en el nivell $L = n-1$, únicament hi ha un grup format per tots els individus de la mostra. Aquesta manera de formar nous grups té la particularitat que si en un nivell s'agrupen dos clústers, aquest queden jeràrquicament agrupats pels altres nivells.

Hi ha diferents estratègies que es poden aplicar a l'hora d'unir els diferents grups en els nivells mitjançant un procediment jeràrquic:

Mètode de la distància mínima, mètode de la mitjana entre grups, mètode de la mediana o mètode de Ward.

En aquest estudi s'utilitza el mètode de Ward, aquest mètode és un procediment jeràrquic el qual, en cada etapa, s'uneixen els dos clústers que tingui un menor increment en el valor total de la suma dels quadrats de les diferencies dins de cada clúster. [11]

Sigui x_{ij}^{k} el valor de la variable j-èsima sobre el individu i-èsim del clúster k, suposant que el clúster té n_k individus, m^k el centrodide del clúster k, amb components m_j^k i E_k la suma de quadrats dels errors del clúster k, distància de cada individu del clúster al centroide d'aquest.

$$E_k = \sum_{i=1}^{n_k} \sum_{j=1}^{n} \left(x_{ij}^{k} - m_j^k \right)^2$$

I on E és la suma del E_k. El procés comença amb *m* clústers, cada un d'ells està format per un únic individu, pel que cada observació és el centre del clúster i per tant E=0. L'objectiu del mètode de Ward és trobar en cada etapa aquells dos clústers, els quals la seva unió proporcioni un menor augment de E.

Matemàticament es pot demostrar que el menor increment dels errors quadràtics és proporcional a la distància euclidiana al quadrat dels centroides dels clústers units. La suma E és no decreixent i el mètode, en conseqüència, no presenta problemes que poden generar altres mètodes d'agregació.

Per acabar cal destacar que per cada iteració hi ha un total de $\binom{n}{2}$ possibles combinacions on n és el nombre de clúster existent en aquella iteració.

3.5.1 Nombre de Clústers

Un cop calculat l'arbre jeràrquic resultant de l'agrupació dels clústers cal definir el nombre de clústers final, dit en altres paraules en quin moment s'ha de "tallar" l'arbre.

Es defineix com a inèrcia d'un clúster la suma de les distàncies al quadrat de cada individu del clúster al centre d'aquest.

Es calcula la diferència d'inèrcia en passar d'*n* clústers a *n-1* clústers. En el moment que aquesta diferència sigui més petita que una fita, s'haurà trobat el nombre òptim de clústers. En el cas d'aquest estudi el valor fita és de 0.05.

3.5.2 Estadístic ETA2

L'estadístic ETA2 serveix per calcular quines dimensions del AC identifiquen millor els diferents clústers, calcula l'arrel quadrada del coeficient de correlació entre les variables i les dimensions. L'estadístic pot prendre valors entre el 0 i 1, com més elevat la dimensió identifica millors els clústers.

3.6 *PROFILING DELS CLÚSTERS*

Per tal de realitzar el profiling dels Clústers es realitzarà una sèrie de contrastos d'hipòtesis entre la mitjana de la variable en el conjunt de les dades versus la de dins el clúster.

3.6.1 Contrast d'hipòtesis.

Per les variables numèriques es vol contrastar que la mitjana global sigui igual a la mitjana dins del clúster.

$$H_0: \mu_G = \mu_{clúster}$$

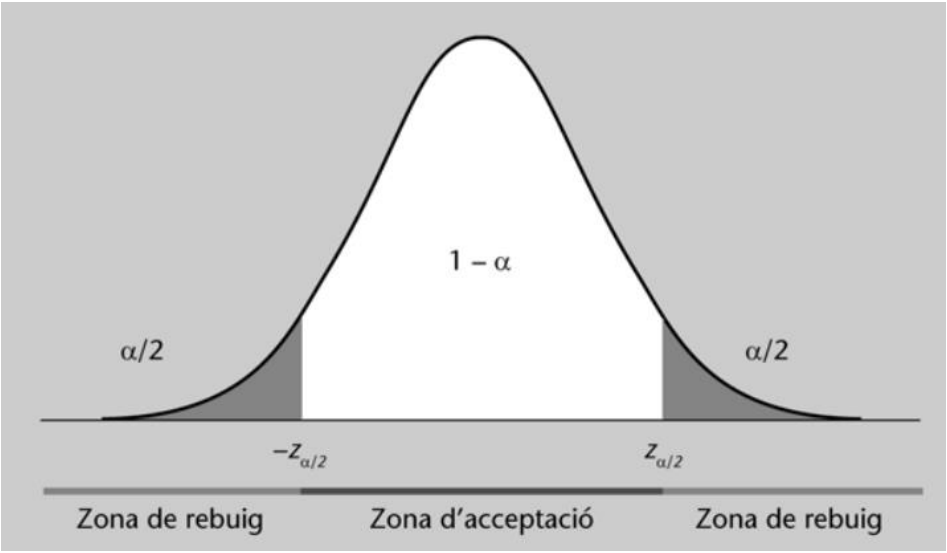
$$H_A: \mu_G \neq \mu_{clúster}$$

Se sap a priori que les mitjanes segueixen una distribució normal així que l'estadístic de contrast serà el següent:

$$z^* = \frac{(\mu_G - \mu_{clúster})}{\sqrt{\frac{\sigma_G^2}{n_G} + \frac{\sigma_{cluster}^2}{n_{cluster}}}} \sim N(0,1)$$

La figura 3.9 mostra la distribució de l'estadístic amb les seves zones de rebuig.

Figura 3.9: Zona de rebuig i acceptació



En el cas d'aquest estudi el nivell de significació alpha és de 0,05.

Cal definir un nou paràmetre anomenat p.valor que indica quina és la probabilitat d'obtenir un valor almenys tan extrem com l'estadístic de contrast obtingut.

En el moment que el p.valor sigui més petit que el nivell de significació alpha es rebutja la hipòtesi nul·la (H_0).

Per les variables categòriques es realitza un contrast per veure si la categoria de cada una de les variables està sobre o infra representada en el clúster.

En aquest cas el contrast d'Hipòtesis no es basa en la mitjana sinó en la proporció.

$$H_0: \frac{n_{mc}}{n_c} = \frac{n_m}{n}$$

$$H_A: \frac{n_{mc}}{n_c} \neq \frac{n_m}{n}$$

On n_{mc} és el nombre d'individus en el clúster c amb la característica m, n_c és el nombre d'observacions en el clúster c, n_m és el nombre total d'individus amb la característica m i per últim n és el nombre total d'observacions.

Sota H_0 les dades segueixen una distribució hipergeomètrica amb paràmetres: H(n_c, $\frac{n_m}{n}$,n) i el p.valor en aquest cas es defineix com $P(N_{mc} \geq n_{mc})$.

Tota aquesta informació queda recollida en una taula resum per clúster on es mostra les 20 característiques més sobre representades. Aquesta taula té per columnes un seguit d'indicadors que ajuden de forma visual a interpretar el clúster.

La primera columna mostra el % de cops que hi ha un projecte amb la categoria corresponent en el clúster respecte al total de projectes amb aquesta característica. La segona columna mostra el % respecte al clúster, és a dir, el % d'observacions amb aquesta característica dins del clúster, en canvi, la següent columna mostra el % sobre el total dels projectes. Les últimes dues columnes fan referència a l'estadístic de contrast i el p.valor corresponent.

4 RESULTATS

Un cop explicada la metodologia a seguir per l'estudi es procedeix a analitzar les dades, primer d'una manera univariant, seguidament realitzant AC i per últim el clúster jeràrquic.

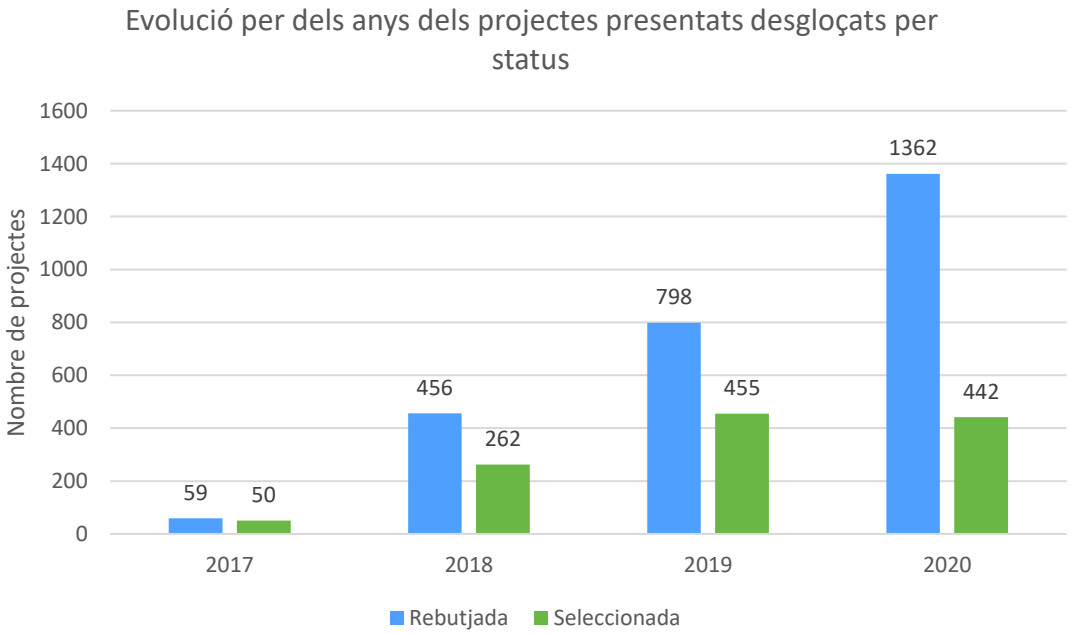
4.1 BASE DE DADES

Abans de res cal recordar que la base de dades està formada per 3.884 observacions, cada un dels projectes, amb 870 paraules i té com a taula auxiliar les principals característiques de cada un d'ells.

Per començar es realitzarà una descriptiva dels projectes i les seves característiques.

Primer es graficarà l'evolució per any del nombre de projectes presentats estratificats per si han estat seleccionats o no.

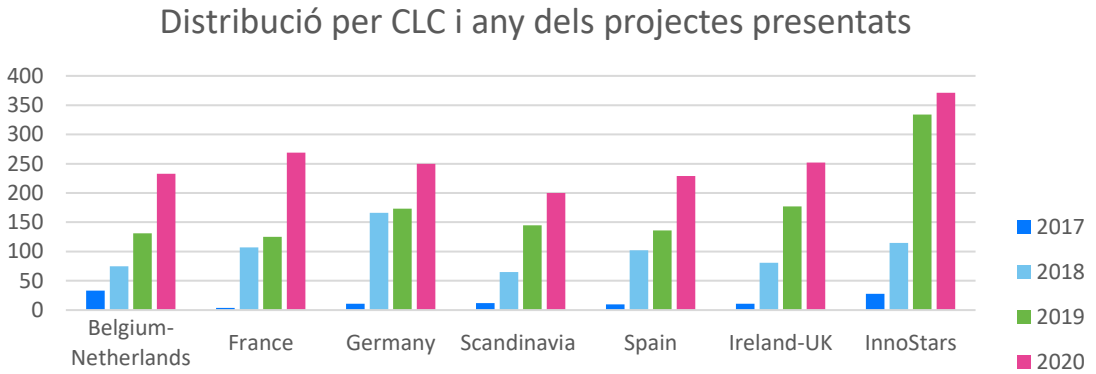
Figura 4.1: Evolució del nombre de propostes



Tal com es pot observar en el gràfic anterior el nombre de propostes no ha parat de créixer durant els últims anys arribant al 2020 a un total de 1804 propostes, de les quals 1362 (75,5%) van ser rebutjades i 442 (24,5%) seleccionades. Tal com es pot veure el ratí d'acceptació ha baixat fent que augmenti la competitivitat de les convocatòries.

Seguidament es veure el nombre de projectes per CLC presentat i com ha variat durant els anys.

Figura 4.2: Evolució del nombre de propostes per CLC



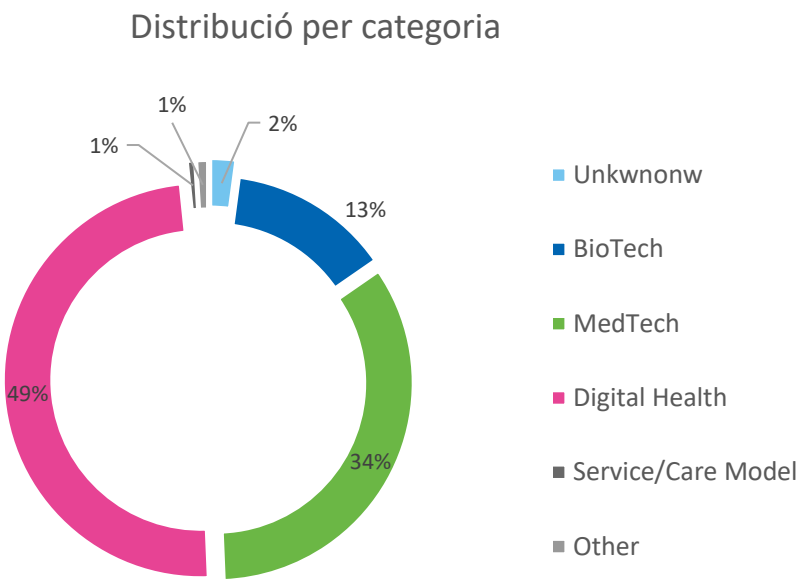
	Belgium-Netherlands	France	Germany	Scandinavia	Spain	Ireland-UK	InnoStars
2017	33	4	11	12	10	11	28
2018	75	107	166	65	102	81	115
2019	131	125	173	145	136	177	334
2020	233	269	250	200	229	252	371

Es pot observar mitjançant el gràfic i amb l'ajuda de la taula inferior que els CLC que reben més propostes són els de Ireland-UK, Germany i Innostar, cal recordar que Innostar agrupa la resta de països de la UE que no tenen un CLC propi i per això és el que gestiona més projectes.

De forma més concreta sobre el total, el 1% de les propostes no van deixar constància del CLC que les gestionava, el 12,15% va ser gestionades per BE-NL, seguit per un 13% al CLC francès, 15,45% a Alemanya, que també s'encarrega de les propostes de Suïssa, el 10,87% ha estat gestionat per els 4 països escandinaus, Dinamarca, Suècia, Finlàndia i Noruega, Espanya el 12,18%, Irlanda i el Regne Unit representen el 13,41% de les propostes i per últim el CLC Innostar s'ha encarregat del 21,83%.

Finalment per acabar aquest apartat més descriptiu es veure la distribució de les propostes segon la seva categoria.

Figura 4.3: Distribució del les propostes per categoria



En la figura 4.3 es pot veure de manera molt clara com la majoria de propostes són categoritzades com Digital Health seguit per MedTech. S'entenen com propostes de Digital Health aquelles que connecten les tecnologies amb la salut per millorar l'eficiència de l'assistència sanitària per fer la medicina més personalitzada i precisa, per altra banda les propostes MedTech són aquelles propostes que implementen la tecnologia dins de la medicina per facilitar la detecció o el tractament d'alguna malaltia.

Un cop coneguda com és la distribució dels projectes en les seves característiques auxiliars es procedeix a realitzar una descriptiva de les paraules que forma el glossari lèxic amb el qual es treballarà.

4.1.1 Glossari

El glossari està format per un total de 870 paraules, tal i com s'ha explicat en la metodologia aquestes paraules han passat per un procés de filtratge i neteja per tal que no hi hagin duplicats ni paraules similars (ex patient i patients).

Aquestes 870 paraules surten un total de 148 580 cops, s'està davant d'un corpus molt repetitiu on les 10 paraules més freqüents acumulen l'11,16% del total d'ocurrències. En tractar-se de projectes de salut dins aquest top 10 apareixen paraules com "patient","medic","health"..., aquestes en concret apareixen un total de 3410, 1426 i 1359 cops, en la taula 4.1 es mostren les paraules i la seva freqüència.

Taula 4.1: Freqüència de les 80 primeres paraules

PARAULA	FREQÜÈNCIA	PARAULA	FREQÜÈNCIA
patient	3410	offer	561
use	2103	person	561
develop	1627	inform	558
data	1501	high	536
medic	1426	tool	534
health	1359	detect	530
provide	1332	enable	519
product	1320	manga	519
device	1288	doctor	518
solute	1216	process	514
system	1116	make	509
clinic	1037	new	505
base	974	people	505
platform	970	model	499
technology	967	application	483
care	964	risk	475
treatment	881	software	475
user	881	active	474
time	841	custom	474
improv	805	design	461
diagnose	774	access	460
allow	767	algorithm	456
test	761	target	456
need	760	cancer	442
monitor	756	one	437
app	717	work	436
analyse	711	result	433
market	695	effect	425
reduce	638	drug	420
hospital	633	cell	419
disease	624	therapy	414
support	624	home	413
help	604	create	412
first	584	increase	411
healthcare	576	project	409
current	574	company	404
cost	572	integer	398
digit	569	measure	395
service	568	prevent	391

En la taula 4.1 es poden observar les 80 paraules més freqüents en el corpus, la primera de totes és "*patient*" amb 3410 aparicions seguida per "*use*" i "*develop*" amb 2103 i 1627, tots els mots apareguts, com era d'esperar, estan relacionat amb la medicina i en el desenvolupament de nous productes en aquest camp. Per tenir una idea més global sobre l'aparició de les paraules s'utilitza la figura 4.4, un *wordcloud* on la mida de les paraules és proporcional a la seva freqüència, permet de forma visual tenir una idea de quines paraules surten més.

Figura 4.4: *WordCloud*

En primer lloc cal remarcar que la localització geogràfica de les paraules no està relacionada amb res, simplement és per fer més amena la lectura.

A primer cop d'ull es pot observar com la paraula més gran és "*patient*" tal com s'havia dit abans és la paraula més freqüent, amb una mida similar es troben *"medic","data","develop","use"* que si es compara amb la taula 3 es veu que són les següents més freqüents, altres paraules que criden l'atenció a primer cop d'ull poden ser *"system","product","clínic","reduce","user","monitor*" etc.... Clarament es pot veure com totes les paraules que apareixen en aquest gràfic fan referència a salut i a nous dispositius o tractament mèdic. Tal com s'ha definit en l'estudi es vol trobar projectes relacionats que es puguin englobar sota les mateixes premisses.

Acabat amb la descriptiva de les paraules present en el glossari es procedeix a realitzar l'anàlisi de correspondències(AC).

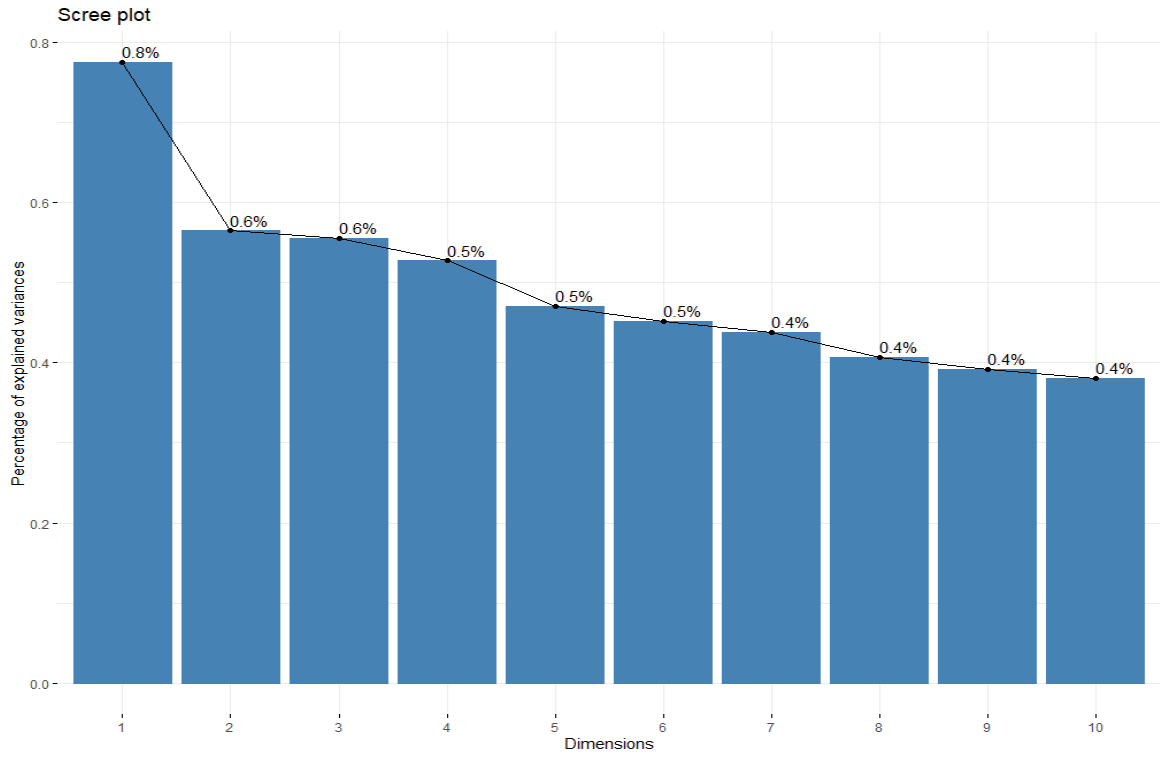
4.2 IDENTIFICACIÓ DELS TEMES A TRAVES DE LES METAKEYS

4.2.1 Anàlisi de correspondències

A traves del AC s'identificaran les *metakeys* tal com s'ha explicat en la metodologia anterior. Primer es presentaran els resultats de l'anàlisi de correspondències i tot seguit les *metakeys* i els temes trobats a partir d'aquests.

Com a tot anàlisi de correspondències primer es presenten els valors propis de cada dimensió. La figura 4.5 mostra el tant per cent de variància explicada per cada una de les 10 primeres dimensions del AC, es pot observar com la primera explica el 0,8% de la inèrcia total i ràpidament cau fins al 0,6% en la segona, els *eigenvalues* d'aquestes dues dimensions són 0,221 i 0,162. Encara que sembli que representen molt poc s'ha de tenir en compte que es té un total de 870 dimensions i que les primeres 30 únicament expliquen 11,39% de la variància total.

Figura 4.5: %Variància explicada per Dimensió

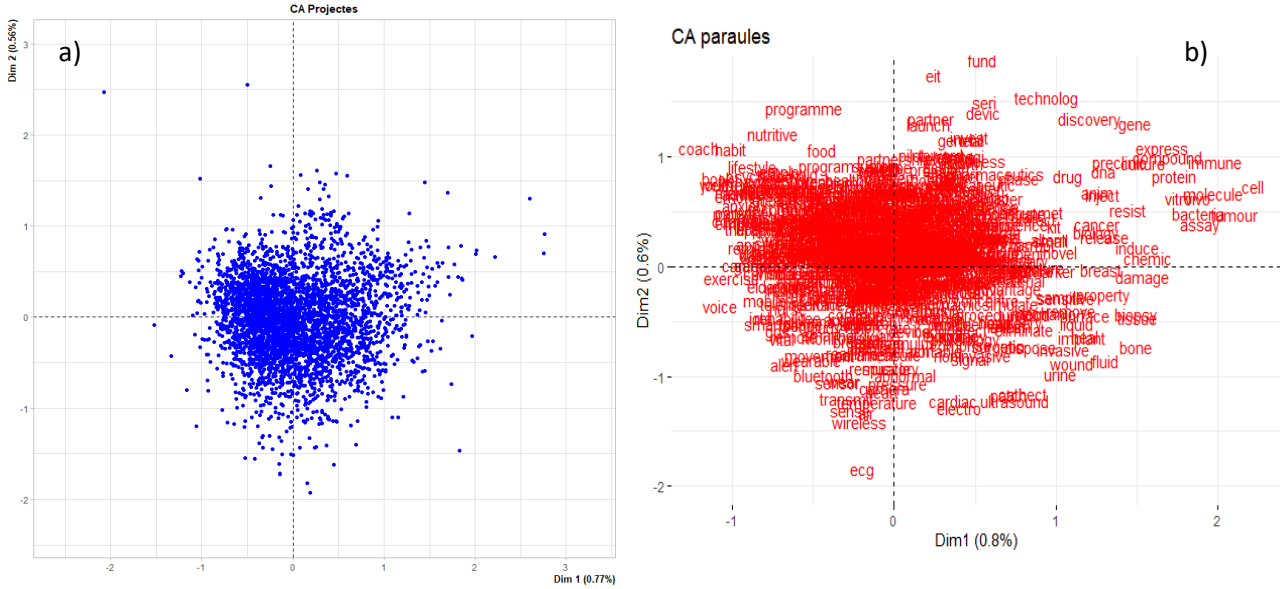


Aquesta dèbil representació pot ser donada, ja que les paraules presenten forta associació entre petits grups i els projectes també tenen aquesta casuística més endavant és veure si aquest supòsit és cert o no. Es un resultat molt habitual en l'AC aplicat en anàlisi textual [12].

Un cop explicades les dimensions i les seves aportacions es construeix el primer pla factorial definit per les dues primeres i el núvol de punts associat als projectes i a les paraules.

En el núvol de punts, figura 4.6a), cada punt representa un projecte sobre el primer pla factorial, es pot observar com la gran majoria de punts estan concèntrics en l'origen de coordenades, però hi ha alguns que es mouen en el primer eix positiu, en canvi en el núvol de punts de les paraules, figura 4.7b), la dispersió és major i es pot observar alguns grups com el cas del primer eix positiu.

Figura 4.7. a) Projecció de projectes sobre el primer pla factoria. b) projecció de les paraules sobre el primer pla.



L'objectiu de construir aquest núvols de punts és l'obtenció de *metakeys* per això amb la figura 4.8 es vol veure quines son les paraules que més contribueixen en els eixos.

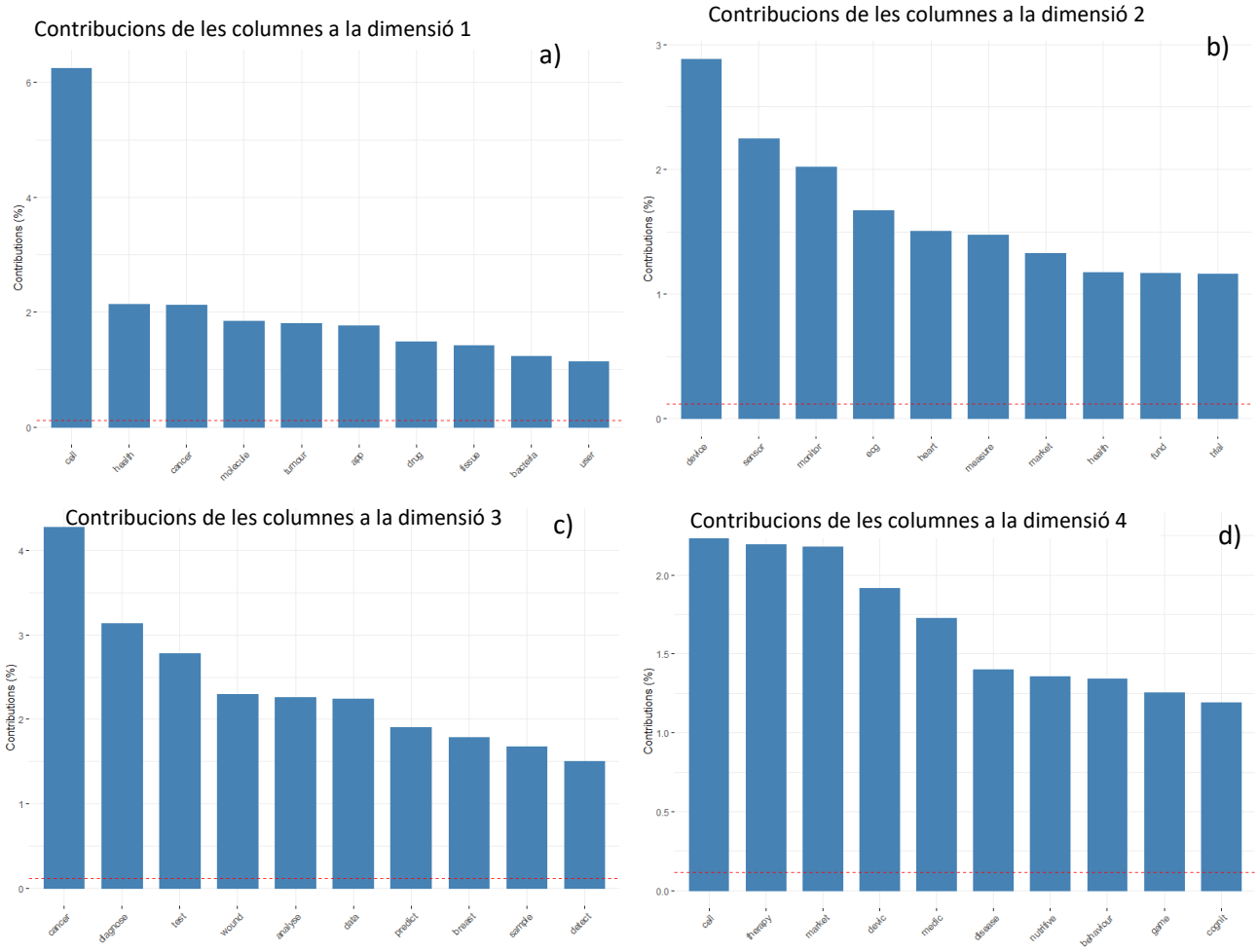


Figura 4.8. a) Top 10 paraules amb mes contribució a l'eix1. b)) Top 10 paraules amb mes contribució a l'eix2. c)) Top 10 paraules amb mes contribució a l'eix3. d) Top 10 paraules amb mes contribució a l'eix4.

Amb aquests 4 gràfics es pot observar les 10 paraules que més contribueixen en les quatre primeres dimensions, les línies vermelles horitzontals discontínues indiquen la contribució mitjana de paraules a l'eix, tot i que no hi ha distinció entre banda positiva i negativa es comença a distingir certes relacions entre paraules.

4.2.2 Identificació de les metakeys

A partir de la metodològica dissenyada per Kerbaol i explicada amb anterioritat s'identifiquen les *metakeys*. Gràcies a les paraules amb una contribució superior a 6 vegades la mitjana es construeixen les *metakeys*, que juntament amb els projectes definiran els temes.

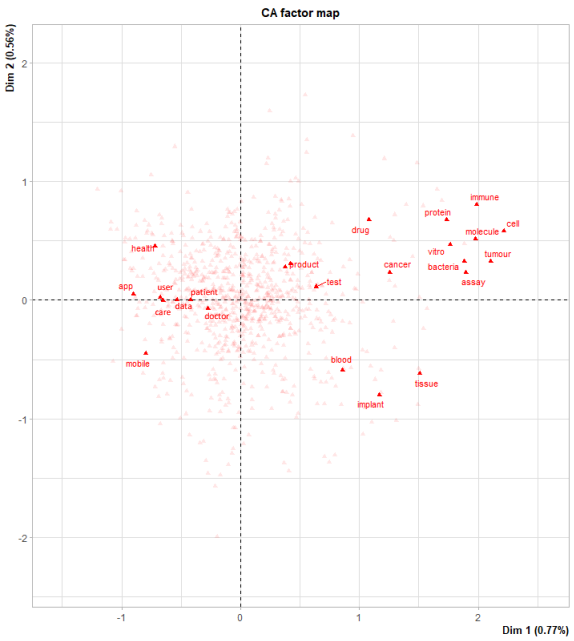
El procediment a realitzar és el mateix per tots els eixos, així que s'explicarà detalladament pel primer eix i es farà de manera anàloga per tots els eixos fins que es trobi en algun eix la impossibilitat de definir un tema per aquest.

Els passos a seguir per l'obtenció de les *metakeys* per cada costat, positiu i negatiu, d'una dimensió són els següents:

- Selecció de les paraules que més contribueixen a cada eix desglossat per pol positiu i negatiu.
- Definició del tema per cada banda positiva i negativa.
- Selecció de projectes que més contribueixen en l'eix desglossat per positiu i negatiu.

En el figura 4.9 es pot veure quines són aquelles paraules que aporten més al primer eix tant per l'esquerra com per la dreta.

Figura 4.9: Projecció sobre el primer pla factorial de les paraules que contribueixen més de 6 cops en la primera dimensió costat negatiu i positiu.



Taula 4.2: Paraules que contribueixen més de 6 cops a l'eix positiu i negatiu, amb gris les següents que contribueixen més en l'eix 1 negatiu.

EIX 1 POSITIU	EIX 1 NEGATIU
cell	health
cancer	app
molecule	user
tumour	care
drug	data
tissue	patient
bacteria	mobile
protein	doctor
test	digit
vitro	service
blood	manga
implant	people
assay	profession
product	person
develop	platform
	mental

Un cop identificades les paraules que més aporten a l'eix cal concretar el tema que tracta aquest, en el primer eix positiu el tema que tracta és *Cancer and cell tractament,* en canvi en el costat negatiu com que no es veu de manera clara quin tema podria englobar aquestes paraules es decideix agafar les següents paraules amb més aportació (marcades en gris fosc) per concretar el tema amb més precisió , en l'eix 1 negatiu es parla sobre *Digital health/ehealth*.

Un cop definit el tema es compara amb la llista de projectes que més contribueixen en l'eix, en la taula 5 es mostren aquells projectes amb una contribució 6 vegades superior a la mitjana per l'eix 1 positiu i pel negatiu.

Taula 4.3: ID dels projectes que contribueixen més de 6 cops a l'eix positiu i negatiu.

CANCER/CELL TRACTAMENT	DIGITAL HEALTH/EHEALTH
2018-HS-0394	2018-BCamp-0234
12564	2018-HS-0191
2020-HS-0650	
2018-HS-0034	
2018-BCamp-0544	

En la taula 4.3 es mostra els projectes que més contribueixen a cada pol del primer eix, ajuden a confirmar si el tema proposat per etiquetar l'eix és el correcte o no. Per exemple, parlant de l'eix 1 positiu la proposta amb ID 2018-HS-0034 feta per l'empresa APTATARGETS S.L consisteix en un tractament per les cèl·lules que han mort en un ictus i la proposta amb ID 2018-BCamp-0544 de l'empresa OPSYON tracta sobre bloquejar cèl·lules canceroses, amb aquestes dues descripcions dels projectes es pot confirmar que l'etiquetatge ha estat el correcte. En relació a l'eix 1 negatiu el projecte amb ID 2018-BCamp-0234 proposat per l'empresa Astonishing Ltd parla sobre la implantació de tecnologia RV per combatre la soledat, per tant també es dóna per bo el seu etiquetatge.

Aquest procés es repeteix per cada una de les dimensions del AC, fins a arribar a la novena dimensió on no s'aconsegueix definir un tema clar per l'eix negatiu. Finalment la taula 4.4 recull la informació sobre cada dimensió i el tema que tracta.

Taula 4.4: Tema tractat per cada pol de cada dimensió

DIMENSIÓ	POL	TEMA
1	+	Cancer/cell therapy
	-	Digital health/ehealth
2	+	Clinical trials
	-	Medical devices
3	+	Implants and operations
	-	Imaging/laboratory test
4	+	Mental health
	-	Surgery
5	+	Medical devices for sleep
	-	Oncology
6	+	Rehabilitate
	-	Dermatology
7	+	Nutrition
	-	Chronic illness
8	+	Early detection
	-	Drug development
9	+	Elderly care

Taula 4.5: Les *metakeys* de cada pol de l'eix.

1		2		3		4	
Cancer/cell tractament	**Digital health/ehealth**	**Clinical trials**	**Medical devices**	**Implants and operations**	**Imaging/laboratory test**	**Mental health**	**Surgery**
cell	health	market	device	wound	cancer	cell	market
cancer	app	health	sensor	implant	diagnose	therapy	devic
molecule	user	fund	monitor	product	test	disease	medic
tumour	care	trial	ecg	heal	analyse	nutritive	surgery
drug	data	clinic	heart	surgery	data	behaviour	surgeon
tissue	patient	project	measure	people	predict	game	clinic
bacteria	mobile	eit	detect	sleep	breast	cognit	launch
protein	doctor	drug	wearable	game	sample	people	sale
test	digit	devic	temperature	skin	detect	mental	hospital
vitro	service	partner	pressure	bone	image	exercise	implant
blood	manga	nutritive	cardiac	stimulus	biomarker	protein	product
implant	people			robot	genetic	stress	trial
assay	profession			active	algorithm	active	fund
product	person			pain	screen	immune	operation
develop	platform			market	score	cancer	print
	mental			infect		***	
						emote	

5		6		7		8	
Medical devices for sleep	**Oncology**	**Rehabilitate**	**Dermatology**	**Nutrition**	**Chronic illness**	**Early detection**	**Drug development**
devic	surgery	reality	wound	user	patient	test	cell
sleep	cell	virtual	bacteria	food	treatment	nutritive	drug
technolog	tumour	game	infect	cell	chronic	detect	data
test	healthcare	cognit	care	nutritive	disease	breast	patient
fund	tissue	rehabilitee	skin	culture	surgery	diagnose	trial
breath	surgeon	train	health	technology	reduce	people	monitor
sensor	patient	therapist	patch	app	care	surgery	clinic
detect	implant	surgery	monitor	product	stroke	urine	therapy
eit	virtual	brain	covid	healthy	clinic	risk	ecg
clinic	procedure	surgeon	prevent	sample	cancer	sample	culture
valid	provide	exercise	heal	mobile	devic	food	devic
project	bone	image	nurse	dna	pain		rehabilitee
heart	service	simul	diabetes	friend	hospital		
market	medic	augment	blood	sensor	risk		
ecg	access	vision	pressure		infect		
monitor	doctor	children			cause		
launch		anxiety					
trial							

9		10	
Elderly care	**names_9_neg**	**Names_10_pos**	**names_10_neg**
care	nutritive	implant	wound
bacteria	wound	bone	heal
fall	skin	heart	skin
cognit	implant	nutritive	patch
people	devic	surgery	diabetes
nurse	surgery	surgeon	game
elder	heal	air	test
brain	lifestyle	ecg	stimulus
home	food	tumour	diagnose
older	bone	cardiac	chronic
dementia	surgeon	cell	reality
covid	person	cathect	
robot	healthy	fall	
detect	coach	human	
molecule	breast	food	
sample	diabetes		

Taula completa a l'annex

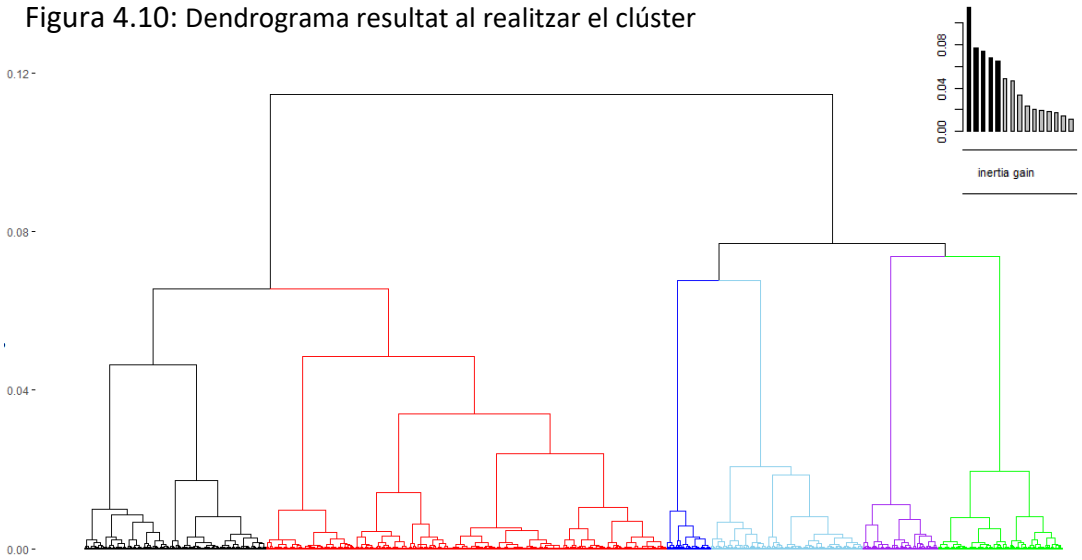
4.3 CLÚSTER JERÀRQUIC

Un cop realitzat l'anàlisi de correspondències i identificats els 17 temes dels quals parlen les propostes es procedeix a realitzar un clúster jeràrquic sobre aquests per veure si es poden identificar grups d'individus que comparteixin característiques comunes.

El clúster s'obté a partir de les 9 primeres components principals de l'AC. Tal com s'ha detallat a la metodologia el mètode utilitzat és el de Ward. Es té un total de 3884 projectes.

La figura 4.10 mostra el dendrograma resultant, amb un total de 6 clústers.

Figura 4.10: Dendrograma resultat al realitzar el clúster



Taula 4.6: Observacions a cada clúster

1	2	4	5	6	3
717	1404	609	565	271	318

Abans de res cal relacionar cada color del gràfic amb el número de clúster corresponent, 1 negre, 2 vermell, 3 verd, 4 blau, 5 blau cel i 6 lila.

En la figura 4.10 s'observa el dendrograma generat a partir del clúster de les components principals de l'AC i la taula 4.6 mostra com estan distribuïts els projectes en els diferents clústers. A primer cop d'ull es veu com clúster numero 2, de color vermell, és el més nombrós amb 1404 projectes en ell, seguit pel primer clúster amb 717 observacions, el clúster menys nombrós és el número 6 que agrupa 271 projectes.

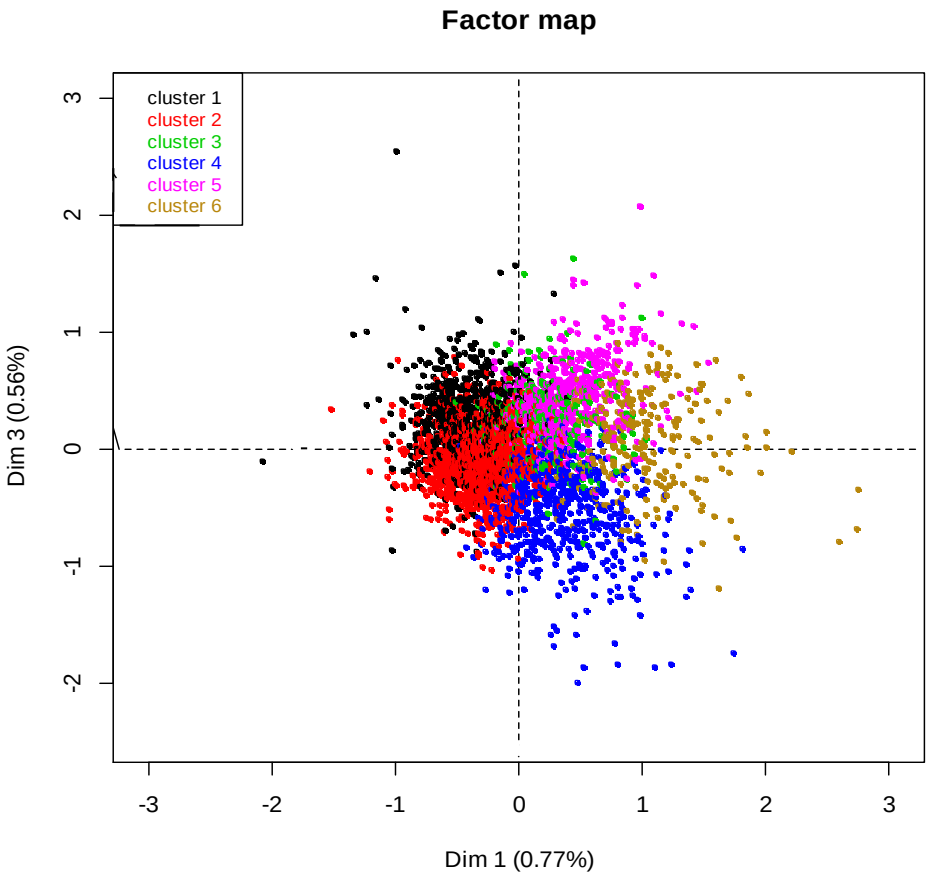
Un cop coneixent el nombre de clústers es vol veure quines són les dimensions que diferencien millor els clústers. Gràcies a l'estadístic ETA2, es pot veure quina dimensió ajusta millor la representació, la taula 4.7 mostra els resultats.

Taual 4.7: Estadístic ETA 2

Dim	*Eta 2*
Dim 1	0.639
Dim 3	0.491
Dim 4	0.481
Dim 2	0.296
Dim 5	0.290
Dim 8	0.249
Dim 6	0.147
Dim 9	0.067
Dim 7	0.042

S'observa com la primera dimensió és la que millor explica els clústers, seguit de la tercera tot i que no hi ha grans diferències entre aquesta última i la quarta.

Figura 4.11: Pla factorial de la primera i tercer dimensió



La figura 4.11 mostrà el pla factorial format per la primera i tercera dimensió, les dues que millor expliquen els clústers, diferenciant cada projecte en el clúster en el qual es troba, en aquesta representació el clúster numero tres no acaba de poder-se distingir amb claredat, però en els altres 5 sí que hi ha una diferenciació més o menys clara. S'ha de tenir present que al ser una representació en dues dimensions es perd molta informació i hi ha punts que es solapen.

Un cop s'ha vist que sis és el nombre òptim de clústers en aquestes dades, s'estudia el comportament de cada clúster, quines són les seves característiques comunes i en què es diferencien entre ells.

4.3.1 Profiling dels clústers

Obtinguts els sis clústers cal estudiar quines són les característiques de cada un i sobre quines temàtiques parlen cada un d'ells.

Abans d'iniciar el *profiling* cal comentar que igual que en el cas dels temes de cada dimensió es detallarà tot el procés pel primer clúster i en els següents es farà de manera més resumida.

4.3.1.1 Clúster 1:

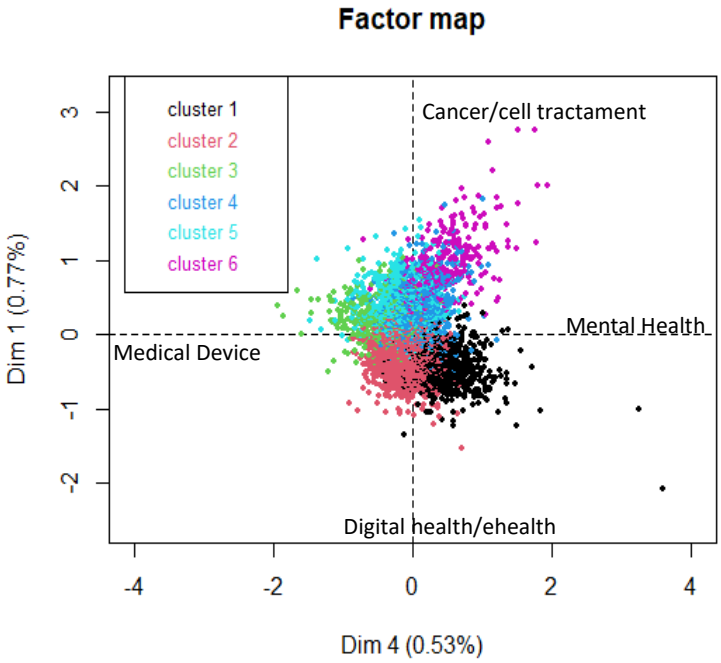
La taula 4.8 mostra les dimensions que millor representen aquest primer clúster. S'observa com l'eix 4 positiu i l'eix 1 negatiu, que fan referència a Mental Health i a Digital Health respectivament, tenen el v.test més extrem, es consideren els pols que més defineixen el clúster 1. La taula també indica que la dimensió 4 és la més dominant, es pot suposar que aquest grup de projectes parla sobre la salut mental amb l'ajuda de la tecnologia.

Taula 4.8: Aportació del clúster 1 a cada dimensió.

	V.TEST	MEAN IN CATEGORY	OVERALL MEAN	P.VALUE
DIM.4	183,9657165	0,458702708	0,007151477	0
DIM.6	114,5024563	0,270978488	0,014412391	0
⋮	⋮	⋮	⋮	⋮
DIM.9	-28,9161955	-0,05804263	0,003972351	7,47E-184
DIM.1	-135,6682451	-0,34723568	0,014765234	0

Taula completa a l'annex

Figura 4.12: Factor map.



En la figura 4.12 es pot veure de manera visual com els projectes que conformen el primer clúster, representats en color negre, es troben en l'espai factorial definit pels temes Mental Health i Digital Health (dim 4 + i dim 1 -).

Cal confirmar que es tracten de projectes relacionats amb la salut mental amb l'ajuda de les paraules més rellevants. És important destacar que aquest anàlisi no es basa en les paraules més freqüents per clúster, sinó en aquelles que estan sobrerepresentades.

La taula 4.9 mostra les paraules del primer clúster amb un v.test superior a 5. Les columnes corresponen al % d'ocurrències dins del clúster, el % respecte al global de registres, la freqüència interna en el clúster i en el total i finalment el p-valor i el v.test del contrast d'hipòtesis.

Taula 4.9: Conjunt de paraules més representades en el primer clúster.

	INTERN %	GLOB %	INTERN FREQ	GLOB FREQ	P.VALUE	V.TEST
MENTAL	0,586863	0,124512	156	185	7,79E-86	19,63485
BEHAVIOUR	0,601911	0,142684	160	212	1,77E-74	18,25842
EXERCISE	0,519148	0,115763	138	172	1,52E-70	17,75703
GAME	0,4251	0,08413	113	125	9,90E-70	17,65158
COGNIT	0,511624	0,116436	136	173	1,88E-67	17,35289
PEOPLE	0,951772	0,339884	253	505	1,13E-60	16,43176
THERAPIST	0,364909	0,072015	97	107	2,79E-60	16,37705
NUTRITIVE	0,432624	0,102302	115	152	4,50E-54	15,48322
APP	1,124821	0,482568	299	717	8,47E-50	14,83678
USER	1,286585	0,592947	342	881	2,51E-48	14,60773
COACH	0,285908	0,058554	76	87	9,64E-45	14,03411
CHILDREN	0,368671	0,093552	98	139	1,05E-41	13,52936
HABIT	0,259574	0,052497	69	78	1,60E-41	13,49811
STRESS	0,353623	0,089514	94	133	3,37E-40	13,27188
ANXIETY	0,237002	0,047113	63	70	4,80E-39	13,07131
⋮	⋮	⋮	⋮	⋮	⋮	⋮
WORK	0,462719	0,293445	123	436	1,51E-07	5,251438
SITUATE	0,142954	0,061246	38	91	1,95E-07	5,204069
INTERACT	0,267098	0,147395	71	219	3,08E-07	5,118233
CONTENT	0,161764	0,074707	43	111	3,82E-07	5,077489

Taula completa a l'annex

De forma molt clara es veu com *Mental* és la paraula més "important" en aquest clúster, per tant, el supòsit inicial es dóna per correcte i les paraules com *behaviour, cognit, game, app, stress, anxiety* concorden amb els resultats previs. Amb tota aquesta informació es pot dir que el primer clúster engloba aquells projectes enfocats a la salut mental amb l'ajuda de les tecnologies.

Un cop definit el tema sobre el que tracta el clúster, cal veure quines són les característiques comunes dins d'aquest. Això es farà amb el suport de les variables auxiliars que defineixen com és cada projecte.

De la mateixa manera que s'han trobat les paraules més representades en el primer clúster es farà amb les característiques dels projectes. Es mostraran aquelles categories

de les variables què la seva representació es diferenciïn més entre la de dins del clúster i la del total dels projectes, es busca trobar les categories més sobrerepresentades.

Primerament es farà aquesta caracterització per les variables categòriques i a continuació per les numèriques.

En la taula 4.10 es mostren les 20 característiques amb el v.test més elevat, a partir d'aquí es comentarà la taula i és realitzarà la descripció de les propostes que conformen el clúster.

Taula 4.10: Conjunt de característiques més representades en el primer clúster.

	CLA/MOD	MOD/CLA	GLOBAL	P.VALUE	V.TEST
CATEGORY=3	26,13232	76,87909	52,63293	0	Inf
PROGRAMME=BOOTCAMP	26,12216	20,75465	14,21456	3,61E-229	32,32046
COUNTRY=RO	42,67625	4,691144	1,966617	4,50E-217	31,44776
STATUS=0	19,6342	79,71936	72,64033	1,87E-188	29,27995
COUNTRY=CZ	57,34649	1,967497	0,613811	2,19E-157	26,72865
GENDER=2	23,11972	27,5111	21,28887	5,70E-157	26,69286
PROGRAMME=BRIDGEHEAD	36,35795	4,371379	2,15103	1,39E-138	25,05895
VALUATION=1	28,06183	8,878188	5,66025	7,13E-124	23,67127
FUNDING.STAGE=1	21,6371	33,95907	28,07915	4,88E-119	23,19705
COUNTRY=GB	25,55783	10,34158	7,239198	1,98E-94	20,61592
COUNTRY=DK	30,58143	3,581371	2,095168	9,62E-68	17,39121
COUNTRY=SK	59,27052	0,733579	0,22143	9,03E-63	16,72221
FUNDING.STAGE=0	21,21254	22,85005	19,27177	5,80E-58	16,04914
COUNTRY=EE	29,1933	3,607704	2,21093	8,90E-58	16,02252
CLC=4	21,3411	13,63705	11,43223	2,48E-34	12,2183
CLC=6	20,81439	16,36446	14,06582	1,26E-31	11,70069
CATEGORY=4	29,34456	1,835829	1,119262	1,82E-30	11,472

COUNTRY=AE	61,44578	0,191859	0,055862	1,71E-18	8,774868
GENDER=4	30,10336	0,876533	0,520931	1,16E-16	8,287584
CATEGORY=0	24,81001	2,087879	1,505586	1,68E-16	8,243

La primera columna mostra el % de cops que hi ha un projecte amb la variable igual a un valor dins el clúster 1 respecte al total de projectes amb la variable igual a aquest valor, és a dir, en la primera fila, sobre el total de projectes amb la variable category igual a 3 el 26,13% es troben dins el clúster 1. La segona columna mostra el % respecte al clúster, és a dir, de tots els projectes del clúster 1 el 76,8% tenen la variable category igual a 3, en canvi la següent columna mostra aquest % sobre el total dels projectes. Amb aquesta diferència es construeix el contrast d'hipòtesis. Cal recordar que es un clúster format per 717 projectes.

Un cop entès que vol dir cada valor de la taula 16 es procedeix a comentar les característiques principals.

Com era d'esperar la majoria de les propostes d'aquest clúster pertanyen a Digital Health (category = 3), un 76,87%, a més el programa Bootcamp està sobrerepresentat, un 20,7% de les propostes han aplicat a aquest programa, això fa pensar que els projectes es troben en estat més embrionari, fet que es confirma amb la també sobrerepresentació de Funding Statge igual a 1 (Pre-Seed) i Valuation igual a 1 ([1,250]).

Per altra banda la proporció de projectes seleccionats és inferior en aquest clúster que en el total, quasi el 80% de propostes d'aquest clúster són rebutjades i en canvi en el total el rati es troba al voltant del 70%. Pel que fa a la localització dels projectes presentats, en països com Romania, Txèquia i Eslovàquia el seu gruix de projectes es troba dins d'aquest clúster, el 42% de projectes presentats per Romania, el 57,8% dels de Txèquia i el 59,27% d'Eslovàquia. De manera no tan contundent països com Gran Bretanya i Dinamarca també tenen una sobrerepresentació en aquest clúster, passant d'un 7% sobre el total a un 10% en el cas de GB i d'un 2% a un 3,5% en el cas de Dinamarca.

Finalment cal destacar la sobrerepresentació per part femenina en aquest clúster on augmenta quasi 5 punts respecte al global.

Pel que fa a la variables numèriques no s'observen grans diferencies entre el clúster i el global.

Finalment per veure les característiques del clúster en la taula 4.11 es mostraran les descripcions dels projectes que es troben més a prop del centre del clúster en el pla factorial. En la taula es pot veure com totes 3 són de la categoria Digital Health(3), tenen status 0 (no seleccionades) i dues d'elles presentades per dones i el Funding Stage no supera el Seed.

Taula 4.11: Descripció dels projectes més apropa del baricentre del clúster 1

ID	CARACTERÍSTIQUES	DESCRIPCIÓ
2020-HS-0232	Program = Headstart Country = CH Category = 3 Funding stage = 1 Gender = 2 Status = 0	Despite being common, mental illness is underdiagnosed by doctors. […]. But their is a lack of a scientific wearable for the same which people can use in day2day life and take necessary actions accordingly. [...] The headband measures the electrical activity of your brain and the mobile app based on the sensor data recommends you scientific meditation techniques to achieve your desired state
2019-HS-0105	Program = Headstart Country = FR Category = 3 Funding stage = 1 Gender = 2 Status = 0	Our digital solution is to prevent WMSDs thanks to video-based physical exercises. Either accessible by employees through general routines, or prescribed and configured by physiotherapists according each individuals situation measured by a unique functional assessment realised in the field.
2020-HS-0491	Program = Headstart Country = ES Category = 3 Funding stage = 2 Gender = 1 Status = 0	Psious' product consists of an extensive platform (50+VR environments) and includes features such as measurement of subjective stress, therapist-patient interaction within the Virtual Reality and session reports to monitor progress. [...]anxiety levels across the world are dramatically rising, while people are less likely to be able to see their mental health professionals. Therefore, we have developed a mobile App [...] that therapists can safely treat patients from home, connected via a 4-digit code

4.3.1.2 Clúster 2:

La taula 4.12 mostra les dues dimensions amb major importància en la representació del clúster 2. En aquest cas el pol 6 negatiu i el 1 negatiu.

Taula 4.12: Aportació màxima del clúster 2.

	V.TEST	MEAN IN CATEGORY	OVERALL MEAN	P.VALUE
DIM.6	-110,926	-0,13464	0,014412	0
DIM.1	-160,772	-0,24249	0,014765	0

Taula completa a l'annex

Amb l'ajuda de la taula 4.5 es concreten els temes que parlen aquests dos pols Dermatology i Digital Health, en compartir dimensió amb el primer clúster és evident que tindran característiques comunes, a més a més com es pot veure en el dendrograma (figura 4.10), com aquests dos clústers estan un al costat de l'altre fet que implica un nivell d'associació elevat.

La taula 4.13 mostra les paraules amb un v.test superior a 5 en el segon clúster.

Taula 4.13: Conjunt de paraules més representades en el segon clúster.

	INTERN %	GLOB %	INTERN FREQ	GLOB FREQ	P.VALUE	V.TEST
PATIENT	3,332417	2,29506	1940	3410	8,89E-99	21,09475
DATA	1,690257	1,01023	984	1501	3,54E-95	20,69899
CARE	1,135427	0,648809	661	964	4,06E-76	18,46348
DOCTOR	0,688814	0,348634	401	518	3,27E-70	17,71404
MEDIC	1,48241	0,959752	863	1426	1,00E-59	16,2992
SERVICE	0,699121	0,382286	407	568	2,94E-55	15,65777
HEALTHCARE	0,680225	0,38767	396	576	1,51E-46	14,32559
PLATFORM	1,015185	0,652847	591	970	1,56E-42	13,66894
MANGA	0,611516	0,349307	356	519	1,23E-41	13,51786
MONITOR	0,812491	0,508817	473	756	2,01E-38	12,96215
INFORM	0,632129	0,375555	368	558	3,45E-37	12,74221
HEALTH	1,286588	0,914659	749	1359	1,43E-32	11,88418
HOSPITAL	0,673354	0,426033	392	633	9,98E-31	11,52407
REMOTE	0,28858	0,145376	168	216	1,32E-30	11,49972
⋮	⋮	⋮	⋮	⋮	⋮	⋮
APP	0,606363	0,482568	353	717	5,68E-08	5,42866
ECG	0,115089	0,067977	67	101	6,27E-08	5,410945
UPDATE	0,066992	0,034325	39	51	1,35E-07	5,272483
INTERFACE	0,125395	0,076726	73	114	1,43E-07	5,261767
CONTINUE	0,202693	0,139992	118	208	4,60E-07	5,042463

Taula completa a l'annex

En aquest cas al ser un clúster molt nombrós (1404) i més heterogeni s'ha difuminat el tema tractat per les dimensions i la realitat, llegit el llistat de paraules amb més representació es pot dir que aquest clúster parla sobre noves maneres d'afrontar els reptes sobre la salut amb l'ajuda de les noves tecnologies.

Per realitzar el profiling es mostrarà en la taula 4.14 les 20 característiques amb un v.test més elevat a partir d'aquí es comentarà la taula i és realitzarà la descripció de les propostes que conformen el clúster.

Taula 4.14: Conjunt de característiques més representades en el segon clúster.

	CLA/MOD	MOD/CLA	GLOBAL	P.VALUE	V.TEST
CATEGORY=3	56,97808	76,5391	52,63293	0	Inf
PROGRAMME=HEADSTART	43,23112	59,45616	53,88679	8,77E-263	34,63054
CLC=7	46,367	24,11536	20,37825	7,03E-179	28,51789
STATUS=0	41,34755	76,6559	72,64033	4,08E-173	28,04923
COUNTRY=GR	66,84337	2,382507	1,396554	1,04E-144	25,615
COUNTRY=DE	46,98282	15,74138	13,12761	9,80E-125	23,7548
FUNDING.STAGE=2	42,78568	43,48976	39,82636	2,97E-118	23,1192
COUNTRY=FR	46,38691	14,75368	12,46197	1,05E-100	21,3037
CLC=2	45,9512	14,62141	12,46736	3,16E-89	20,02776
CATEGORY=4	60,91401	1,740071	1,119262	3,57E-72	17,96636
COUNTRY=HR	77,44681	0,625258	0,316328	3,64E-64	16,91256
PROGRAMME=START-UP MEET PHARMA	50,48935	4,519376	3,507202	1,74E-63	16,8201
VALUATION=3	44,0258	17,70304	15,75515	7,58E-61	16,45613
COUNTRY=IT	48,87784	5,574069	4,4683	2,54E-60	16,38278
COUNTRY=NG	100	0,249072	0,097591	8,90E-60	16,30631
GENDER=-1	43,20929	21,92009	19,87683	4,53E-56	15,7763
COUNTRY=US	72,08333	0,594338	0,323058	2,43E-48	14,60997

PROGRAMME=INNOSTARS AWARDS	51,02228	2,872063	2,205546	1,02E-43	13,86602
COUNTRY=BE	49,56765	3,347877	2,646386	9,37E-41	13,36746
COUNTRY=IN	83,41969	0,276556	0,129896	1,50E-36	12,62686

Com s'ha dit abans en ser el clúster més nombrós i heterogeni les interpretacions són més difícils, ja que aquelles categories de les variables amb poca representació es troben en aquest clúster com és el cas de Nigerià, Índia i Estats Units.

Com era d'esperar la categoria dominant és Digital Health on el 76,5% dels projectes dins del clúster s'han classificat d'aquesta manera. Els programes HeadStart i Stat-Up Meet Pharma tenen una sobrerepresentació notable en aquest clúster i de manera similar a l'anterior, la ràtio de projectes seleccionats és inferior a la mitjana global, en canvi els projectes dins aquest clúster tenen un nivell de maduresa superior al del primer.

Per altra banda el CLC que predomina en aquest grup és el Innostars(7) seguit del francès(2) i en conseqüència França està sobrerepresentada en aquest clúster.

La taula 4.15 mostra la variable numèrica amb més variació en el clúster respecte a la mitjana global.

Taula 4.15: Variable numèrica

	V.TEST	MEAN IN CATEGORY	OVERALL MEAN	SD IN CATEGORY	OVERALL SD	P.VALUE
TRL	47,51377	5,543287	5,170332	2,493171	2,428503	0

Tal com s'havia dit en les variables categòriques els projectes d'aquest clúster tenen un grau de maduresa econòmica superior i en tractar-se de noves formes d'enfocar la medicina fent ús de les noves tecnologies, el nivell tecnològic(TRL) també és superior a la mitjana global.

Finalment per veure les característiques del clúster en la taula 4.16 es mostren les descripcions dels projectes que es troben més a prop del centre del clúster. En la taula es pot veure com totes 3 són de la categoria Digital Health (3), tenen status 0 (no seleccionades) i dues d'elles van participar a HeadStart.

Taula 4.16: Descripció dels projectes més apropa del baricentre del clúster 2

ID	CARACTERÍSTIQUES	DESCRIPCIÓ
2019-BC-1124-4076	Program = Bootcamp Country = EE Category = 3 Funding stage = 1 Gender = 1 Status = 0	We will perform a scientific study [..]which can assess medical conditions based on online forms, information provided by video calls and medical history.[...] We use communication technology to provide a fully functioning online queue system for emergency rooms (ER) so that people can stay at home and therefore reduce the time people have to spend in the ER. Hospitals get valuable information about arriving workload hours earlier and can plan their work based on that.
2020-HS-0549	Program = Headstart Country = NL Category = 3 TRL = 8 Funding stage = 3 Status = 0	Pipple has developed a capacity tool for the planning of medical professionals in a COVID-19 hospital under the analysis of patient scenarios. This tool is currently quite basic yet has proved to be already be of meaningful value to hospitals. [...] The tool could proof to be useful for hospitals that are in the need for a more centralized approach to employee planning than regularly used.
2020-HS-0491	Program = Headstart Country = ES Category = 3 Funding stage = 2 Gender = 1 Status = 0	We will deliver a user-friendly software platform that enables players from business, government, the press and academia to deploy sophisticated epidemiological prediction tools. [...]The open-source model is that of the MRC Ctr for Global Infect Disease Analysis, with whom we have a track record of collaboration;[...]

4.3.1.3 Clúster 3:

La taula 4.17 mostra les dues dimensions amb major importància en la representació del clúster 3. En aquest cas el pol 5 positiu i el 2 positiu.

Taula 4.17: Aportació màxima del clúster 3.

	V.TEST	MEAN IN CATEGORY	OVERALL MEAN	P.VALUE
DIM.5	158,2074	0,552667	0,016235	0
DIM.2	156,0967	0,560482	-0,02944	0

Taula completa a l'annex

Amb l'ajuda de la taula 4.5 es concreten els temes que parlen aquests dos pols, Medical Devices i Clinical Trials.

La taula 4.18 mostra les paraules amb un v.test superior a 5 en el tercer clúster.

Taula 4.18: Conjunt de paraules més representades en el tercer clúster.

	INTERN %	GLOB %	INTERN FREQ	GLOB FREQ	P.VALUE	V.TEST
DEVIC	0,944095	0,07538	102	112	2,66E-103	21,58182
MARKET	2,212144	0,467761	239	695	2,14E-95	20,7231
PRODUCT	2,665679	0,88841	288	1320	1,28E-63	16,83835
CLINIC	2,054795	0,697941	222	1037	7,87E-48	14,52958
LAUNCH	0,66642	0,086149	72	128	2,47E-47	14,45115
FUND	0,573862	0,063266	62	94	5,54E-47	14,39529
TRIAL	0,953351	0,187105	103	278	3,23E-45	14,11141
TECHNOLOG	0,472047	0,044421	51	66	1,40E-44	14,00777
PROJECT	1,092188	0,275273	118	409	5,18E-39	13,06563
EIT	0,490559	0,057881	53	86	4,73E-38	12,89623
VALID	0,740466	0,183739	80	273	2,25E-27	10,83909
FIRST	1,1107	0,393054	120	584	8,65E-25	10,28023
SALE	0,453536	0,091533	49	136	1,43E-21	9,540095
INVEST	0,305442	0,041728	33	62	2,57E-21	9,479037
⋮	⋮	⋮	⋮	⋮	⋮	⋮
PHASE	0,268419	0,085476	29	127	5,64E-08	5,429835
DELAY	0,157349	0,036344	17	54	3,02E-07	5,121896
POSIT	0,342466	0,135281	37	201	3,27E-07	5,107404
MONTH	0,286931	0,103648	31	154	4,11E-07	5,063861
GOAL	0,286931	0,104321	31	155	4,78E-07	5,034982

Taula completa a l'annex

Clarament les paraules més rellevants estan sota el mateix paraigües Clinical Trials, per tant aquest clúster parla sobre nous medicaments que ha de passar un seguit de fases per poder ser de domini públic.

Per realitzar el profiling es mostrarà en la taula 4.19 les 20 característiques amb un v.test més elevat a partir d'aquí es comentarà la taula i és realitzarà la descripció de les propostes que conformen el clúster.

Taula 4.19: Conjunt de característiques més representades en el tercer clúster.

	CLA/MOD	MOD/CLA	GLOBAL	P.VALUE	V.TEST
FUNDING.STAGE=3	18,89681	29,93336	11,51837	0	Inf
VALUATION=6	36,99107	43,30803	8,513259	0	Inf
PROGRAMME=START-UP RESCUE INSTRUMENT	78,84179	41,83636	3,858527	0	Inf
CATEGORY=2	10,61194	45,82562	31,40059	6,14E-233	32,58762
CATEGORY=5	32,29974	4,627916	1,041863	3,56E-186	29,10024
COUNTRY=IL	29,00763	3,868937	0,969848	7,36E-137	24,90042
CLC=1	11,19887	19,1595 7	12,44044	3,01E-95	20,70673
COUNTRY=CH	17,76799	5,599778	2,291695	1,12E-93	20,53159
FUNDING.STAGE=4	24,32035	3,06368	0,916005	5,69E-86	19,65083
COUNTRY=IE	12,37723	12,01407	7,05815	9,61E-83	19,26993
PROGRAMME=BRIDGEHEAD GLOBAL	19,01141	4,165124	1,593081	6,02E-79	18,81206
PROGRAMME=BRIDGEHEAD EUROPE	14,97025	4,424287	2,149011	4,52E-51	15,03211
CLC=6	9,550696	18,47464	14,06582	8,13E-40	13,20575
COUNTRY=BR	45,03311	0,629397	0,101629	6,71E-37	12,69013
PROGRAMME=GO GLOBAL INBOUND	20,74534	1,545724	0,541796	1,04E-34	12,28859
CATEGORY=1	9,564979	16,17919	12,29977	1,09E-34	12,28481
COUNTRY=NL	9,783474	11,37542	8,454705	1,82E-27	10,85837
FUNDING.STAGE=5	21,31716	1,138467	0,388343	4,48E-27	10,77574

COUNTRY=MX	100	0,212884	0,01548	6,43E-27	10,74253
CLC=0	19,60784	1,295816	0,480549	9,26E-27	10,70876

En aquest tercer clúster i de manera lògica en tractar-se d'assajos clínics i productes que estan a punt de sortir al mercat són projectes molt desenvolupats a escala econòmica. La gran majoria de projectes en Series A, B o C es troben en aquest grup. A més el 43% de les propostes d'aquest clúster tenen un producte valorat en 6 milions d'euros o més.

El 41,84% dels projectes d'aquest grup han participat en el Start-up Rescue instrument, això significa que el 78,84% de tots els projectes que han participat en aquest programa es troben en aquest clúster.

Per altra banda un bon grapat de les propostes internacionals, fora de la UE, es troben en aquest clúster, participants del Go Global Inbund, propostes de països com Mèxic, Israel o Brasil.

En l'àmbit europeu el 20% de les propostes d'aquest clúster van venir a partir del CLC Belgium-Netherlands i un 18% per part del CLC Ireland-UK.

Per acabar aquest profiling de les variables categòriques el 42% de les propostes d'aquest clúster es classifiquen com "MedTech" mentre que un 16% restant com "BioTech".

La taula 4.20 mostra les variables numèriques amb més variació en el clúster respecte la mitjana global.

Taula 4.20: Variables numèriques

	V.TEST	MEAN IN CATEGORY	OVERALL MEAN	SD IN CATEGORY	OVERALL SD	P.VALUE
YEAR	43,85722	2019,589	2019,247	0,780517	0,839969	0
TRL	26,34885	5,763143	5,170332	2,587114	2,428503	5,29E-153

Tot i que no s'aprecia directament sembla que les propostes d'aquest clúster són més recents i tenen un nivell tecnològic superior a la del conjunt de projectes.

Finalment per veure les característiques del clúster en la taula 4.21 es mostraran les descripcions dels projectes que es troben més a prop del centre del clúster. En la taula es pot veure com totes 3 van participar l'any 2020, són projectes amb un gran valor de mercat i dues d'elles van participar a Start-up Rescue.

Taula 4.21: Descripció dels projectes més apropa del baricentre del clúster 3

ID	CARACTERÍSTIQUES	DESCRIPCIÓ
12495	Program = Start-up Rescue instrument Year = 2020 Category = 3 Funding stage = 3 TRL = 7 Valuation = 6	ReHub platform is the first Digital Recovery Therapy solution that delivers effective, personalized home rehabilitation for MSD sufferers. An end-to-end solution built on solid medical and biomechanical evidence, The ReHub Launch project at this stage aims to accelerate the market entry and market traction of ReHub solution. [...]
12540	Program = Start-up Rescue instrument Year = 2020 Category = 3 Funding stage = 2 TRL = 8 Valuation = 4	As the B2B sales and marketing to hospitals has longer sales cycles and heavily delayed by Covid-19. The Head MDM patient version is now planned to a implemented and marketed during the Covid-19 crisis.[...] The existing application optimised for hospital use is based on and tablet which is planned to released with a simplified user interface and on a smartphone base . 1) Application adaption including simplified UI and smartphone format. 1) IOS then android evaluation
8847	Program = European Health Catapult Year = 2020 Category = 2 Funding stage = 2 TRL = 5 Valuation = 5	PHLECS Full Body Blue Light is a UV-free phototherapy device for the treatment of psoriasis patient. The CE certification is expected in Q2 2020 so we will launch immediately in Europe, followed by the US. In order to accelerate market penetration, we want to invest in health economics studies and the development of a low cost device.

4.3.1.4 Clúster 4:

La taula 4.22 mostra les dues dimensions amb major importància en la representació del clúster 4. En aquest cas el pol 3 negatiu i el 8 positiu.

Taula 4.22: Aportació màxima del clúster 4.

	V.TEST	MEAN IN CATEGORY	OVERALL MEAN	P.VALUE
DIM.3	-213,613	-0,54308	0,00628	0
DIM.8	108,7992	0,242184	-0,00291	0

Taula completa a l'annex

Amb l'ajuda de la taula 4.5 es concreten els temes que parlen aquests dos pols, Imaging/laboratory test i Early detection.

La taula 4.23 mostra les paraules amb un v.test superior a 5 en el quart clúster.

Taula 4.23: Conjunt de paraules més representades en el quart clúster.

	INTERN %	GLOB %	INTERN FREQ	GLOB FREQ	P.VALUE	V.TEST
DIAGNOSE	1,964577	0,520931	457	774	3,69E-167	27,55659
TEST	1,921589	0,512182	447	761	2,80E-162	27,14627
DETECT	1,39283	0,35671	324	530	7,93E-125	23,76371
CANCER	1,139197	0,297483	265	442	1,78E-99	21,17077
SAMPLE	0,722208	0,14403	168	214	4,18E-92	20,3552
BIOMARKER	0,511564	0,100283	119	149	3,81E-67	17,3121
ANALYSE	1,216576	0,47853	283	711	4,11E-54	15,48915
IMAGE	0,782392	0,255754	182	380	7,60E-49	14,68884
BREAST	0,305219	0,056535	71	84	8,65E-44	13,87766
BLOOD	0,696415	0,239602	162	356	4,43E-40	13,25132
EARLY	0,520162	0,16826	121	250	1,76E-33	12,05823
SCREEN	0,477173	0,149414	111	222	1,92E-32	11,85941
ACCURACY	0,309518	0,074034	72	110	1,64E-31	11,67855
PREDICT	0,567449	0,205277	132	305	3,35E-30	11,41927
⋮	⋮	⋮	⋮	⋮	⋮	⋮
AUTO	0,275127	0,14403	64	214	2,35E-07	5,16959
DEEP	0,15046	0,061246	35	91	2,39E-07	5,165751
INDICT	0,171954	0,07538	40	112	3,34E-07	5,103309
MOLECULE	0,214943	0,104321	50	155	4,28E-07	5,055989
MEASURE	0,434185	0,26585	101	395	5,25E-07	5,016874

Taula completa a l'annex

Tal com indicava el tema de cada un dels pols de les dimensions clarament aquest clúster engloba els projectes sobre detecció i diagnòstic de malalties.

Per realitzar el profiling es mostrarà en la taula 4.24 les 20 característiques amb un v.test més elevat a partir d'aquí es comentarà la taula i és realitzarà la descripció de les propostes que conformen el clúster.

Taula 4.24: Conjunt de característiques més representades en el quart clúster.

	CLA/MOD	MOD/CLA	GLOBAL	P.VALUE	V.TEST
CATEGORY=1	33,59781	26,39498	12,29977	0	Inf
PROGRAMME=DIGITAL SANDBOX	50,94917	7,153297	2,198142	0	Inf
CLC=5	22,57518	17,16963	11,90739	5,18E-147	25,82082
CATEGORY=2	19,29697	38,70261	31,40059	3,62E-146	25,74547
STATUS=1	19,59115	34,23609	27,35967	1,63E-139	25,14433
COUNTRY=ES	22,10153	16,35715	11,58702	1,36E-124	23,74093
GENDER=1	17,38202	64,71069	58,28577	2,48E-105	21,79692
COUNTRY=MC	88,26816	0,679219	0,120474	1,30E-102	21,50837
COUNTRY=HU	30,88347	3,486373	1,767398	2,92E-86	19,68461
FUNDING.STAGE=0	19,3162	23,77698	19,27177	6,84E-77	18,55946
CLC=4	20,10479	14,68059	11,43223	8,22E-61	16,45122
PROGRAMME=GOLD TRACK	24,2965	4,565386	2,94185	5,07E-51	15,02451
COUNTRY=CH	25,22761	3,692718	2,291695	5,38E-48	14,55562
COUNTRY=EE	24,14003	3,408993	2,21093	3,24E-37	12,74699
COUNTRY=SG	72,27723	0,313817	0,067977	1,02E-36	12,65763
COUNTRY=CY	37,02032	0,705012	0,298156	7,43E-28	10,93984
VALUATION=-1	16,68463	51,8958	48,697	2,25E-26	10,62625
COUNTRY=FI	27,45098	1,384232	0,789474	6,47E-25	10,30817

PROGRAMME=INVESTOR NETWORK	21,16753	3,709913	2,743976	4,19E-21	9,427719
PROGRAMME=EUROPEAN HEALTH CATAPULT	22,53467	2,514831	1,747207	2,15E-20	9,254776

En aquest clúster que agrupa els projectes sobre detecció de malalties, es troben el 50% de les propostes inscrites a DigitalSand Box, també estan sobrerepresentats programes com GoldTrack i InvestorNetwork. Per altra banda el 34% dels projectes dins del clúster pertanyen a "BioTech" i el 20% a "MedTech", totes dues categories també sobrerepresentades.

La proporció de projectes seleccionats augmenten, en el global es troba vora el 27% i en aquest grup arriba quasi el 35%. A més a més els projectes tendeixen a estar presentats per homes, augmentant en 6 punts el percentatge respecte al total.

Per últim des del CLC Espanya un gran gruix dels seus projectes presentats es troba en aquest clúster, 23%. Països com Hongria, Suïssa i Estònia gairebé el 25% de les seves propostes estan en aquest grup.

Pel que fa a les variables numèriques no hi ha cap que estigui molt diferenciada respecte a la mitjana global.

Finalment per veure les característiques del clúster en la taula 4.25 es mostraran les descripcions dels projectes que es troben més a prop del centre del clúster. En la taula es pot veure com totes 3 són de la categoria MedTech (2), són projectes seleccionats i presentat per homes.

Taula 4.25: Descripció dels projectes més apropa del baricentre del clúster 4

ID	CARACTERÍSTIQUES	DESCRIPCIÓ
2019-GT1001-3683	Program = Gold Track Gender = 1 Category = 2 Country = EE Status = 1	Selfdiagnostics sees opportunity here to introduce new diagnostics products onto the market to serve patients, doctors and ecosystem stimulating prevention and faster treatment to minimize economical burden of communicable diseases. Has developed a unique proprietary technology to detect pathogeneses' DNA and RNA to recognise diseases at any place at any time which is more non-invasive, more easy to use and accurate comparing other products on the market. [...]
2019-HS-0348	Program = Headstart Gender = 1 Category = 4 Country = DE Status = 1	Our saliva epigenetics-KIT aims at detecting histone modifications, thoroughly, from individuals. Current efforts to develop personalized medicine are hampered by tools that are expensive, non-thorough, with low throughput and unspecific for histone modifications. These tools are unable to function on non-invasive samples and cannot characterize more than one epigenetic mark at a time. [...]
12881	Program = Bootcamp Gender = 1 Category = 2 Country = NL Status= 1	We are developing the next generation of urodynamics, a diagnostic technique that is used to determine the cause of incontinence in a patient. Our solution will be more accurate, quicker and easier to use than existing systems.

4.3.1.5 Clúster 5:

La taula 4.26 mostra les dues dimensions amb major importància en la representació del clúster 5. En aquest cas el pol 3 positiu i el 2 negatiu.

Taula 4.26: Aportació màxima del clúster 5.

	V.TEST	MEAN IN CATEGORY	OVERALL MEAN	P.VALUE
DIM.3	141,7441	0,410211	0,00628	0
DIM.2	-108,853	-0,35128	-0,02944	0

Taula completa a l'annex

Amb l'ajuda de la taula 4.5 es concreten els temes que parlen aquests dos pols, *Implants and Operations i Medical devices.*

La taula 4.27 mostra les paraules amb un v.test superior a 5 en el cinquè clúster.

Taula 4.27:Conjunt de paraules més representades en el cinquè clúster.

	INTERN %	GLOB %	INTERN FREQ	GLOB FREQ	P.VALUE	V.TEST
SURGERY	1,192571	0,222776	244	331	4,41E-135	24,73575
IMPLANT	0,718475	0,118455	147	176	6,33E-96	20,78173
WOUND	0,576735	0,083457	118	124	6,91E-93	20,44319
SURGEON	0,518084	0,089514	106	133	1,96E-65	17,08382
CATHECT	0,386119	0,057208	79	85	3,04E-60	16,37186
TISSUE	0,58651	0,137973	120	205	4,10E-50	14,8854
DEVICE	1,779081	0,866873	364	1288	4,54E-42	13,59082
SKIN	0,625611	0,175663	128	261	5,54E-42	13,57621
BONE	0,288368	0,049132	59	73	1,42E-37	12,81098
HEAL	0,268817	0,045094	55	67	8,68E-36	12,488
PROCEDURE	0,400782	0,093552	82	139	5,34E-35	12,34265
REMOVE	0,28348	0,055862	58	83	6,16E-31	11,56554
MANUFACTURE	0,395894	0,103648	81	154	9,07E-30	11,33238
INVASIVE	0,268817	0,055862	55	83	1,45E-27	10,87882
⋮	⋮	⋮	⋮	⋮	⋮	⋮
SIMUL	0,146628	0,050478	30	75	3,76E-08	5,501841
CARDIAC	0,16129	0,061246	33	91	1,28E-07	5,282228
DECREASE	0,14174	0,050478	29	75	1,56E-07	5,2451
CHEMIC	0,107527	0,032979	22	49	2,53E-07	5,155673
HIGH	0,576735	0,360748	118	536	2,78E-07	5,137628

Taula completa a l'annex

Tal com indicava el tema de cada un dels pols de les dimensions clarament aquest clúster engloba els projectes sobre operacions i tractaments mèdics.

Per realitzar el profiling es mostrarà en la taula 4.28 les 20 característiques amb un v.test més elevat a partir d'aquí es comentarà la taula i és realitzarà la descripció de les propostes que conformen el clúster.

Taula 4.28:Conjunt de característiques més representades en el cinquè clúster.

	CLA/MOD	MOD/CLA	GLOBAL	P.VALUE	V.TEST
CATEGORY=2	35,12164	80,08798	31,40059	0	Inf
STATUS=1	18,70065	37,15543	27,35967	1,64E-238	32,97876
COUNTRY=IE	25,04053	12,8348	7,05815	6,05E-225	32,0184
GENDER=-1	19,58487	28,26979	19,87683	8,62E-214	31,20679
COUNTRY=NL	23,19694	14,24242	8,454705	2,13E-196	29,89734
CLC=1	20,04436	18,1085	12,44044	1,02E-140	25,25427
CLC=6	18,90521	19,31085	14,06582	7,16E-111	22,37335
PROGRAMME=HEADSTART	15,39	60,22483	53,88679	4,51E-86	19,66265
VALUATION=-1	15,37579	54,37439	48,697	1,49E-68	17,49777
FUNDING.STAGE=3	17,60547	14,7263	11,51837	6,00E-51	15,0134
COUNTRY=IL	26,71756	1,88172	0,969848	1,37E-38	12,99133
CATEGORY=5	24,677	1,867058	1,041863	1,49E-30	11,48946
FUNDING.STAGE=1	15,27325	31,1437	28,07915	2,16E-25	10,41311
COUNTRY=IR	70,68966	0,200391	0,039036	8,74E-23	9,825582
PROGRAMME=INVESTOR NETWORK	18,81285	3,748778	2,743976	1,19E-19	9,069817
PROGRAMME=GO GLOBAL INBOUND	24,7205	0,97263	0,541796	1,11E-16	8,292535
COUNTRY=LV	21,53013	1,554252	0,994077	3,55E-16	8,153169
PROGRAMME=MENTORING AND COACHING NETWORK	20,03854	2,033236	1,397227	2,55E-15	7,91132

COUNTRY=ES	15,71794	13,22581	11,58702	8,43E-15	7,760949
VALUATION=2	16,20553	7,614858	6,470588	2,31E-12	7,014062

Aquest conjunt de projectes que parlen sobre operacions i tractaments mèdics, el 80% d'ells són de la categoria "MedTech" i la proporció de propostes seleccionades puja quasi fins a un 40%.

El programa HeadStart és el més freqüent on quasi un 60% dels projectes presentats s'han inscrit en ell. Econòmicament parlant els projectes d'aquest grup es troben en un estat per sobre de l'embrionari, però encara no estan llestos per sortir al mercat (Series A i Seed).

El 25% de les propostes presentades des d'Irlanda i Holanda es troba en aquest grup fent que els CLC dels seus respectius països estiguin molt sobrerepresentats.

Pel que fa a les variables numèriques no hi ha cap que estigui molt diferenciada respecte a la mitjana global.

Finalment per veure les característiques del clúster en la taula 4.28 es mostraran les descripcions dels projectes que es troben més a prop del centre del clúster. En la taula es pot veure com totes 3 són de la categoria MedTech (2), han participat al HeadStart i no passen d'un finançament inicial.

Taula 4.29: Descripció dels projectes més apropa del baricentre del clúster 5

ID	CARACTERÍSTIQUES	DESCRIPCIÓ
2020-HS-0060	Program = Headstart Status = 0 Category = 2 Country = IE Funding Stage = 1	RelEase is a patient-centric drainage technology that will have the capability to accelerate the time to catheter liberation to less than 30 days in patients with advanced cancers. User-centric design that will allow up to 90% of patients to drain themselves and when required based on their clinically- recommended drainage schedule - promoting better quality of life and clinical outcomes [...] Active element designed to shorten the treatment duration and allowing the catheter to be removed in 30 days or less
2020-HS-0250	Program = Headstart Status = 0 Category = 2 Country = NL Funding Stage = 2	This Variable Venous Stent - the WStent (V + V = W) - provides more flexibility, higher radial force and porosity and better positioning with minimal stent material. Advantages are better patency rates and clinical results, better healing and less need for anti-coagulation. Animal studies have shown that less stent material against the vein wall improves ingrowth in te wall.[...] Customers will be vascular surgeons and interventional radiologists treating patient with i.a. deep vein thrombosis and post thrombotic syndrome[...]
2018-HS-0158	Program = Headstart Status = 1 Category = 2 Country = IE Funding Stage = 2	Medical staff must attend to gas bubbles in IV tubing as the patient is at risk from embolism Adverse outcomes due to induced gas include; slow recovery, hypertension, heart arrythmia, heart attack, stroke and death. Clinicians are supported currently by bubble alarms that stop the infusion causing downtime until a clinician expels the gas. [...]

4.3.1.6 Clúster 6:

La taula 4.30 mostra les dues dimensions amb major importància en la representació del clúster 6. En aquest cas el pol 1 positiu i el 8 negatiu.

Taula 4.30: Aportació màxima del clúster 6.

	V.TEST	MEAN IN CATEGORY	OVERALL MEAN	P.VALUE
DIM.1	227,1356	1,096236	0,014765	0
DIM.8	-145,329	-0,56433	-0,00291	0

Taula completa a l'annex

Amb l'ajuda de la taula 4.5 es concreten els temes que parlen aquests dos pols, *Cancer/cell tractament i Drug development*.

La taula 4.31 mostra les paraules amb un v.test superior a 5 en el sisè clúster.

Taula 4.31:Conjunt de paraules més representades en el sisè clúster.

	INTERN %	GLOB %	INTERN FREQ	GLOB FREQ	P.VALUE	V.TEST
CELL	3,468021	0,282003	321	419	1,10E-294	36,68893
DRUG	1,998704	0,282676	185	420	7,04E-107	21,95945
MOLECULE	0,918323	0,104321	85	155	8,13E-60	16,31182
PROTEIN	0,702247	0,074707	65	111	1,52E-48	14,64166
IMMUNE	0,486171	0,039036	45	58	1,43E-42	13,67532
TUMOUR	0,702247	0,090187	65	134	1,30E-41	13,51336
CULTURE	0,486171	0,04644	45	69	5,23E-37	12,70953
BACTERIA	0,583405	0,077399	54	115	7,49E-34	12,12817
COMPOUND	0,432152	0,042401	40	63	2,48E-32	11,83803
THERAPEUTIC	0,723855	0,129223	67	192	5,41E-32	11,77249
VIVO	0,345722	0,031633	32	47	1,49E-27	10,87653
TARGET	1,069576	0,306905	99	456	1,76E-27	10,86121
GENE	0,432152	0,05317	40	79	5,23E-27	10,76148
VITRO	0,45376	0,0599	42	89	1,03E-26	10,69886
⋮	⋮	⋮	⋮	⋮	⋮	⋮
CHEMIC	0,162057	0,032979	15	49	3,37E-07	5,101569
CAUSE	0,334918	0,119128	31	177	3,59E-07	5,089722
PHARMACEUTICS	0,22688	0,06192	21	92	3,59E-07	5,08926
STROKE	0,216076	0,057208	20	85	4,12E-07	5,063252
MODEL	0,669836	0,335846	62	499	4,53E-07	5,045279

Taula completa a l'annex

Tal com indicava el tema de cada un dels pols de les dimensions clarament aquest clúster engloba els projectes sobre càncer i tractaments cel·lulars.

Per realitzar el profiling es mostrarà en la taula 4.32 les 20 característiques amb un v.test més elevat a partir d'aquí es comentarà la taula i és realitzarà la descripció de les propostes que conformen el clúster.

Taula 4.32:Conjunt de característiques més representades en el sisè clúster.

	CLA/MOD	MOD/CLA	GLOBAL	P.VALUE	V.TEST
CATEGORY=1	39,64979	78,28436	12,29977	0	Inf
STATUS=1	8,855871	38,89369	27,35967	1,86E-136	24,86316
PROGRAMME=START-UP MEET PHARMA	14,96834	8,426966	3,507202	2,31E-116	22,93038
FUNDING.STAGE=3	9,746406	18,02074	11,51837	2,08E-80	18,98959
PROGRAMME=GOLD TRACK	13,61245	6,428263	2,94185	2,57E-72	17,98462
CLC=2	9,387821	18,78781	12,46736	4,48E-72	17,95385
COUNTRY=FR	9,310866	18,62576	12,46197	8,04E-69	17,53289
COUNTRY=CH	14,06755	5,175022	2,291695	1,97E-62	16,67568
PROGRAMME=INVESTOR NETWORK	12,65636	5,574762	2,743976	7,40E-53	15,30215
COUNTRY=CY	27,08804	1,296456	0,298156	1,85E-43	13,8232
COUNTRY=CZ	18,9693	1,869058	0,613811	4,38E-39	13,07834
COUNTRY=PL	12,49149	3,964996	1,977386	1,55E-36	12,62432
PROGRAMME=INVESTOR NETWORK	13,99254	2,430856	1,082245	1,84E-29	11,27031
COUNTRY=NO	25	0,821089	0,204604	1,16E-25	10,47173
CLC=3	7,711075	20,82973	16,82797	4,36E-25	10,34617
VALUATION=5	8,413403	11,25756	8,335577	9,06E-24	10,0514
PROGRAMME=INNOSTARS AWARDS	10,71102	3,792135	2,205546	8,85E-23	9,824241
COUNTRY=AT	13,40909	1,912273	0,88841	1,92E-21	9,509165

COUNTRY=SE	8,619366	8,491789	6,137434	1,69E-20	9,280033
COUNTRY=LT	14,07895	1,156007	0,511509	6,28E-15	7,79821

Aquest últim clúster parla sobre tractaments cel·lulars i de càncer, la gran majoria pertanyen a "BioTech" i la ràtio de seleccionades augmenta com en el cas del clúster anterior fins el 40%.

Països com França i Noruega estan sobrerepresentats on el 25% de les propostes Noruegues es troben en aquest grup i el 9% de les Franceses.

De manera remarcable cal destacar la poca presència de propostes per part del CLC Ireland-UK i Belgium Netherland.

Pel que fa a les variables numèriques no hi ha cap que estigui molt diferenciada respecte a la mitjana global.

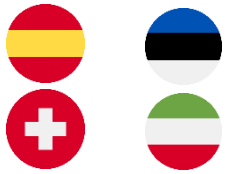
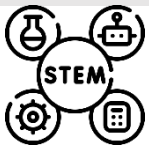
Finalment per veure les característiques del clúster en la taula 4.33 es mostraran les descripcions dels projectes que es troben més a prop del centre del clúster. En la taula es pot veure com totes 3 són de la categoria BioTech (1) i dues de les propostes van ser franceses.

Taula 4.33: Descripció dels projectes més apropa del baricentre del clúster 6

ID	CARACTERÍSTIQUES	DESCRIPCIÓ
2019 BRIDGEHEAD 1001 – 3719	Program = Bridgehead Status = 0 Category = 1 Country = FR Funding Stage = 3	ukarys has developed a[...]system called C3P3, which autonomously synthesizes target messenger RNA in host cells and therefore, the protein of interest at a high rate. [...] called synthetic gene therapy, designed to address the issues of safety, efficacy, tolerance, and cost of production of all existing technologies. Synthetic gene can be used for treatment of monogenic and multifactorial human diseases and has demonstrated its efficacy in animals studies with our first treatments,[...] new cellular tools for the bioproduction of recombinant proteins and viruses.
2018-BCAMP-0463	Program = Bootcamp Status = 1 Category = 1 Country = SE Funding Stage = 0	[...] Our approach will be to utilize novel disease-coupled orphan G Protein-Coupled Receptors (GPCRs) as drug targets in multiple diseases pathways were we are targeting GPCRs. The underlying technology of cell screening assays of constitutively active orphan receptors found in the brain that have very strongly correlated to neurodegenerative diseases is the medicinal approach. [...]
2020-HS-0068	Program = Headstart Status = 0 Category = 2 Country = FR Funding Stage = 1	In Sickle Cell Disease, due to a mutation in the adult globin, fetal hemoglobin (HbF) is a crucial protein, since it moderates its severity by acting on its primary cause in a dose-dependent way. [...] As HbF expression is not distributed homogeneously, HbF level is not a relevant biomarker for the disease. The method we developed allows the quantification of HbF in each cell, making possible to determine protective thresholds of HbF and the proportion of cell containing these thresholds. [...]This product will significantly change the management of patients.

4.3.2 Resum dels Clústers

CLÚSTER 1	CLÚSTER 2	CLÚSTER 3	CLÚSTER 4	CLÚSTER 5	CLÚSTER 6
Salut mental	***EHealth***	**Productes a punt de sortir al mercat**	**Detecció i diagnòstic de malalties**	**Operacions i tractaments mèdics**	**Tractaments cel·lulars**
Projectes sobre Digital Health	Digital Health és la categoria principal		Categories dominats Bio Tech i Med Tech	Med Tech	Bio Tech
20% de les propostes inscrites en algun Bootcamp	Headstarts i Starup Pharma amb una gran representació	Projectes que han participat a Starup Rescue	50 dels de sand box es troben en aquest clúster	60% del clúster a HeadStart	
La ràtio de rebutjades és superior	Rebutjades		La ràtio de seleccionats augmenta	La ràtio de seleccionats augmenta	La ràtio de seleccionats augmenta
Romania, Txèquia i Eslovàquia presenten alt % de les seves propostes en aquest clúster igual GB amb 1 de cada 4 de les seves	El 40 % dels projectes presentat des del CLC Innostars ens troben en aquest clúster	Un gran nombre de les propostes presentades per països de fora la UE es troben en aquest clúster	Des del CLC Spain presenten el 20% de projectes en aquest clúster. Hongria, Suïssa i Estonià es troben molt representats	25% de les propostes de Irlanda i Països Baixos aquí	El 25% de propostes per part de Noruega es troba en aquest Clúster i el 9% de les franceses
Son projectes en estat molt Embrionari (Pre-Seed)	Projectes més madurs econòmicament	Projectes a punt de sortir al mercat		Estat embrionari	
Projectes presentats majoritàriament per dones	TRL elevat	Presentats en els darrers anys	Projectes presentats majoritàriament per homes		

CLÚSTER 1	CLÚSTER 2	CLÚSTER 3	CLÚSTER 4	CLÚSTER 5	CLÚSTER 6
Salut mental	***EHealth***	**Productes a punt de sortir al mercat**	**Detecció i diagnòstic de malalties**	**Operacions i tractaments mèdics**	**Tractaments cel·lulars**
			 		
20% de les propostes inscrites en algun Bootcamp	Headstart i Start-ups Meet Pharma amb una gran representació	Projectes que han participat a Start-up Rescue	50 dels de Digital Sandbox es troben en aquest clúster	60% del clúster a HeadStart	20% de les propostes inscrites en algun Bootcamp
					
					
					

5 CONCLUSIONS

Un cop finalitzat l'estudi sobre les descripcions dels projectes presentats a EIT Health cal recapitular i fer una valoració sobre aquest.

En primer lloc es volia demostrar que l'Anàlisi de Correspondències (CA) combinat amb una classificació jeràrquica podria ser de gran utilitat a l'hora de tractar amb dades textuals, determinant temaris, classificant els objectes i descrivint les seves característiques. Al llarg d'aquest treball s'ha demostrat amb escreix que aquests dos mètodes combinats són de gran utilitat, permeten trobar relacions i descriure les diferents característiques de les dades textuals. S'ha vist que entre les propostes presentades hi ha 6 grans grups: projectes sobre salut mental, d'EHealth, productes a punt de sortir al mercat, sobre detecció i diagnòstic de malalties, projectes d'operacions i tractaments mèdics i per últim sobre tractaments cel·lulars. D'aquesta manera tan clara es confirma la hipòtesi inicial i dóna lloc a que grans quantitats de dades textuals puguin ser processades i analitzades amb certa facilitat.

D'altra banda, el segon objectiu era agrupar diferents projectes de EIT Health sota característiques comunes i tipologia similar, amb l'ajuda del clúster i la seva agrupació jeràrquica també s'ha pogut dur a terme aquest objectiu. De manera similar al paràgraf anterior, s'han determinat 6 grups de projectes amb diferents característiques comunes. El primer grup engloba aquells projectes sobre salut mental com a característiques més destacables s'ha vist que té una ràtio de rebutjades superior a la mitjana, països com Romania, Txèquia, Eslovàquia i Regne Unit presenten un alt percentatge de les seves propostes en aquest grup i finalment, els projectes son proposats majoritàriament per dones. El segon clúster té certa relació amb el primer, té una ràtio de rebutjades més elevada que en el global, el CLC Innostars presenta un gran número de les seves propostes en aquest clúster i tenen un pressupost superior a la mitjana. A continuació, el següent clúster que engloba les propostes de productes a punt de sortir al mercat, tal com indica la seva descripció són projectes amb un bon pressupost que busquen en EIT Health un impuls definitiu per vendre el seu producte; la gran majoria de propostes vénen de països de fora de la UE. El quart grup agrupa els projectes que tracten sobre detecció i diagnòstic de malalties, la ràtio de seleccionats augmenta notablement i el vint per cent de les propostes espanyoles es troben en aquest grup, normalment són projectes presentats per homes. El següent clúster amb característiques similars a l'anterior engloba els projectes d'operacions i tractaments mèdics, la ràtio de seleccionats augmenta tot i que en aquest clúster els projectes estan en un estat més embrionari. Finalment, el sisè clúster agrupa les propostes sobre tractaments cel·lulars, de la mateixa manera que en els dos anteriors, el percentatge de seleccionades és superior a la mitjana. Països com Noruega i França tenen una bona part dels projectes en aquest grup.

A nivell personal aquest estudi m'ha semblat molt enriquidor, ja que he pogut descobrir i conèixer com tractar amb dades textuals. Durant el transcurs d'aquest, sense cap mena de dubte, la part més feixuga i complexa ha sigut el processament de les dades i trobar quin format necessitava cada software. A més a més, al tractar-se de dades textuals no estava familiaritzat amb la metodologia a seguir per processar-les, fet que ha dificultat encara més la feina. Amb la dificultat afegida de ser un text recollit sense cap tractament previ ha fet que tot fos més lent i dens, per exemple una de les problemàtiques més importants va ser com eliminar les "", ja que al llenguatge R li costa molt tractar aquest tipus de signe de puntuació. Tal com he dit amb anterioritat aquest treball m'ha obert una nova línia d'aprenentatge que trobo molt interessat i m'agradaria seguir aprofundint en ella en un futur.

Un cop explicats els aspectes més concrets sobre el treball cal reflexionar sobre les limitacions d'aquest i possibles actuacions a futur. Pel que fa a les limitacions cal ser conscients que la base de dades del treball no era la més idònia, ja que la manera com s'han recollit les dades no ha sigut homogènia al llarg del temps, fet que ha implicat que les propostes del 2016 no hagin pogut ser estudiades, no tenien el camp *"project description"*. A més a més el llindar de freqüència mínima és totalment subjectiu i modificar aquest podria haver fet que els resultats fossin diferents, de fet si es canvia d'un llindar de l'1% a un 5% es passa de 870 paraules a únicament 160.

D'altra banda, la continuació d'aquest estudi es pot dividir en dos blocs, un primer que seria continuar analitzant les propostes i un segon on es definirien possibles accions a fer des de EIT Health. Per seguir amb l'estudi seria possible fer una anàlisi factorial múltiple de les taules de contingència per determinar l'evolució temporal de les característiques dels projectes al llarg del temps, fet que podria ser molt interessant en el moment que la línia temporal fos una mica més àmplia (deu anys), també podria ser d'interès dissenyar un petit algoritme per classificar directament aquelles propostes presentades en funció de la descripció dels projectes, que automatitzaria el procés de classificació. Des de EIT Health amb la informació presentada en aquest estudi podrien optimitzar els seus recursos de manera més eficient com per exemple organitzar un nou programa per projectes de Salut Mental posant la seu al CLC *Ireland-UK*, també podrien dissenyar o impulsar més recerca en temes que no s'han detectat en les *metakeys* per trobar noves maneres d'innovar.

6 BIBLIOGRAFIA

Per consultar el codi utilitzat: arnaureynals/TFG: Codi en R: Anàlisi de Correspondències i Clúster jeràrquic per dades textuals (github.com)

[1] Public Spending on Health: A Closer Look at Global Trends, HO/HIS/HGF/HFWorkingPaper/18.3,

[2] Vairaprakash Gurusamy, and Subbu Kannan, "Preprocessing Techniques for Text Mining", Conference Paper, October 2014

[3] Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.

[4] Porter, M.F. (1980), "An algorithm for suffix stripping", Program: electronic library and information systems, Vol. 14 No. 3, pp. 130-137. https://doi.org/10.1108/eb046814

[5] Lebart, L., Salem, A., Bécue M. (2000). Análisis estadístico de textos. Milenio.

[6] Yule, G. U. 1944. The Statistical Study of Literary Vocabulary. Cambridge University Press.

[7] Statistical Analysis of Textual Data: Benzécri and the French School of Data Analysis

[8] ROBERT GIBRAT L'analyse des données Journal de la société statistique de Paris, tome 119, no 3 (1978), p. 201-228

[9] Coatrieux, JL.,Bansard, JY.,Kerbaol, M. (2004) An analysis of IEEE publications. IEEE Engineering in Medicine and Biology Magazine.

[10] NTENSIVE USE OF FACTORIAL CORRESPONDENCE ANALYSIS FOR TEXT MINING: APPLICATION WITH STATISTICAL EDUCATION PUBLICATIONS Annie Morin IRISA, Université de Rennes 1, France

[11] Romesburg, H.C. (1984). Cluster Analysis for researchers. Lifetime Learning Publications

[12] Fundamentos en Humanidades Universidad Nacional de San Luis – Argentina Año XI – Número II (22/2010) 75/87 pp; Del análisis textual al análisis multidimensiona

Becue-Bertaut, M. and Pages, J. (2004). A principal axes method for comparing múltiple contingency tables: MFACT. Computational Statistics and Data Analysis, 45:481–503.

Becue-Bertaut, M. and Pages, J. (2008). Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. Computational Statistics and Data Analysis,52:3255–3268.

Becue-Bertaut, M. and Pages, J. (2014). Correspondence analysis of textual data involvingcontextual information: CA-GALT on principal components. Advances in Data Analysis and Classification.

Becue-Bertaut, M., Pages, J., and Kostov, B. (2014). Untangling the influence of several contextual variables on the respondents' lexical choices. a statistical approach. SORT, 38:285–302.

Benzecri, J. P. (1981). Pratique de l'analyse des donnees: Linguistique & lexicologie, volume 3. Dunod.

Benzecri, J. P. (1983). Analyse de l'inertie intraclasse par l'analyse d'un tableau de contingence. Les Cahiers de l'Analyse des Donnees, 8(3):351–358.

Escofier, B. and Drouet, D. (1983). Analyse des differences entre plusieurs tableaux de frequence. Les Cahiers de l'Analyse des Donnees, 8(4):491–499.

Escofier, B. and Pages, J. (1988). Analyses factorielles simples et multiples. Dunod.

Greenacre, M. (1984). Theory and Applications of Correspondence Analysis. Academic Press.

Kostov, B., Becue-Bertaut, M., and Husson, F. (2015). Correspondence analysis on generalised aggregated lexical table (CA-GALT) in the FactoMineR package. The R Journal.

Lebart, L., Morineau, A., and Piron, M. (1997). Statistique exploratoire multidimensionnelle.

Lebart, L., Salem, A., and Berry, L. (1998). Exploring Textual Data. Kluwer Academic

Publishers.

Murtagh, F. (2005). Correspondence Analysis and Data Coding with Java and R. Chapman

& Hall / CRC Press.

Husson F., Lê S. & Pagès J. (2017) Exploratory Multivariate Analysis by Example Using R 2nd edition, 230 p., CRC/Press

7 ANNEX

Taula 4.5

4

Mental health	Surgery
cell	market
therapy	devic
disease	medic
nutritive	surgery
behaviour	surgeon
game	clinic
cognit	launch
people	sale
mental	hospital
exercise	implant
protein	product
stress	trial
active	fund
immune	operation
cancer	print
breath	
anxiety	
healthy	
children	
therapeutic	
molecule	
person	
tumour	
coach	
brain	
emote	

7.1.1.1 Taula 4.8

	V.TEST	MEAN IN CATEGORY	OVERALL MEAN	P.VALUE
DIM.4	183,9657	0,458703	0,007151	0
DIM.6	114,5025	0,270978	0,014412	0
DIM.3	109,744	0,277505	0,00628	0
DIM.7	47,76222	0,101068	-0,00455	0
DIM.5	41,16486	0,110935	0,016235	0
DIM.8	38,40311	0,080224	-0,00291	0

DIM.2	32,40246	0,053643	-0,02944	2,53E-230
DIM.9	-28,9162	-0,05804	0,003972	7,47E-184
DIM.1	-135,668	-0,34724	0,014765	0

7.1.1.2 Taula 4.9

	INTERN %	GLOB %	INTERN FREQ	GLOB FREQ	P.VALUE	V.TEST
MENTAL	0,586863	0,124512	156	185	7,79E-86	19,63485
BEHAVIOUR	0,601911	0,142684	160	212	1,77E-74	18,25842
EXERCISE	0,519148	0,115763	138	172	1,52E-70	17,75703
GAME	0,4251	0,08413	113	125	9,90E-70	17,65158
COGNIT	0,511624	0,116436	136	173	1,88E-67	17,35289
PEOPLE	0,951772	0,339884	253	505	1,13E-60	16,43176
THERAPIST	0,364909	0,072015	97	107	2,79E-60	16,37705
NUTRITIVE	0,432624	0,102302	115	152	4,50E-54	15,48322
APP	1,124821	0,482568	299	717	8,47E-50	14,83678
USER	1,286585	0,592947	342	881	2,51E-48	14,60773
COACH	0,285908	0,058554	76	87	9,64E-45	14,03411
CHILDREN	0,368671	0,093552	98	139	1,05E-41	13,52936
HABIT	0,259574	0,052497	69	78	1,60E-41	13,49811
STRESS	0,353623	0,089514	94	133	3,37E-40	13,27188
ANXIETY	0,237002	0,047113	63	70	4,80E-39	13,07131
PERSON	0,869009	0,377574	231	561	9,91E-38	12,83904
FOOD	0,300956	0,078745	80	117	9,12E-33	11,92172

PHYSIC	0,379956	0,117109	101	174	5,96E-32	11,7643
MOTIVE	0,214431	0,04644	57	69	4,16E-31	11,59918
EMPLOY	0,300956	0,082111	80	122	8,44E-31	11,53849
EMOTE	0,180573	0,036344	48	54	2,10E-29	11,25849
LIFESTYLE	0,274622	0,074034	73	110	1,02E-28	11,11837
PSYCHOLOGY	0,180573	0,037017	48	55	1,36E-28	11,09254
HEALTHY	0,334813	0,102975	89	153	2,10E-28	11,05379
REHABILITEE	0,304717	0,088841	81	132	3,33E-28	11,01228
SLEEP	0,319765	0,097591	85	145	1,62E-27	10,86878
THERAPY	0,632007	0,278638	168	414	5,99E-27	10,74908
REALITY	0,248288	0,065958	66	98	1,20E-26	10,68504
ACTIVE	0,665864	0,31902	177	474	1,99E-23	9,973714
PARENT	0,176811	0,041728	47	62	7,53E-23	9,840502
TRAIN	0,470243	0,205277	125	305	7,89E-21	9,361162
DEMENTIA	0,180573	0,04644	48	69	1,23E-20	9,313994
DISORDER	0,237002	0,073361	63	109	2,85E-20	9,224375
HELP	0,759913	0,406515	202	604	5,97E-20	9,144948
WELLB	0,15424	0,037017	41	55	1,32E-19	9,058315
LIFE	0,432624	0,191143	115	284	8,74E-19	8,850196
BREATH	0,195621	0,0599	52	89	3,17E-17	8,440196
HEALTH	1,384395	0,914659	368	1359	5,53E-17	8,374757
INDIVIDUAL	0,417576	0,19047	111	283	5,72E-17	8,370923

CHANGE	0,372432	0,170952	99	254	4,56E-15	7,83846
VIRTUAL	0,244526	0,092879	65	138	7,05E-15	7,783538
EDUCATE	0,229479	0,085476	61	127	1,58E-14	7,681133
PLAY	0,142954	0,040382	38	60	1,65E-14	7,675304
DAILY	0,221955	0,082111	59	122	2,92E-14	7,6016
SESSION	0,13543	0,039036	36	58	2,07E-13	7,344386
PROGRAMME	0,161764	0,052497	43	78	3,16E-13	7,287336
MOVEMENT	0,218193	0,08413	58	125	4,67E-13	7,234694
STIMULUS	0,199383	0,074034	53	110	7,49E-13	7,170269
BRAIN	0,278384	0,124512	74	185	3,11E-12	6,972633
IMPAIR	0,127906	0,038363	34	57	5,10E-12	6,902861
SKILL	0,139192	0,045094	37	67	1,43E-11	6,755139
RECOMMEND	0,240764	0,104994	64	156	2,59E-11	6,668058
DISABLE	0,13543	0,044421	36	66	4,43E-11	6,588838
WEIGHT	0,11662	0,034998	31	52	4,78E-11	6,57758
SERIOUS	0,112858	0,033652	30	50	7,97E-11	6,501112
MOBILE	0,451433	0,258447	120	384	3,06E-10	6,295577
ENGAGE	0,214431	0,093552	57	139	3,36E-10	6,281102
PERSONALISE	0,173049	0,06865	46	102	4,36E-10	6,240592
SOCIAL	0,206907	0,089514	55	133	4,76E-10	6,226687
FEEL	0,112858	0,035671	30	53	5,98E-10	6,190838
MUSCLE	0,120382	0,039709	32	59	6,89E-10	6,168631

TRACK	0,304717	0,155472	81	231	7,21E-10	6,161447
FAMILY	0,203145	0,088841	54	132	1,11E-09	6,092606
PROGRAM	0,248288	0,118455	66	176	1,24E-09	6,074492
LIVE	0,285908	0,146722	76	218	3,31E-09	5,915321
PROGRESS	0,206907	0,096917	55	144	1,49E-08	5,663055
CREATE	0,443909	0,277292	118	412	1,13E-07	5,305205
WORK	0,462719	0,293445	123	436	1,51E-07	5,251438
SITUATE	0,142954	0,061246	38	91	1,95E-07	5,204069
INTERACT	0,267098	0,147395	71	219	3,08E-07	5,118233
CONTENT	0,161764	0,074707	43	111	3,82E-07	5,077489

7.1.1.3 Taula 4.12

	V.TEST	MEAN IN CATEGORY	OVERALL MEAN	P.VALUE
DIM.9	43,19968	0,059531	0,003972	0
DIM.2	-24,4704	-0,06707	-0,02944	3,05E-132
DIM.7	-32,6304	-0,04783	-0,00455	1,52E-233
DIM.3	-68,6759	-0,0955	0,00628	0
DIM.5	-70,232	-0,08065	0,016235	0
DIM.8	-79,735	-0,10643	-0,00291	0
DIM.4	-97,4257	-0,13625	0,007151	0
DIM.6	-110,926	-0,13464	0,014412	0
DIM.1	-160,772	-0,24249	0,014765	0

7.1.1.4 Taula 4.13

	INTERN %	GLOB %	INTERN FREQ	GLOB FREQ	P.VALUE	V.TEST
PATIENT	3,332417	2,29506	1940	3410	8,89E-99	21,09475
DATA	1,690257	1,01023	984	1501	3,54E-95	20,69899
CARE	1,135427	0,648809	661	964	4,06E-76	18,46348
DOCTOR	0,688814	0,348634	401	518	3,27E-70	17,71404
MEDIC	1,48241	0,959752	863	1426	1,00E-59	16,2992
SERVICE	0,699121	0,382286	407	568	2,94E-55	15,65777
HEALTHCARE	0,680225	0,38767	396	576	1,51E-46	14,32559
PLATFORM	1,015185	0,652847	591	970	1,56E-42	13,66894
MANGA	0,611516	0,349307	356	519	1,23E-41	13,51786
MONITOR	0,812491	0,508817	473	756	2,01E-38	12,96215
INFORM	0,632129	0,375555	368	558	3,45E-37	12,74221
HEALTH	1,286588	0,914659	749	1359	1,43E-32	11,88418
HOSPITAL	0,673354	0,426033	392	633	9,98E-31	11,52407
REMOTE	0,28858	0,145376	168	216	1,32E-30	11,49972
PROVIDE	1,240209	0,896487	722	1332	1,33E-28	11,09483
SECURE	0,242201	0,123839	141	184	9,82E-25	10,26802
ACCESS	0,496427	0,309598	289	460	2,07E-24	10,19561
NURSE	0,249072	0,130569	145	194	1,68E-23	9,99035
COMMUNICATE	0,267968	0,151434	156	225	1,07E-19	9,081755
CONNECT	0,386492	0,240275	225	357	1,52E-19	9,043267
CLOUD	0,214718	0,11509	125	171	4,73E-19	8,918367

RECORD	0,247355	0,139319	144	207	1,73E-18	8,77348
PROFESSION	0,346984	0,215372	202	320	8,61E-18	8,591195
SHARE	0,185516	0,101629	108	151	1,63E-15	7,966943
NETWORK	0,211282	0,121147	123	180	4,84E-15	7,831046
EMERGE	0,142572	0,073361	83	109	9,02E-15	7,752429
WEB	0,152879	0,082784	89	123	1,77E-13	7,364994
DIGIT	0,529064	0,382959	308	569	7,30E-13	7,173667
APPOINT	0,077298	0,034325	45	51	9,62E-13	7,135847
VITAL	0,103064	0,050478	60	75	1,18E-12	7,107514
TELEMEDICINE	0,089323	0,042401	52	63	4,09E-12	6,933985
HOME	0,398516	0,277965	232	413	4,21E-12	6,929808
MOBILE	0,374468	0,258447	218	384	4,75E-12	6,91281
ALERT	0,120242	0,063266	70	94	7,89E-12	6,840526
DOCUMENT	0,128831	0,069323	75	103	9,13E-12	6,819544
CONSULT	0,127113	0,069996	74	104	7,95E-11	6,501624
COVID	0,297169	0,201238	173	299	1,02E-10	6,464102
INTEGER	0,376185	0,267869	219	398	2,29E-10	6,340235
SOFTWARE	0,436306	0,319693	254	475	3,91E-10	6,257699
REPORT	0,226742	0,147395	132	219	4,60E-10	6,232016
SAA	0,075581	0,036344	44	54	4,77E-10	6,226485
INSURE	0,135702	0,078072	79	116	5,60E-10	6,201395
GUIDELINE	0,082452	0,041055	48	61	7,16E-10	6,162491

WEARABLE	0,226742	0,148741	132	221	1,08E-09	6,097744
SOLUTE	0,99629	0,818414	580	1216	1,75E-09	6,019661
COMMUNITY	0,12196	0,069323	71	103	1,92E-09	6,00464
ELDER	0,14429	0,086149	84	128	2,66E-09	5,951516
AUTHOR	0,073863	0,036344	43	54	3,02E-09	5,930254
PRIVATE	0,123677	0,071342	72	106	4,06E-09	5,881799
VIA	0,240484	0,162202	140	241	4,56E-09	5,862488
CAREGIVER	0,127113	0,074034	74	110	4,75E-09	5,855691
BOOK	0,066992	0,032306	39	48	5,97E-09	5,817546
REALTIME	0,171774	0,108359	100	161	6,94E-09	5,792312
QUALITY	0,322935	0,232871	188	346	1,64E-08	5,646165
DECISION	0,194105	0,127204	113	189	1,65E-08	5,644767
ETC	0,176927	0,113743	103	169	1,77E-08	5,632803
VISIT	0,089323	0,048459	52	72	2,70E-08	5,560031
TEAM	0,178645	0,115763	104	172	2,75E-08	5,556673
DASHBOARD	0,070427	0,035671	41	53	3,43E-08	5,517779
SMART	0,231895	0,15951	135	237	4,68E-08	5,463244
TRANSPAR	0,066992	0,033652	39	50	5,11E-08	5,447513
APP	0,606363	0,482568	353	717	5,68E-08	5,42866
ECG	0,115089	0,067977	67	101	6,27E-08	5,410945
UPDATE	0,066992	0,034325	39	51	1,35E-07	5,272483
INTERFACE	0,125395	0,076726	73	114	1,43E-07	5,261767

CONTINUE	0,202693	0,139992	118	208	4,60E-07	5,042463

7.1.1.5 Taula 4.17

	V.TEST	MEAN IN CATEGORY	OVERALL MEAN	P.VALUE
DIM.5	158,2074	0,552667	0,016235	0
DIM.2	156,0967	0,560482	-0,02944	0
DIM.3	71,51989	0,266798	0,00628	0
DIM.1	32,93481	0,144289	0,014765	6,98E-238
DIM.6	18,06744	0,074081	0,014412	5,75E-73
DIM.7	9,65274	0,026908	-0,00455	4,79E-22
DIM.8	6,054119	0,016403	-0,00291	1,41E-09
DIM.9	-7,27714	-0,01903	0,003972	3,41E-13
DIM.4	-145,46	-0,51908	0,007151	0

7.1.1.6 Taula 4.18

	INTERN %	GLOB %	INTERN FREQ	GLOB FREQ	P.VALUE	V.TEST
DEVIC	0,944095	0,07538	102	112	2,66E-103	21,58182
MARKET	2,212144	0,467761	239	695	2,14E-95	20,7231
PRODUCT	2,665679	0,88841	288	1320	1,28E-63	16,83835
CLINIC	2,054795	0,697941	222	1037	7,87E-48	14,52958
LAUNCH	0,66642	0,086149	72	128	2,47E-47	14,45115
FUND	0,573862	0,063266	62	94	5,54E-47	14,39529
TRIAL	0,953351	0,187105	103	278	3,23E-45	14,11141
TECHNOLOG	0,472047	0,044421	51	66	1,40E-44	14,00777
PROJECT	1,092188	0,275273	118	409	5,18E-39	13,06563
EIT	0,490559	0,057881	53	86	4,73E-38	12,89623

VALID	0,740466	0,183739	80	273	2,25E-27	10,83909
FIRST	1,1107	0,393054	120	584	8,65E-25	10,28023
SALE	0,453536	0,091533	49	136	1,43E-21	9,540095
INVEST	0,305442	0,041728	33	62	2,57E-21	9,479037
EXPAND	0,351722	0,06192	38	92	2,06E-19	9,00997
SUCCESS	0,425768	0,091533	46	136	4,77E-19	8,917537
PILOT	0,286931	0,043074	31	64	1,59E-18	8,783231
DEVELOP	2,027027	1,095033	219	1627	3,24E-18	8,702845
FINAL	0,360977	0,070669	39	105	5,86E-18	8,635232
PARTNER	0,379489	0,080092	41	119	1,92E-17	8,498412
SERI	0,231396	0,032306	25	48	3,98E-16	8,139218
SCALE	0,388745	0,093552	42	139	1,67E-15	7,963481
NOW	0,379489	0,09086	41	135	3,00E-15	7,890874
PLAN	0,610885	0,220756	66	328	9,09E-14	7,453489
EUROPE	0,490559	0,157491	53	234	2,00E-13	7,349086
STRATEGI	0,314698	0,074034	34	110	4,23E-13	7,248137
NEW	0,77749	0,339884	84	505	2,60E-12	6,997856
START	0,379489	0,109705	41	163	3,19E-12	6,969107
NEXT	0,305442	0,07538	33	112	4,20E-12	6,930335
VERSION	0,259163	0,055862	28	83	4,91E-12	6,908283
WORK	0,675676	0,293445	73	436	5,31E-11	6,561856
WANT	0,44428	0,154799	48	230	6,22E-11	6,53835

MARK	0,194372	0,038363	21	57	3,95E-10	6,255914
COUNTRY	0,277675	0,078072	30	116	1,41E-09	6,054739
ACCELER	0,194372	0,041055	21	61	1,70E-09	6,024491
CUSTOM	0,66642	0,31902	72	474	5,87E-09	5,820441
CONDUCT	0,194372	0,043747	21	65	6,36E-09	5,80696
EXPECT	0,240652	0,064612	26	96	6,46E-09	5,804351
SLEEP	0,305442	0,097591	33	145	7,10E-09	5,78852
PHASE	0,268419	0,085476	29	127	5,64E-08	5,429835
DELAY	0,157349	0,036344	17	54	3,02E-07	5,121896
POSIT	0,342466	0,135281	37	201	3,27E-07	5,107404
MONTH	0,286931	0,103648	31	154	4,11E-07	5,063861
GOAL	0,286931	0,104321	31	155	4,78E-07	5,034982

7.1.1.7 Taula 4.22

	V.TEST	MEAN IN CATEGORY	OVERALL MEAN	P.VALUE
DIM.8	108,7992	0,242184	-0,00291	0
DIM.1	84,01505	0,248042	0,014765	0
DIM.5	61,49061	0,163438	0,016235	0
DIM.4	39,58509	0,108259	0,007151	0
DIM.6	37,4061	0,101631	0,014412	3,11E-306
DIM.9	20,93604	0,050695	0,003972	2,52E-97
DIM.7	-18,5767	-0,0473	-0,00455	4,96E-77
DIM.2	-72,3232	-0,22242	-0,02944	0
DIM.3	-213,613	-0,54308	0,00628	0

7.1.1.8 Taula 4.23

	INTERN %	GLOB %	INTERN FREQ	GLOB FREQ	P.VALUE	V.TEST
DIAGNOSE	1,964577	0,520931	457	774	3,69E-167	27,55659
TEST	1,921589	0,512182	447	761	2,80E-162	27,14627
DETECT	1,39283	0,35671	324	530	7,93E-125	23,76371
CANCER	1,139197	0,297483	265	442	1,78E-99	21,17077
SAMPLE	0,722208	0,14403	168	214	4,18E-92	20,3552
BIOMARKER	0,511564	0,100283	119	149	3,81E-67	17,3121
ANALYSE	1,216576	0,47853	283	711	4,11E-54	15,48915
IMAGE	0,782392	0,255754	182	380	7,60E-49	14,68884
BREAST	0,305219	0,056535	71	84	8,65E-44	13,87766
BLOOD	0,696415	0,239602	162	356	4,43E-40	13,25132
EARLY	0,520162	0,16826	121	250	1,76E-33	12,05823
SCREEN	0,477173	0,149414	111	222	1,92E-32	11,85941
ACCURACY	0,309518	0,074034	72	110	1,64E-31	11,67855
PREDICT	0,567449	0,205277	132	305	3,35E-30	11,41927
ACCRUE	0,399794	0,119801	93	178	2,73E-29	11,23551
METHOD	0,395495	0,142684	92	212	1,49E-21	9,535486
BIOPSY	0,18915	0,042401	44	63	1,87E-21	9,511885
RESULT	0,631932	0,291425	147	433	9,62E-21	9,340126
URINE	0,193449	0,045767	45	68	1,97E-20	9,263937
LABORATORY	0,253633	0,072688	59	108	2,58E-20	9,235099
GENETIC	0,275127	0,083457	64	124	3,45E-20	9,203969

NONINVASIVE	0,266529	0,082111	62	122	3,76E-19	8,943845
LAB	0,270828	0,084803	63	126	5,69E-19	8,897998
ASSAY	0,180552	0,044421	42	66	3,43E-18	8,696357
DISEASE	0,79099	0,419976	184	624	4,45E-18	8,666597
ALGORITHM	0,627633	0,306905	146	456	5,33E-18	8,646103
SENSITIVE	0,206345	0,056535	48	84	7,76E-18	8,603095
EXAMINE	0,154759	0,038363	36	57	1,34E-15	7,991244
IDENTIFY	0,438483	0,203258	102	302	1,24E-14	7,711866
LUNG	0,193449	0,059227	45	88	2,31E-14	7,63228
SEQUENCE	0,146161	0,039709	34	59	4,05E-13	7,253885
ULTRASOUND	0,146161	0,041728	34	62	2,86E-12	6,984517
SCORE	0,154759	0,04644	36	69	4,62E-12	6,916859
RISK	0,576047	0,319693	134	475	5,53E-12	6,891387
STAGE	0,257931	0,111051	60	165	1,41E-10	6,414524
PATHOLOGY	0,202046	0,077399	47	115	1,60E-10	6,395756
KIT	0,176253	0,063266	41	94	2,25E-10	6,343357
DNA	0,146161	0,047113	34	70	2,40E-10	6,332996
RELIABLE	0,197747	0,078745	46	117	1,17E-09	6,084613
RAPID	0,163357	0,06192	38	92	6,91E-09	5,792977
ABNORMAL	0,098874	0,028941	23	43	2,25E-08	5,591915
HIGH	0,576047	0,360748	134	536	3,01E-08	5,54081
TECHNIQUE	0,219242	0,101629	51	151	6,17E-08	5,413686

PROPRIETARY	0,180552	0,078072	42	116	1,06E-07	5,315455
PRECIS	0,202046	0,093552	47	139	2,01E-07	5,198103
AUTO	0,275127	0,14403	64	214	2,35E-07	5,16959
DEEP	0,15046	0,061246	35	91	2,39E-07	5,165751
INDICT	0,171954	0,07538	40	112	3,34E-07	5,103309
MOLECULE	0,214943	0,104321	50	155	4,28E-07	5,055989
MEASURE	0,434185	0,26585	101	395	5,25E-07	5,016874

7.1.1.9 Taula 4.26

	V.TEST	MEAN IN CATEGORY	OVERALL MEAN	P.VALUE
DIM.3	141,7441	0,410211	0,00628	0
DIM.1	97,96349	0,316169	0,014765	0
DIM.8	49,63033	0,120975	-0,00291	0
DIM.6	5,243056	0,027959	0,014412	1,58E-07
DIM.7	-31,1942	-0,0841	-0,00455	1,28E-213
DIM.4	-70,4934	-0,19236	0,007151	0
DIM.9	-79,5259	-0,19269	0,003972	0
DIM.5	-87,1645	-0,21498	0,016235	0
DIM.2	-108,853	-0,35128	-0,02944	0

7.1.1.10 Taula 4.27

	INTERN %	GLOB %	INTERN FREQ	GLOB FREQ	P.VALUE	V.TEST
SURGERY	1,192571	0,222776	244	331	4,41E-135	24,73575
IMPLANT	0,718475	0,118455	147	176	6,33E-96	20,78173
WOUND	0,576735	0,083457	118	124	6,91E-93	20,44319
SURGEON	0,518084	0,089514	106	133	1,96E-65	17,08382
CATHECT	0,386119	0,057208	79	85	3,04E-60	16,37186

TISSUE	0,58651	0,137973	120	205	4,10E-50	14,8854
DEVICE	1,779081	0,866873	364	1288	4,54E-42	13,59082
SKIN	0,625611	0,175663	128	261	5,54E-42	13,57621
BONE	0,288368	0,049132	59	73	1,42E-37	12,81098
HEAL	0,268817	0,045094	55	67	8,68E-36	12,488
PROCEDURE	0,400782	0,093552	82	139	5,34E-35	12,34265
REMOVE	0,28348	0,055862	58	83	6,16E-31	11,56554
MANUFACTURE	0,395894	0,103648	81	154	9,07E-30	11,33238
INVASIVE	0,268817	0,055862	55	83	1,45E-27	10,87882
DISPOSE	0,30303	0,072015	62	107	3,62E-26	10,58186
MATERIAL	0,347019	0,096244	71	143	2,79E-24	10,16672
INFECT	0,386119	0,130569	79	194	4,60E-20	9,172912
PRINT	0,205279	0,045767	42	68	1,37E-19	9,054848
REDUCE	0,835777	0,429398	171	638	6,11E-18	8,630484
TECHNOLOGY	1,13392	0,650828	232	967	1,97E-17	8,495404
COMPLICIT	0,244379	0,069323	50	103	6,06E-17	8,364039
MECHANIC	0,239492	0,069323	49	103	3,55E-16	8,153062
PAIN	0,312805	0,109705	64	163	1,42E-15	7,984067
MINIM	0,234604	0,069323	48	103	2,00E-15	7,941384
LIGHT	0,215054	0,063939	44	95	4,04E-14	7,559724
NATURE	0,278592	0,098264	57	146	6,97E-14	7,488426
LAYER	0,14174	0,031633	29	47	7,40E-14	7,480537

PATCH	0,180841	0,049132	37	73	1,34E-13	7,401718
DAMAGE	0,175953	0,048459	36	72	4,98E-13	7,225903
PROPERTY	0,205279	0,063939	42	95	1,11E-12	7,115881
SURFACE	0,156403	0,041055	32	61	1,90E-12	7,041738
LIQUID	0,146628	0,037017	30	55	2,53E-12	7,001536
ELIMINATE	0,156403	0,041728	32	62	3,40E-12	6,959996
CAUSE	0,293255	0,119128	60	177	1,89E-11	6,714292
FLOW	0,210166	0,071342	43	106	2,02E-11	6,70483
OPERATION	0,332356	0,146049	68	217	4,87E-11	6,574786
HEAD	0,127077	0,031633	26	47	5,17E-11	6,565957
REPLACE	0,171065	0,052497	35	78	5,49E-11	6,556964
INSTRUMENT	0,156403	0,045767	32	68	8,14E-11	6,49799
FLUID	0,14174	0,040382	29	60	2,93E-10	6,302542
PATENT	0,307918	0,135954	63	202	3,19E-10	6,289419
PRODUCT	1,280547	0,88841	262	1320	1,22E-09	6,077958
SAFE	0,239492	0,097591	49	145	1,60E-09	6,034
LESS	0,239492	0,097591	49	145	1,60E-09	6,034
INNOVATION	0,434995	0,230852	89	343	3,40E-09	5,91084
FLEXIBLE	0,146628	0,047113	30	70	5,40E-09	5,834324
REQUIRE	0,40567	0,214026	83	318	8,72E-09	5,753852
DESIGN	0,532747	0,310271	109	461	1,66E-08	5,643783
SIMUL	0,146628	0,050478	30	75	3,76E-08	5,501841

CARDIAC	0,16129	0,061246	33	91	1,28E-07	5,282228
DECREASE	0,14174	0,050478	29	75	1,56E-07	5,2451
CHEMIC	0,107527	0,032979	22	49	2,53E-07	5,155673
HIGH	0,576735	0,360748	118	536	2,78E-07	5,137628

7.1.1.11 Taula 4.30

	V.TEST	MEAN IN CATEGORY	OVERALL MEAN	P.VALUE
DIM.1	227,1356	1,096236	0,014765	0
DIM.4	116,5121	0,517468	0,007151	0
DIM.2	79,85617	0,33594	-0,02944	0
DIM.7	51,57843	0,198979	-0,00455	0
DIM.9	48,93697	0,191252	0,003972	0
DIM.3	5,508076	0,030571	0,00628	3,63E-08
DIM.6	-44,3414	-0,16288	0,014412	0
DIM.5	-80,5705	-0,31452	0,016235	0
DIM.8	-145,329	-0,56433	-0,00291	0

7.1.1.12 Taula 4.31

	INTERN %	GLOB %	INTERN FREQ	GLOB FREQ	P.VALUE	V.TEST
CELL	3,468021	0,282003	321	419	1,10E-294	36,68893
DRUG	1,998704	0,282676	185	420	7,04E-107	21,95945
MOLECULE	0,918323	0,104321	85	155	8,13E-60	16,31182
PROTEIN	0,702247	0,074707	65	111	1,52E-48	14,64166
IMMUNE	0,486171	0,039036	45	58	1,43E-42	13,67532
TUMOUR	0,702247	0,090187	65	134	1,30E-41	13,51336
CULTURE	0,486171	0,04644	45	69	5,23E-37	12,70953
BACTERIA	0,583405	0,077399	54	115	7,49E-34	12,12817
COMPOUND	0,432152	0,042401	40	63	2,48E-32	11,83803
THERAPEUTIC	0,723855	0,129223	67	192	5,41E-32	11,77249
VIVO	0,345722	0,031633	32	47	1,49E-27	10,87653
TARGET	1,069576	0,306905	99	456	1,76E-27	10,86121

GENE	0,432152	0,05317	40	79	5,23E-27	10,76148
VITRO	0,45376	0,0599	42	89	1,03E-26	10,69886
EXPRESS	0,388937	0,043074	36	64	1,46E-26	10,66624
THERAPY	0,972342	0,278638	90	414	3,82E-25	10,35868
DISCOVERY	0,367329	0,041728	34	62	1,19E-24	10,24942
RELEASE	0,388937	0,050478	36	75	2,16E-23	9,96545
HUMAN	0,626621	0,169606	58	252	6,90E-18	8,616609
CANCER	0,864304	0,297483	80	442	1,99E-17	8,494503
RESIST	0,259291	0,032306	24	48	1,68E-16	8,243299
INDUCE	0,237684	0,030287	22	45	5,92E-15	7,805693
NOVEL	0,45376	0,116436	42	173	3,75E-14	7,569311
PROTECT	0,378133	0,084803	35	126	7,18E-14	7,484464
BIOLOGY	0,31331	0,0599	29	89	1,16E-13	7,421568
DAMAGE	0,280899	0,048459	26	72	1,42E-13	7,39407
DEVELOP	1,933881	1,095033	179	1627	4,07E-13	7,253134
DISEASE	0,929127	0,419976	86	624	1,26E-11	6,773099
PRECLINIC	0,22688	0,038363	21	57	2,24E-11	6,689455
ADMINISTER	0,205272	0,031633	19	47	3,11E-11	6,641185
EFFICACY	0,605013	0,230179	56	342	8,48E-11	6,491777
STUDY	0,486171	0,164221	45	244	1,23E-10	6,43521
PHASE	0,324114	0,085476	30	127	3,85E-10	6,259937
ACTIVE	0,723855	0,31902	67	474	8,52E-10	6,134984
RESPONSE	0,324114	0,088168	30	131	8,64E-10	6,132675
EFFECT	0,669836	0,286041	62	425	1,05E-09	6,101286
PROPERTY	0,270095	0,063939	25	95	1,07E-09	6,099091
DNA	0,22688	0,047113	21	70	1,84E-09	6,011618
HIGH	0,756266	0,360748	70	536	1,05E-08	5,721658
PROPRIETARY	0,280899	0,078072	26	116	1,97E-08	5,614536
INJECT	0,172861	0,032306	16	48	3,35E-08	5,522114
SMALL	0,270095	0,074707	25	111	3,39E-08	5,520145
LEAD	0,432152	0,164221	40	244	4,48E-08	5,470822
NATURE	0,31331	0,098264	29	146	5,19E-08	5,444568
ANIM	0,194468	0,043747	18	65	1,14E-07	5,302401
TISSUE	0,378133	0,137973	35	205	1,16E-07	5,298916
MECHANIC	0,248487	0,069323	23	103	1,47E-07	5,256054
CHEMIC	0,162057	0,032979	15	49	3,37E-07	5,101569
CAUSE	0,334918	0,119128	31	177	3,59E-07	5,089722
PHARMACEUTICS	0,22688	0,06192	21	92	3,59E-07	5,08926
STROKE	0,216076	0,057208	20	85	4,12E-07	5,063252
MODEL	0,669836	0,335846	62	499	4,53E-07	5,045279
VIRUS	0,183665	0,043747	17	65	6,56E-07	4,974025
ALTERN	0,183665	0,043747	17	65	6,56E-07	4,974025
SKIN	0,410545	0,175663	38	261	2,17E-06	4,736968
SAFETY	0,248487	0,082111	23	122	3,55E-06	4,635882
SELECT	0,248487	0,082784	23	123	4,11E-06	4,6055

CONTAIN	0,205272	0,060573	19	90	4,78E-06	4,574011
PHARMA	0,237684	0,078072	22	116	5,32E-06	4,55158
EYE	0,162057	0,040382	15	60	5,91E-06	4,529414
NEW	0,637424	0,339884	59	505	6,39E-06	4,512968
SHOW	0,324114	0,129896	30	193	7,35E-06	4,483326
MODULE	0,216076	0,06865	20	102	8,76E-06	4,445667
SINGLE	0,22688	0,076726	21	114	1,43E-05	4,338883
APPROACH	0,378133	0,171625	35	255	2,12E-05	4,252371
DEATH	0,140449	0,034325	13	51	2,13E-05	4,251147
PRODUCT	1,318064	0,88841	122	1320	2,34E-05	4,229781
TECHNOLOGY	1,015557	0,650828	94	967	3,37E-05	4,147056
TREATMENT	0,939931	0,592947	87	881	3,81E-05	4,118643
NORMAL	0,151253	0,042401	14	63	5,36E-05	4,039532
PROFILE	0,205272	0,071342	19	106	5,71E-05	4,024536
ASSAY	0,151253	0,044421	14	66	9,30E-05	3,908279
SURFACE	0,140449	0,041055	13	61	0,000163	3,770003
LICENSE	0,162057	0,053843	15	80	0,000226	3,687815
CONCEPT	0,183665	0,067977	17	101	0,000329	3,591605
ISOLATION	0,129646	0,038363	12	57	0,000343	3,580178
AMOUNT	0,140449	0,044421	13	66	0,00038	3,5538
PATHWAY	0,162057	0,056535	15	84	0,000399	3,540785
MATERIAL	0,22688	0,096244	21	143	0,000446	3,511344
DELIVERY	0,356525	0,184412	33	274	0,00048	3,491485
OVERCOME	0,108038	0,028941	10	43	0,0005	3,480659
COMMERCIAL	0,183665	0,072688	17	108	0,000748	3,37129
SPECIFY	0,410545	0,230179	38	342	0,000878	3,326926
SEVER	0,291703	0,145376	27	216	0,000939	3,308102
CHARACTERISE	0,108038	0,031633	10	47	0,001075	3,27008
RECENT	0,118842	0,03769	11	56	0,001188	3,24164
PATHOLOGY	0,183665	0,077399	17	115	0,001567	3,161903
THEREFOR	0,205272	0,091533	19	136	0,00162	3,152217
SIDE	0,172861	0,072015	16	107	0,001944	3,09862
ENGINE	0,118842	0,041728	11	62	0,002877	2,980599
ENTER	0,108038	0,035671	10	53	0,002879	2,980422
SCREEN	0,280899	0,149414	26	222	0,003085	2,959114
DEMONSTRATE	0,129646	0,050478	12	75	0,004488	2,841669
INFECT	0,248487	0,130569	23	194	0,004748	2,823647
APPLY	0,205272	0,100283	19	149	0,004799	2,820254
PHYSIOLOGY	0,118842	0,044421	11	66	0,004838	2,817614
SIMULATE	0,08643	0,026922	8	40	0,005761	2,761081
RESEARCH	0,399741	0,247678	37	368	0,006104	2,742136
BLOOD	0,388937	0,239602	36	356	0,006258	2,733935
UNMET	0,08643	0,027595	8	41	0,006772	2,707838
STRATEGI	0,162057	0,074034	15	110	0,006922	2,700551
THUS	0,248487	0,134608	23	200	0,006983	2,697637

LARGE	0,194468	0,096917	18	144	0,007621	2,66842
ADVANTAGE	0,118842	0,047113	11	70	0,007762	2,662255
MULTIPLY	0,183665	0,09086	17	135	0,008963	2,613447
GENETIC	0,172861	0,083457	16	124	0,009014	2,611521
IDENTIFY	0,334918	0,203258	31	302	0,009288	2,601276
TRIAL	0,31331	0,187105	29	278	0,009643	2,588359
DEFINE	0,08643	0,029614	8	44	0,010644	2,554175
DIFFERENT	0,356525	0,222776	33	331	0,01114	2,53827
SIGNAL	0,162057	0,078072	15	116	0,011441	2,528947
BRAIN	0,22688	0,124512	21	185	0,011795	2,51821
PROCESS	0,507779	0,345942	47	514	0,011911	2,514757
RELY	0,097234	0,037017	9	55	0,013166	2,479254
POTENTIAL	0,270095	0,158837	25	236	0,013889	2,460125
RELIVE	0,172861	0,088168	16	131	0,015279	2,425694
STEP	0,205272	0,111724	19	166	0,015618	2,417723
LOW	0,31331	0,196527	29	292	0,018691	2,351628
SEQUENCE	0,097234	0,039709	9	59	0,020833	2,311006
BARRIER	0,097234	0,039709	9	59	0,020833	2,311006
RAPID	0,129646	0,06192	12	92	0,023613	2,263362
SCALE	0,172861	0,093552	16	139	0,026267	2,222243