GOBIERNO DE ESPAÑA

MINISTERIO DE INDUSTRIA, TURISMO Y COMERCIO

ontsi observatorio nacional de las telecomunicaciones y de la SI

Proyecto Piloto sobre Viabilidad de Internet como Fuente de Datos

Resultados del Proyecto

ÍNDICE

1. Detección automática de B2C
2. Análisis de demanda de profesionales TICC y programas formativos
3. Análisis comparativo de la oferta y la demanda de profesionales TICC.
4. Detección de ofertas de empleo en los sitios web de empresas españolas.
5. Conclusiones

01.

El proyecto

Concepción

Utilizar Internet como fuente de datos (IaD, *Internet as Data source*), con dos objetivos diferenciados:

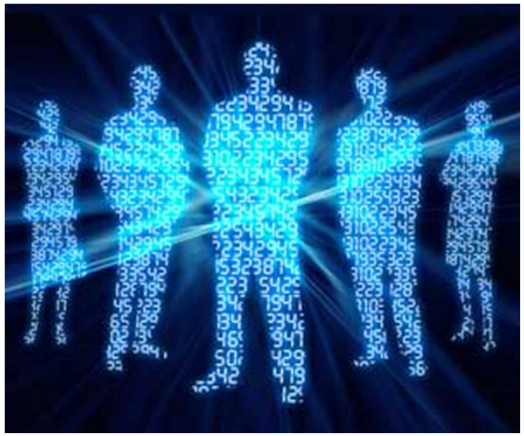
- Hacer el seguimiento del comercio electrónico en empresas españolas (subproyecto 1)
- Hacer el seguimiento de la oferta y la demanda de profesionales TIC en España (subproyecto 2)

Subproyecto 1

- Objetivos:

  1. Detectar la presencia de comercio electrónico en los sitios web de empresas españolas.
  2. Capturar indicadores asociados a la venta electrónica.

Subproyecto 2



- Objetivo:
  - Caracterizar la demanda de profesionales TIC y de los Contenidos, así como la oferta formativa de centros universitarios y de formación profesional
- Estrategias:
  - Acceso a portales centralizados
  - Acceso a la web de empresas y universidades

Subproyecto 2

- Objetivos concretos:

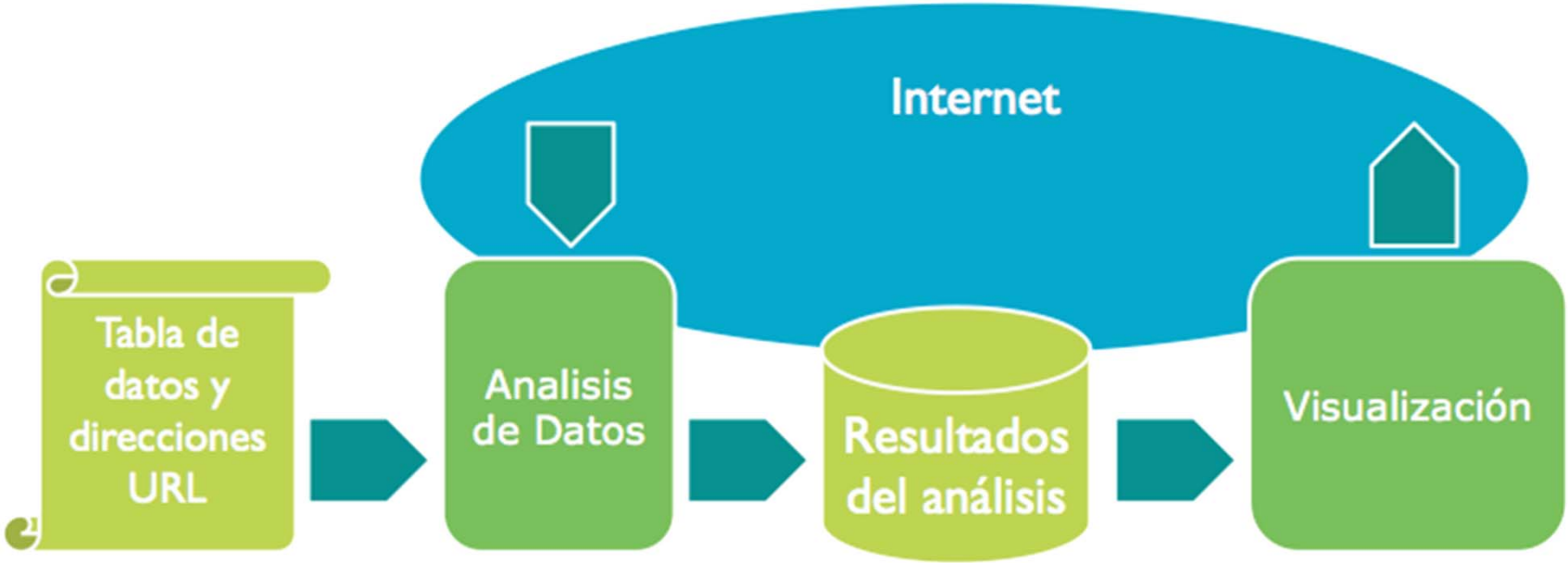
  1. Detectar automáticamente la presencia de ofertas de empleo en los sitios web de empresas.

  2. Analizar comparativamente la oferta laboral y curricular a partir de portales centralizados.

Metodología

1. Desarrollar un SW de captura, análisis y visualización de datos.
2. Aplicar el SW a las tareas de detección, perfilado y matching.
3. Extraer conclusiones sobre la viabilidad del ML para IaD

Procesos


Internet
Tabla de datos y direcciones URL
Analisis de Datos
Resultados del análisis
Visualización

Componentes del análisis

Rastrear

Perfilar

Analizar texto

Comparar

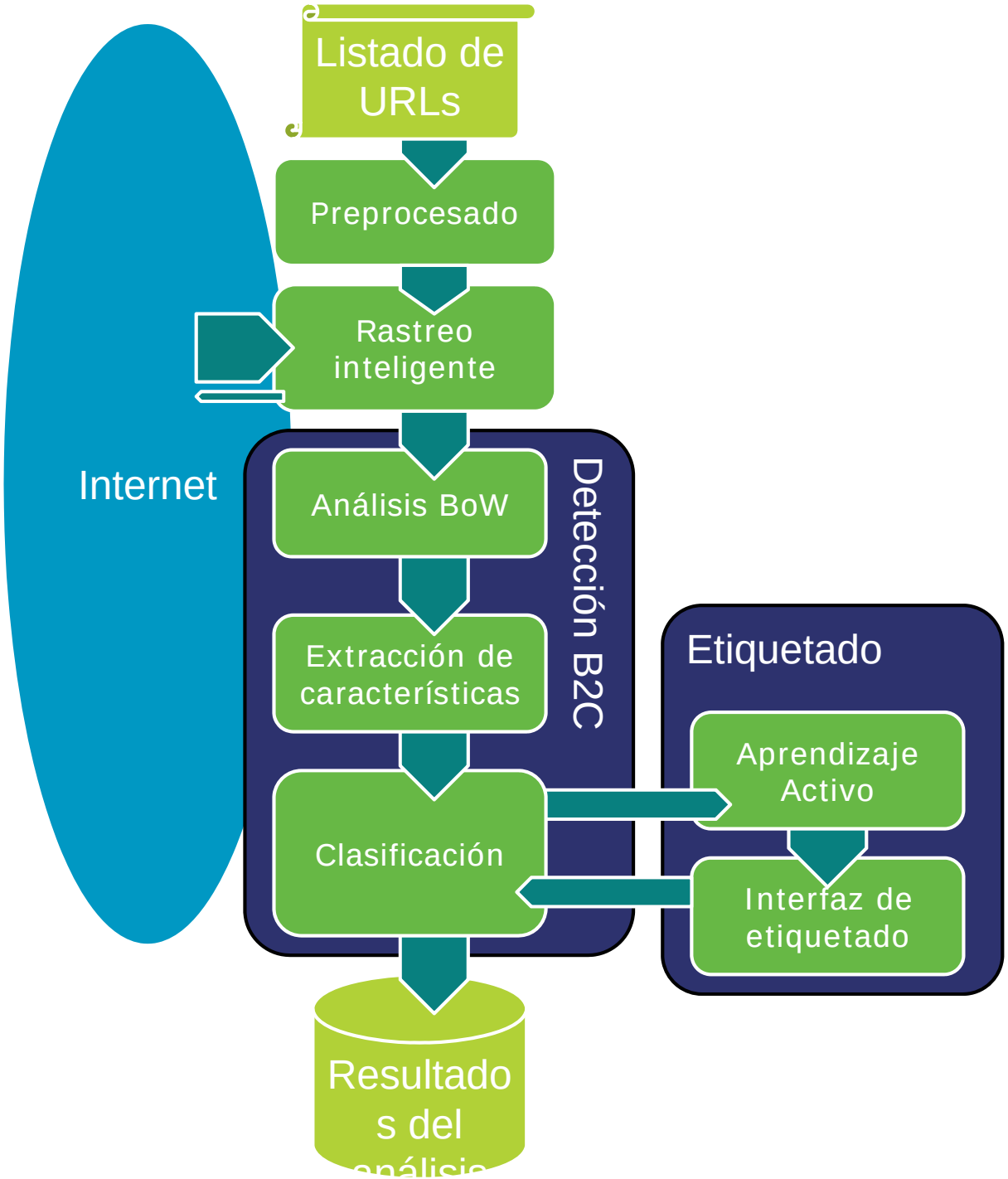
Clasificar

02.

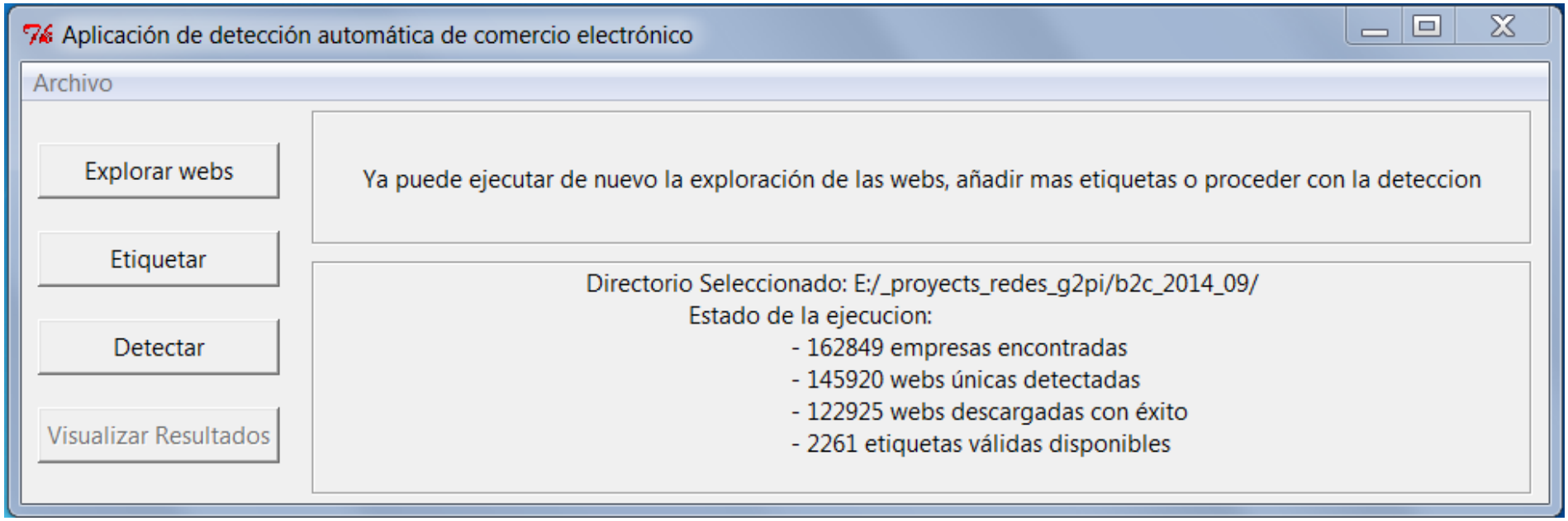
Detección automática de B2C

Detección de B2C

Proceso de detección automática de B2C



La aplicación de exploración y análisis web



Exploración web

FUENTE DE DATOS

162.849

EMPRESAS,

145.920

DOMINIOS WEB diferentes

EXPLORACIÓN WEB

El **CRAWLER** obtiene una representación de cada sitio web basada en

8.763.024

TÉRMINOS, que se reducen a

343.780

EXTRACCIÓN DE CARACTERÍSTICAS

Identifica los

10.000

TÉRMINOS, más relevantes para detección de B2C.

Etiquetado

2.540

PÁGINAS WEB ETIQUETADAS.

¿Cuándo hay B2C?

Condición 1: Variedad de productos, añadir a cesta de la compra, completar pago, finalizar pedido.

Condición 2: Realizar una reserva (de habitaciones de hotel, de entradas, de billetes de viaje), pagar o señalizar la operación

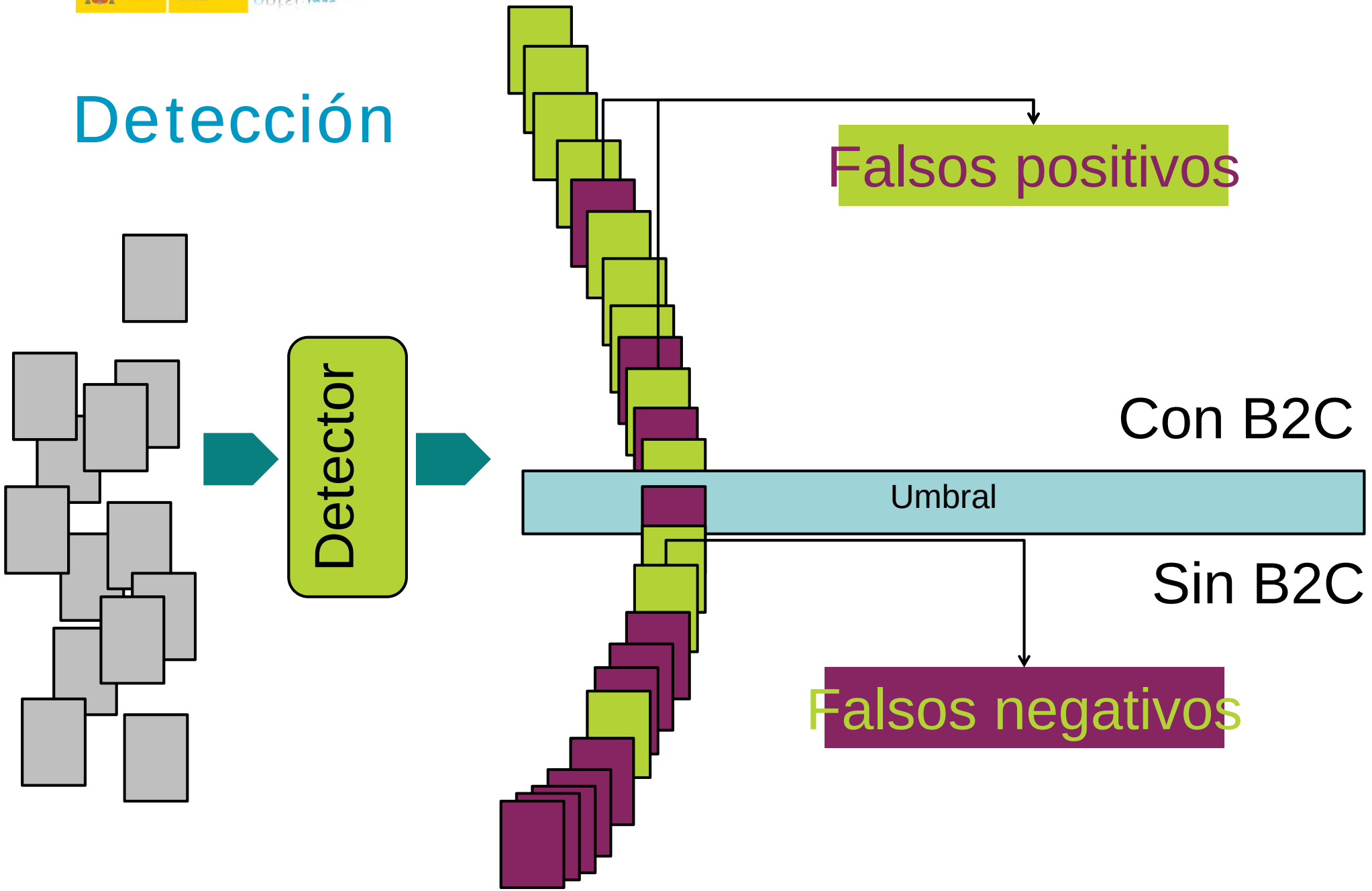
1.606

webs **SIN B2C**

776

webs **CON B2C**

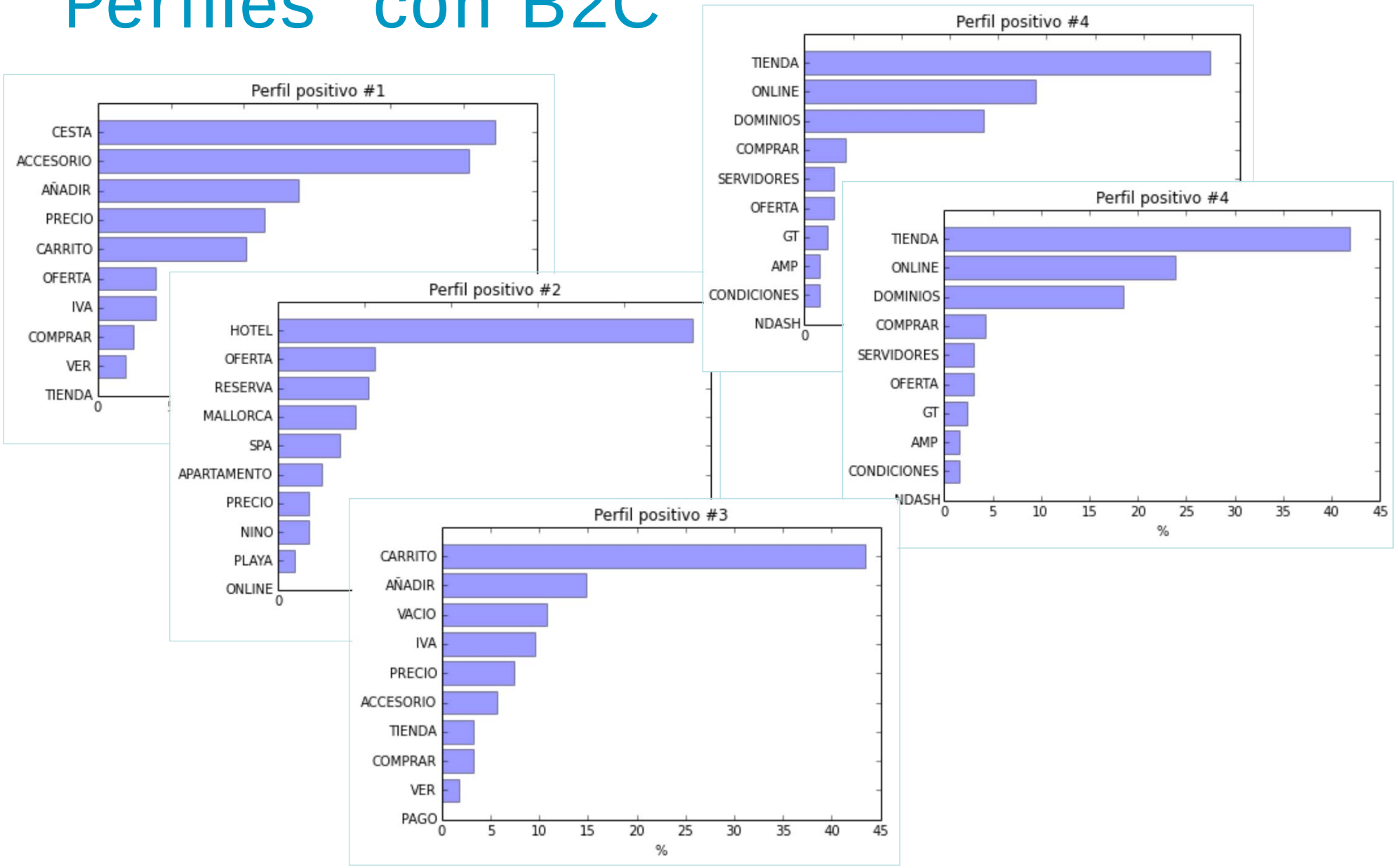
Detección


Detector
Falsos positivos
Con B2C
Umbral
Sin B2C
Falsos negativos

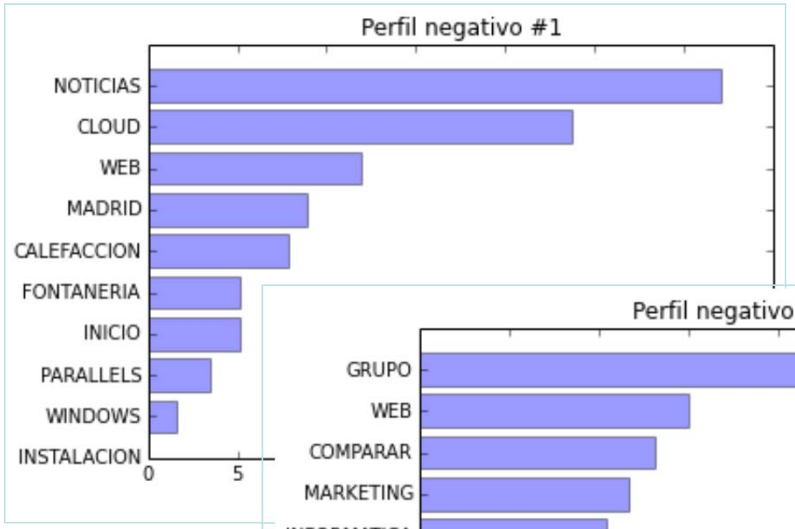
Resultados

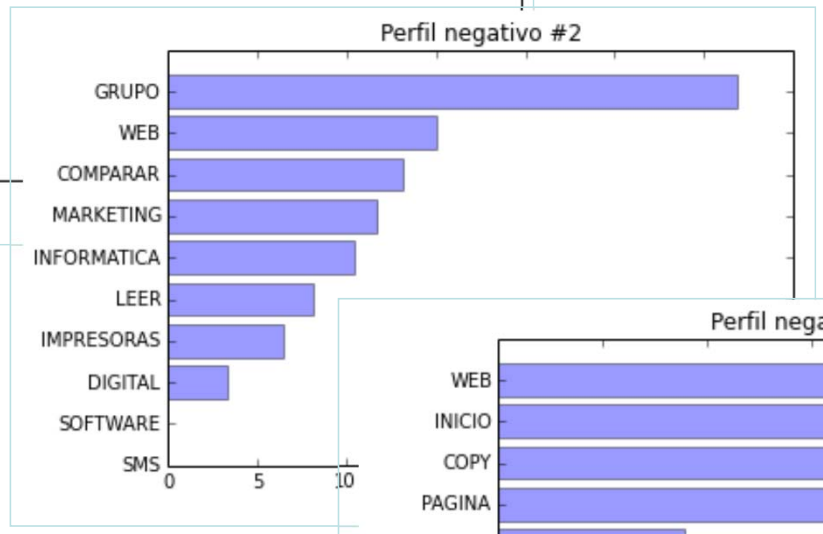
http://IaD.tsc.uc3m.es/

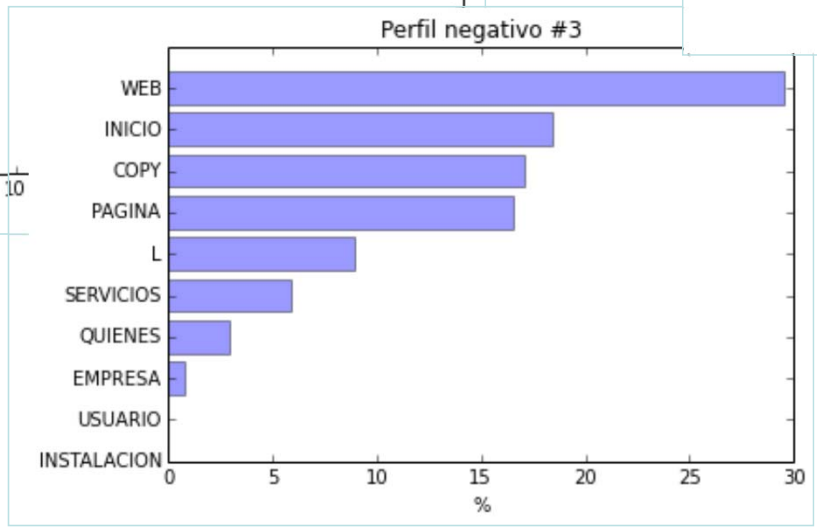
Perfiles “con B2C”

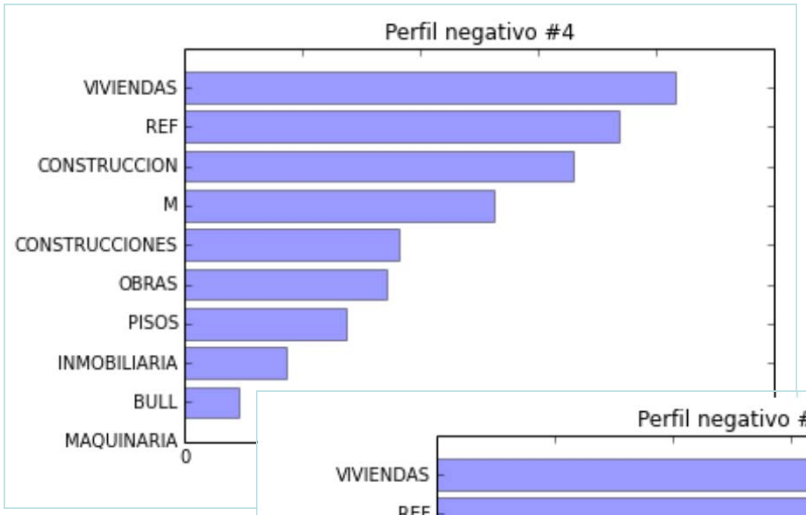


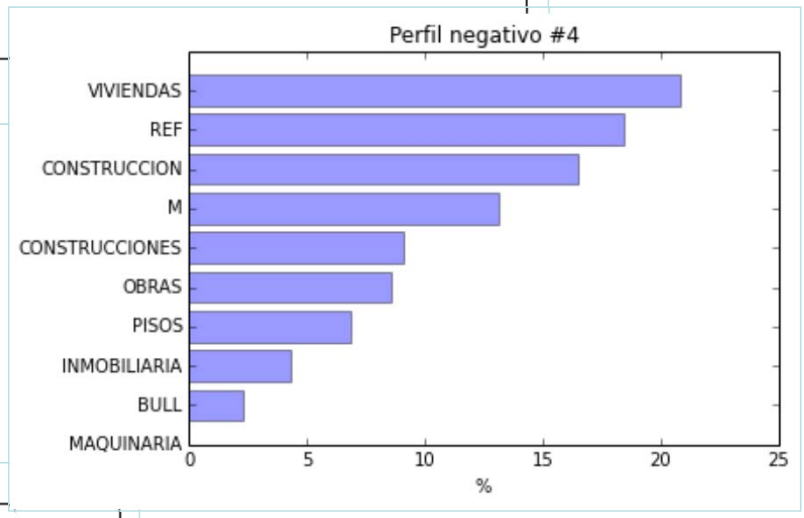
Perfiles “sin B2C”











03.

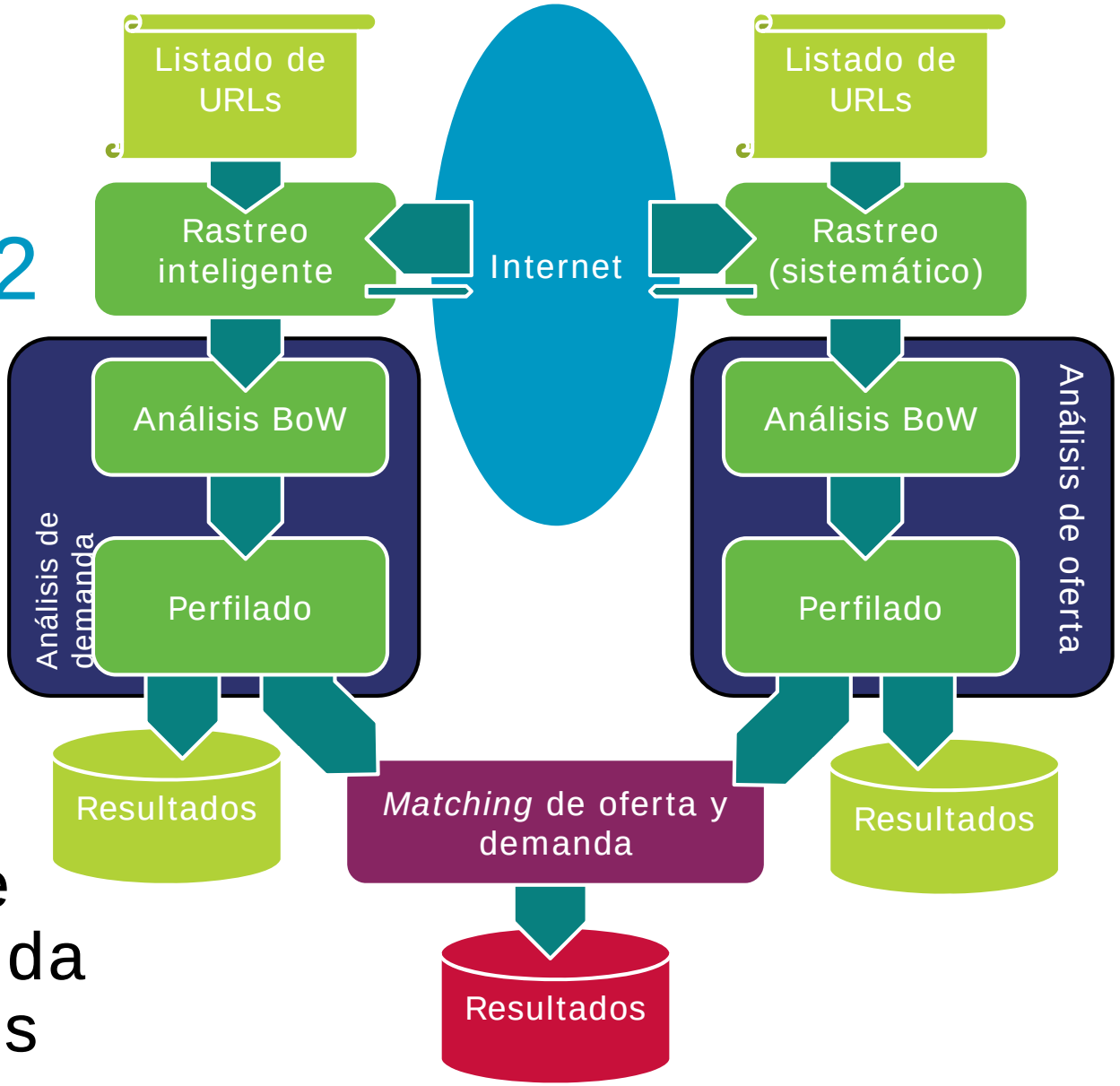
Análisis de demanda de profesionales TICC y programas formativos.

Tareas de perfilado y “matching”

- Objetivo 1:
  Utilizar fuentes de datos disponibles en Internet para estimar y modelar la demanda de profesionales del Sector TICC
- Objetivo 2:
  Ídem para la oferta formativa disponible
- Objetivo 3 (“matching”):
  Analizar la adecuación de la oferta formativa a la demanda de profesionales

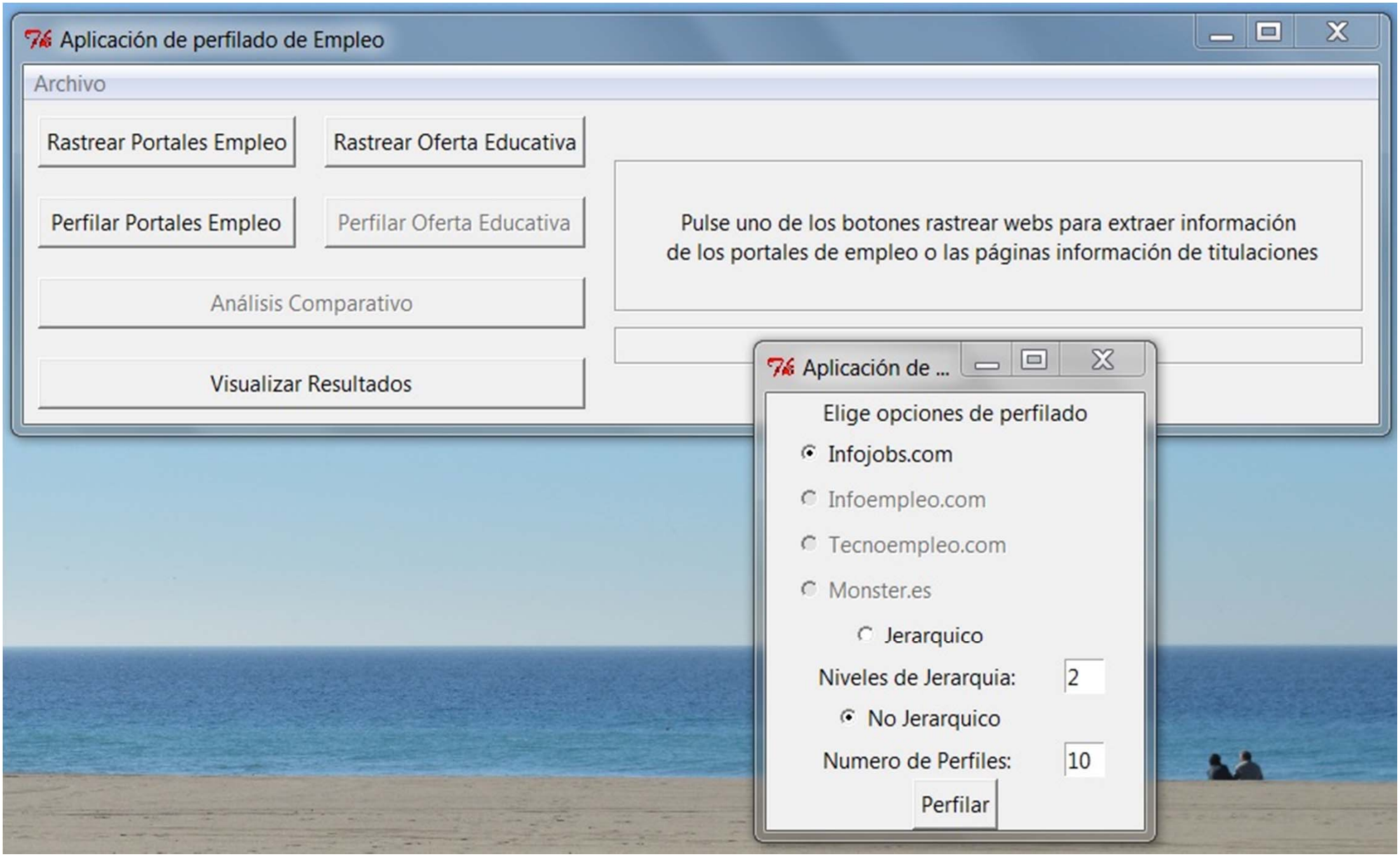
- Se han aplicado técnicas de ML que evitan la necesidad de etiquetado manual y que permiten detectar grupos de términos “coherentes” en los documentos

Estructura global del suproyecto 2

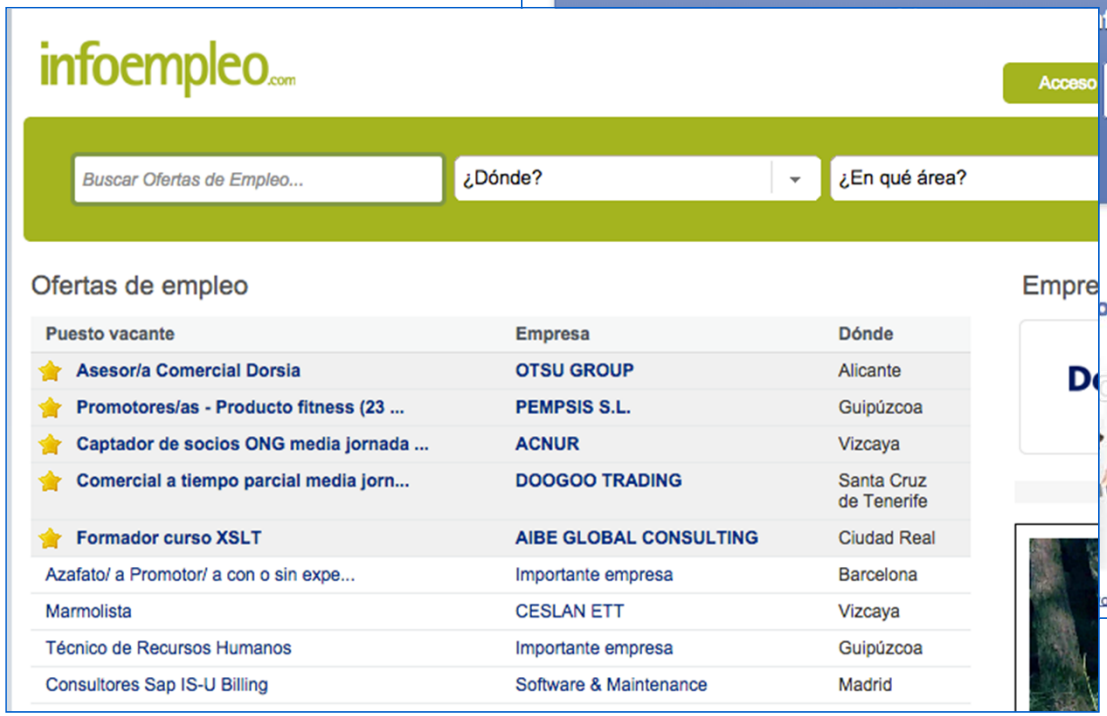


Análisis comparativo de oferta y demanda de profesionales TICC

La aplicación de exploración y análisis web



Fuente de datos



Fuente de datos

Exploración web

	InfoJobs	Tecnoempleo	Infoempleo
# Ofertas crawler	3.079	3.380	1.160
# Descartadas	216	301	151
# Ofertas corpus	2.863	3.079	1.009
Categorías	Inform.+Teleco Diseño Web Marketing+Com	---	Tecnología Telecomunicac. Internet Artes gráficas
Tamaño del vocabulario	1.211	1.881	823
Tamaño promedio de los documentos (en palabras)	14,85	46,05	23,67

Rama del conocimiento	Nivel	Total de títulos	Títulos TICC	Plan de estudios
Ingeniería y arquitectura	G	714	257	180
	M	1164	232	130
Ciencias	G	223	75	40
	M	670	104	45
Arte y Humanidades	G	403	42	20
	M	728	60	27
Ciencias Sociales y Jurídicas	G	923	178	95
	M	2025	187	99

72 cualificaciones profesionales:

- 31 cualificaciones de la rama de Artes Gráficas.
- 18 cualificaciones de Imagen y Sonido.
- 23 cualificaciones de Informática y Comunicaciones.

Perfilado con LDA

- Objetivo: Encontrar “bloques constructivos” para los documentos del corpus. Sin etiquetas.
- Perfiles verticales y transversales
- Asignación “blanda” de documentos a perfiles

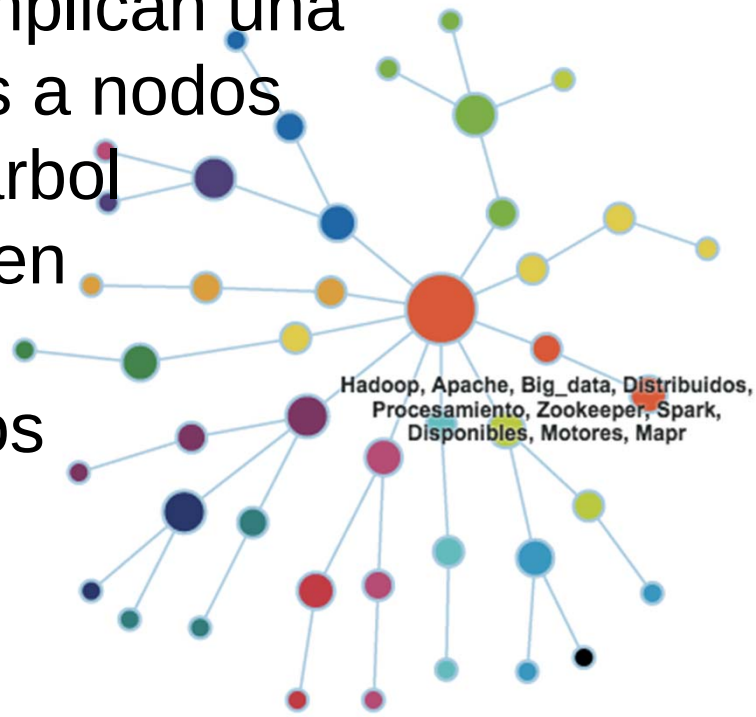
Topic 0 (19,5%)	Topic 1 (13%)	Topic 2 (11,4%)	Topic 3 (9,8%)	Topic 4 (9,4%)
marketing	linux	java	business_intelligence	javascript
nivel	windows	j2ee	power_center	html
frances	sistemas	spring	proyectos	html5
comercial	seguridad	struts	consultor	php
alto	redes	oracle	testing	css
comunicacion	vmware	jsf	funcional	css3
excel	unix	eclipse	telecomunicaciones	jquery
ingles	soporte	maven	sas	web
campañas	administrador	jboss	procesos	diseño
office	servidores	tomcat	business_objects	ajax
Topic 5 (8,5%)	**Topic 6 (8,1%)**	**Topic 7 (7,7%)**	**Topic 8 (7,5%)**	**Topic 9 (5,1%)**
oracle	.net	marketing	sap	sharepoint
pl	asp_net	seo	android	crm
cobol	server	digital	abap	dynamics
db2	visual_studio	online	ios	navision
host	visual_basic	cisco	consultor	red_hat
unix	mvc	manager	liferay	microsoft
aix	programacion	sem	tibco	consultor
cics	vb_net	community	soa	nav
banksphere	entity	redes_sociales	bpm	cognos
dba	windows	comunicacion	moviles	salesforce

Perfilado de Ofertas de empleo

Descripción del Perfil	Competencias (palabras clave)	Perfiles	PAFET VII	Estudio ONTSI	Clasificación ESCO
Experto programador en JAVA	Java / j2ee / spring / struts / jsf / eclipse / maven	Infojobs (T0) Infoempleo (T0) Tecnoemp. (T0, T13)	Programador	Programador SW	2512 – SW developers
Community manager	seo / redes_sociales / manager / sem / community / google / google_analytics	Infojobs (T11)	Community Manager Content Editor	Community Manager	
Especialista en Marketing online	marketing / publicidad / online / comunicación / comercial / estrategias	Infojobs (T4) Infoempleo (T17)	Especialista en Marketing online Content Editor	Técnico de marketing Ingeniero comercial	
Especialista en diseño Web	html / html5 / web / javascript / diseño / css / css3 / .net / php / asp_net / visual_studio	Infojobs (T2, T5) Infoempleo (T1, T4) Tecnoemp. (T1, T2)	Diseñador Diseñador gráfico/web Especialista UX Webmaster	Técnico de la web	2513 – Web and multimedia developers
Diseñador de Contenidos digitales	web / diseño /photoshop / diseño_grafico / maquetacion	Infoempleo (T12)	Diseñador Arquitecto de contenidos	Diseñador gráfico Editor o gestor de contenido	
Administrador de Sistemas	linux / unix / sistemas / administracion / windows / vmware	Infojobs (T6, T8) Infoempleo (T11, T15) Tecnoemp. (T3)			2522 – System administrators
Técnico de Soporte Software	soporte / incidencias / atencion usuario / movilidad	Infoempleo (T10) Tecnoemp. (T5)			
Gestor de proyectos y productos	sap / gestión / proyectos / consultor	Infojobs (T7)	Gestor de productos		
Especialista en Gestión de BBDD	pl / oracle / analisis / forms	Infojobs (T15) Tecnoemp. (T7)		Administrador de BBDD	2521 – Database designers and administrators
Ingeniero de pruebas	testing / tester / calidad / qa	Tecnoemp. (T18)	Ingeniero de pruebas		
Especialista en COBOL en entorno bancario	cobol / db2 / host /cics / banca / banksphere / riesgos	Infojobs (T13) Tecnoemp. (T8)			
Analista de Datos	Business_intelligence / Business_objects / power_center / datos / sas / estadística	Infojobs (T10, T19) Infoempleo (T7) Tecnoemp. (T12)	Periodista de Datos Especialista *Big Data*		
Diseñador de Aplicaciones Móviles	android / ios / moviles / aplicaciones	Infojobs (T12) Infoempleo (T6)	Programador Técnico desarrollador de aplicaciones		2514 – Application programers
Especialista en seguridad TIC / Ingeniero de Redes	redes / cisco / firewall / tcp / ip / routers / seguridad / gestion / infraestructura	Infojobs (T0,T16) Infoempleo (T5, T9) Tecnoemp. (T15)		Especialista en Seguridad TIC	2523 – Computer network professionals

Perfilado jerárquico.

- Cada nivel de la jerarquía se puede considerar un modelo completo de perfiles para el corpus de datos
- Las relaciones jerárquicas no implican una asignación dura de documentos a nodos
- Postprocesado para limpiar el árbol
- Alternativa: Modelos que agrupen de forma aglomerativa los documentos del corpus de datos en base a su "similitud semántica"



Resultados

http://IaD.tsc.uc3m.es/

04.

Análisis comparativo de la oferta y demanda de profesionales TICC.

Exploración web

	InfoJobs	Tecnoempleo	Infoempleo
# Ofertas crawler	3.079	3.380	1.160
# Descartadas	216	301	151
# Ofertas corpus	2.863	3.079	1.009
Categorías	Inform.+Teleco Diseño Web Marketing+Com	---	Tecnología Telecomunicac. Internet Artes gráficas
Tamaño del vocabulario	1.211	1.881	823
Tamaño promedio de los documentos (en palabras)	14,85	46,05	23,67

Rama del conocimiento	Nivel	Total de títulos	Títulos TICC	Plan de estudios
Ingeniería y arquitectura	G	714	257	180
	M	1164	232	130
Ciencias	G	223	75	40
	M	670	104	45
Arte y Humanidades	G	403	42	20
	M	728	60	27
Ciencias Sociales y Jurídicas	G	923	178	95
	M	2025	187	99

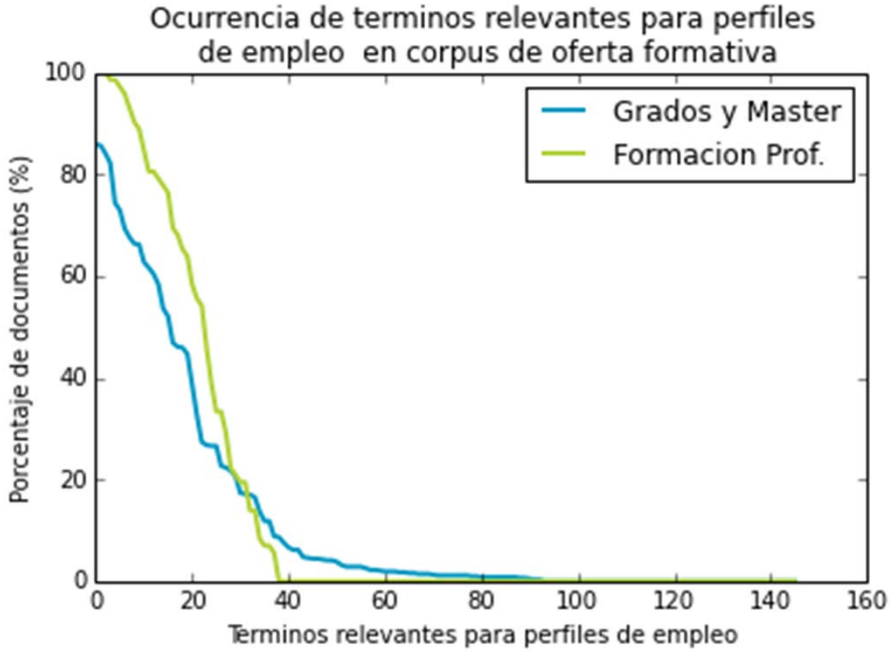
Matching: vocabularios de oferta y demanda

Topic 0 (14,9%)	Topic 1 (13,38%)	Topic 2 (10,93%)	Topic 3 (8,79%)	Topic 4 (6,99%)
software	electronica	artistica	audiovisual	matematicas
computadores	electronico	marketing	audiovisuales	fisica
computing	fisica	bellas_artes	cine	ecuaciones
seguridad	industrial	dra	multimedia	matematicos
inteligentes	comunicaciones	innovacion	ficcion	calculo
sistemas_operativos	circuitos	artisticos	cinematografico	algebra
informaticos	matematicas	investigador	postproduccion	geometria
inteligencia	instrumentacion	estrategica	television	numericos
artificial	automatica	campañas	indiv	ordenador
distribuidos	technology	business	documentales	teorema
Topic 5 (5,71%)	Topic 6 (5,63%)	Topic 7 (5,61%)	Topic 8 (5,24%)	Topic 9 (5,01%)
periodismo	industrial	comunicaciones	biotecnologia	sem
online	electronica	circuitos	biologia	marketing
chat	fisica	telecomunicaciones	molecular	mb
obligat	industriales	señales	genetica	recuperacion
metodo	electrica	telematica	biotecnologicos	periodismo
blackboard	mecanica	electronico	celulares	online
videoconferencia	instrumentacion	procesado	animal	jorge
blogs	automatica	señal	quimica	audiovisual
virtuales	quimica	audio	bioquimica	buscadores
colaborativo	automatizacion	telematicos	vegetal	multimedia
Topic 10 (4,47%)	Topic 11 (4,16%)	Topic 12 (3,79%)	Topic 13 (3,7%)	Topic 14 (1,7%)
publicitaria	online	periodismo	uveg	telefonica
marketing	marketing	periodistica	libres	buzon
publicitarios	videojuegos	periodisticos	etse	asincrona
audiovisual	uca.es	audiovisual	continuado	sincrona
mensajes	potestad	television	voluntaria	wikis
taller	maritima	informativos	presencialmente	moodle
iglesia	antonio	prensa	repaso	blogs
ds	marina	periodista	discusiones	glosarios
sintesis	suma	radio	certificacion	diferir
estrategica	oe	audiovisuales	videojuegos	video

Topic 0 (13%)	Topic 1 (12,37%)	Topic 2 (8,45%)	Topic 3 (8,42%)	Topic 4 (7,95%)
			audiovisuales	tinta
			camaras	entintado
			postproduccion	cilindros
			television	imprimir
			grabacion	primas
			sonido	tirada
			video	secado
			rodaje	tintar
			cinematografico	huecograbado
			cinematograficas	preimpresion
Topic 5 (7,28%)	Topic 6 (6,46%)	Topic 7 (6,04%)	Topic 8 (5,68%)	Topic 9 (5,61%)
tapas	transformar	estampacion	fotograficos	sonido
encuadernacion	carton	obra	fotograficas	musical
primas	ondulado	matrices	vectoriales	espectaculos
pliegos	envasar	bocetos	digitalizacion	sonora
plegado	papeleria	matriz	peliculas	mezcla
fotolitos	embalaje	grabado	procesado	audio
impresoras	complejos	estampas	maquetacion	microfonos
estampacion	colas	pantallas	revelado	vivo
dura	fabricacion	estampar	laboratorio	grabacion
imposicion	apilado	fotolitos	quimicos	radio
Topic 10 (5,06%)	Topic 11 (4,58%)	Topic 12 (4,29%)	Topic 13 (2,41%)	Topic 14 (2,4%)
preimpresion	espectaculos	editorial	embalaje	encuadernacion
transformar	vivo	animacion	envasar	industrial
tintar	ilustraciones	obra	estructurales	libro
fabricacion	artistico	autor	envase	tapas
preventivas	evento	derechos	prototipos	artisticas
calibracion	colectivo	libro	fabricacion	contenedores
emergencia	escena	bocetos	maquetas	lomo
proveedores	artisticas	maquetacion	artes	construccion
recomendaciones	listado	encargo	logistica	pieles
colas	teatros	tipografia	carton	adhesivos

Relaciones ofertas-formación

- Restricción al vocabulario “relevante” para las ofertas de empleo
- Únicamente 40 / 150 términos aparecen en el corpus de titulaciones FP
- 90/150 términos aparecen el corpus de Grados y Master, pero la gran mayoría en muy baja proporción
- 92 documentos del corpus de Grados y Máster no contienen términos de los 150 más importantes del corpus de empleo.

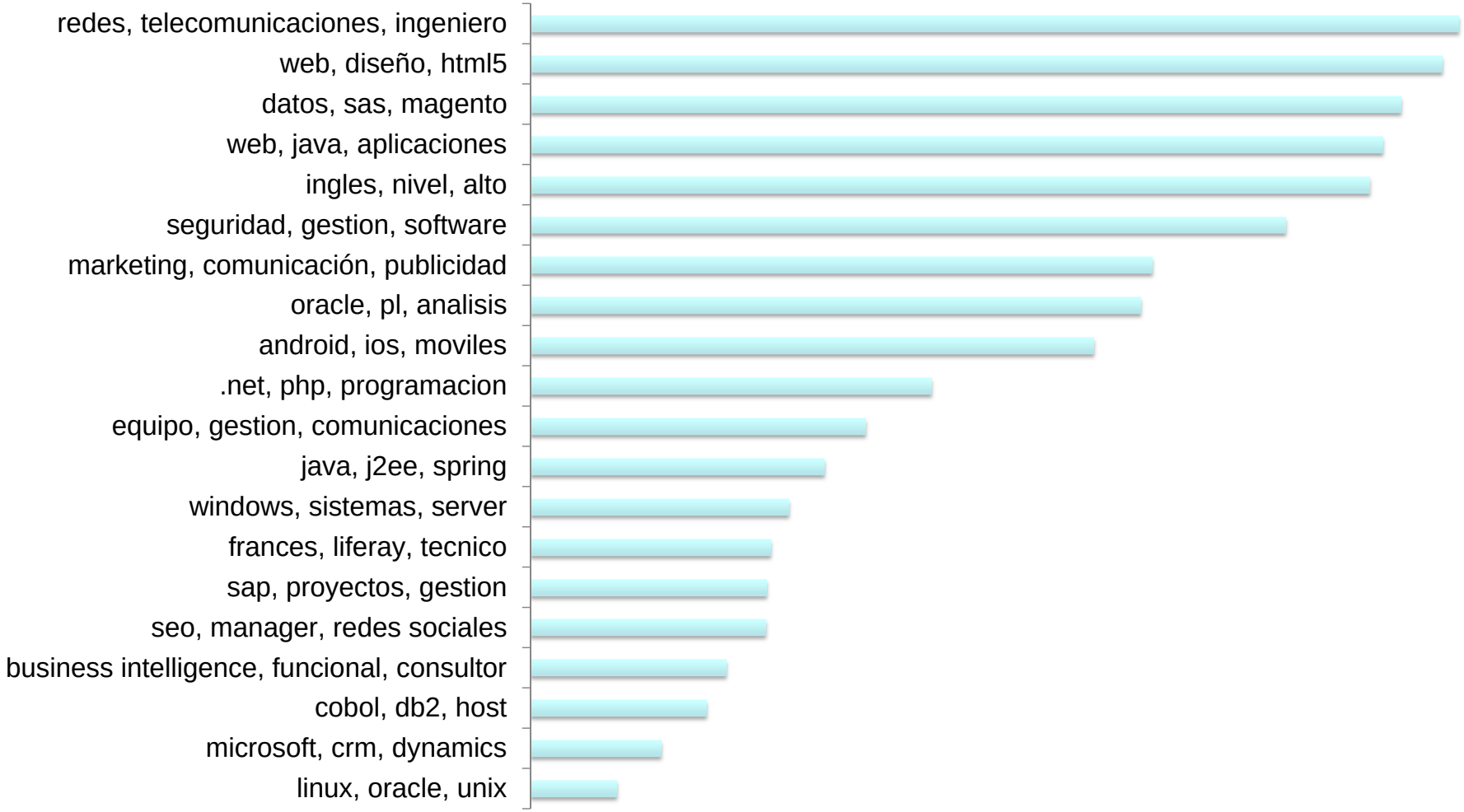


- Conclusión 1: En Sector TIC la oferta laboral parece estar más dirigida a titulados superiores
- Conclusión 2: Las Universidades deberían hacer un esfuerzo por acercar (al menos) sus descripciones de planes de estudio a lo demandado por el mercado laboral

Matching: resultados disponibles

Las técnicas de matching exploradas permiten:

- Establecer similitudes entre documentos y/o perfiles de distintos corpus de documentos
- Proporcionar un ranking de cobertura de la oferta de empleo.



Resultados

http://IaD.tsc.uc3m.es/

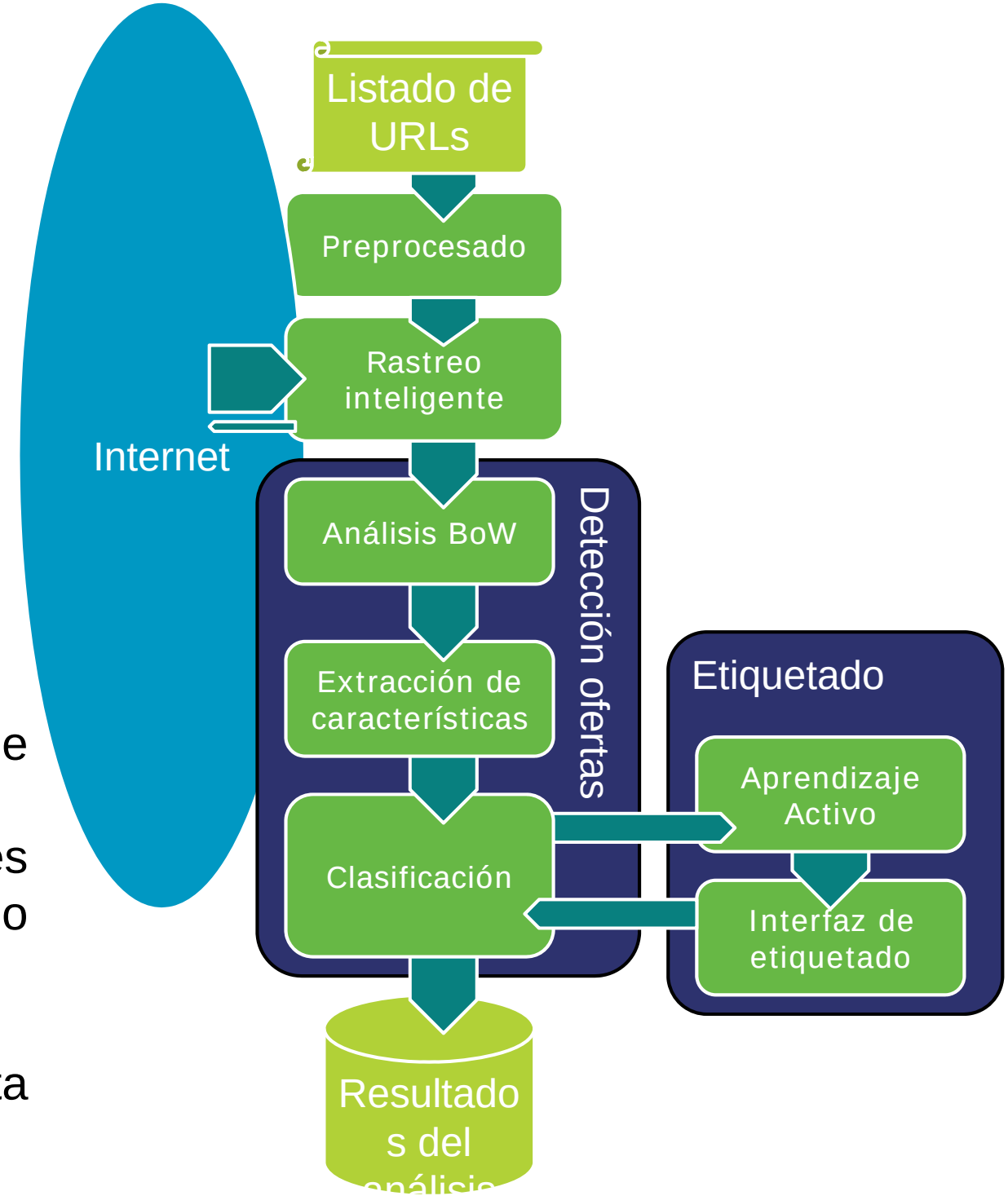
05.

Detección de ofertas de empleo en los sitios web de empresas españolas

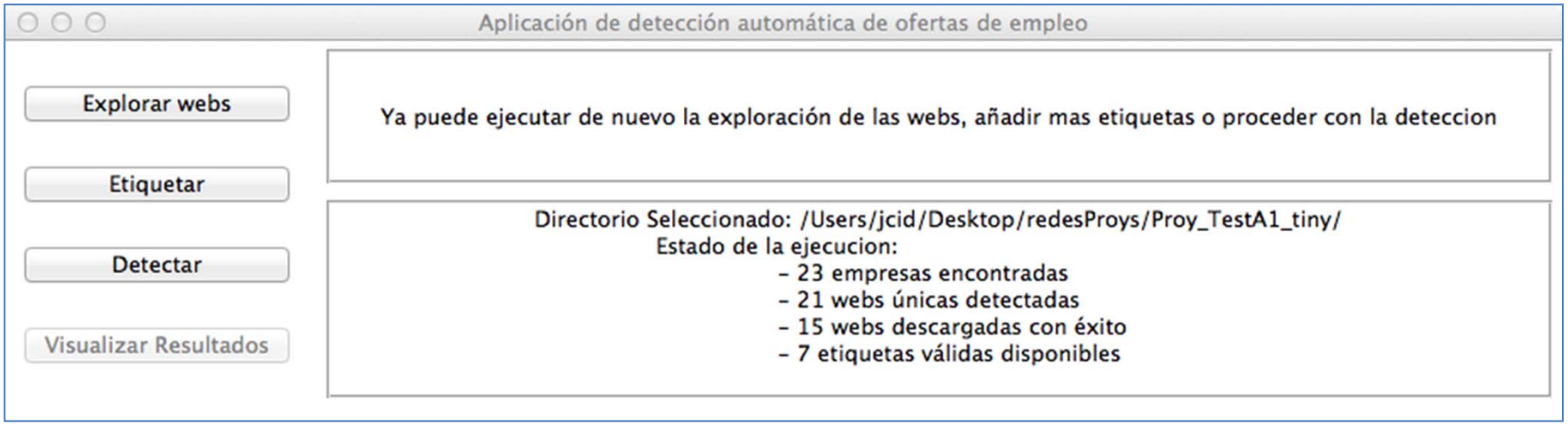
Detección de ofertas de empleo

Objetivos:

1. Detectar la presencia de ofertas de empleo.
2. Localizar las ubicaciones de las ofertas de empleo en la web.
3. Segmentar las ofertas
4. Caracterizar la oferta total.



La aplicación de exploración y análisis web



Exploración web

FUENTE DE DATOS

8.349

EMPRESAS, del sector **TICC**

EXPLORACIÓN WEB

El **CRAWLER** obtiene una representación de cada sitio web basada en

94.673

TERMINOS

EXTRACCIÓN DE CARACTERÍSTICAS

Identifica los

1.000

TÉRMINOS, más relevantes para detección de B2C.

Etiquetado manual

828

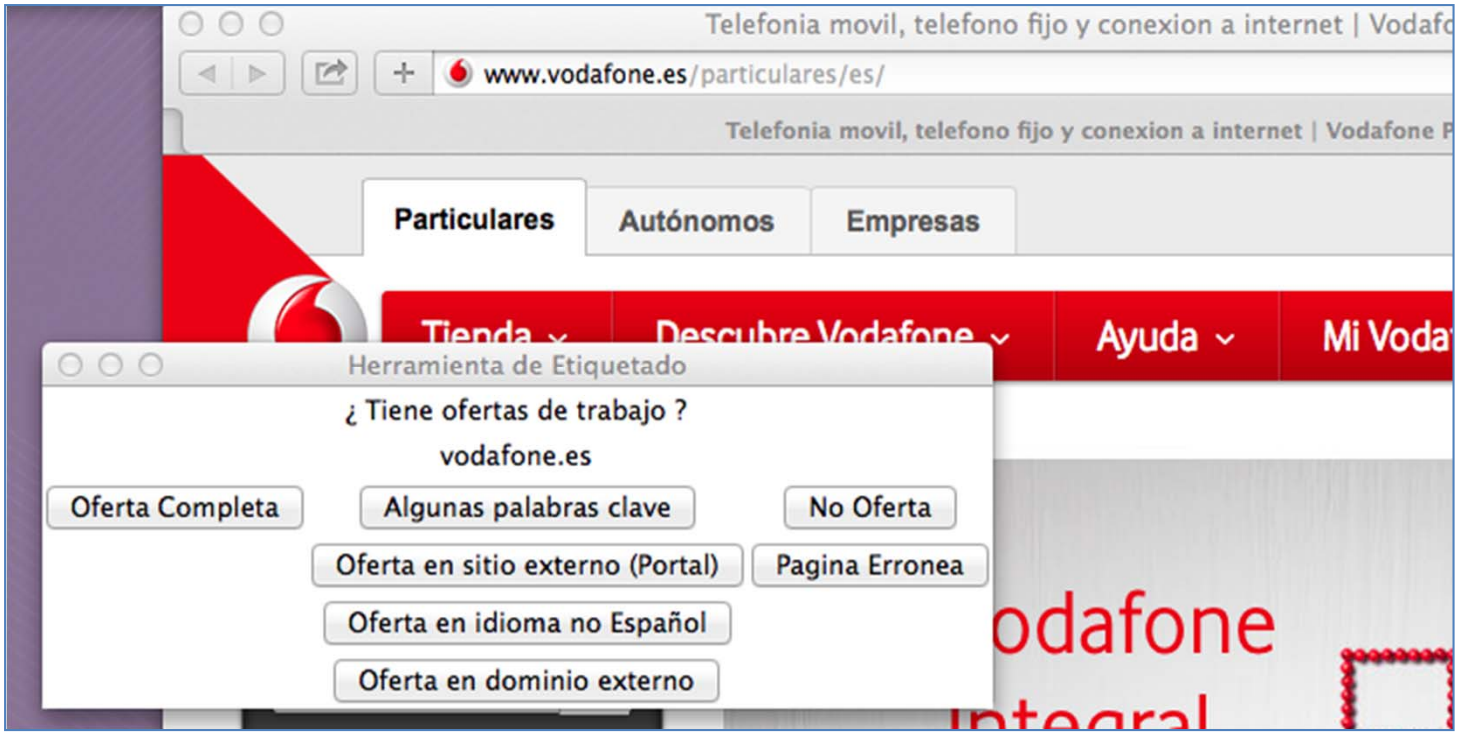
PÁGINAS WEB ETIQUETADAS

803

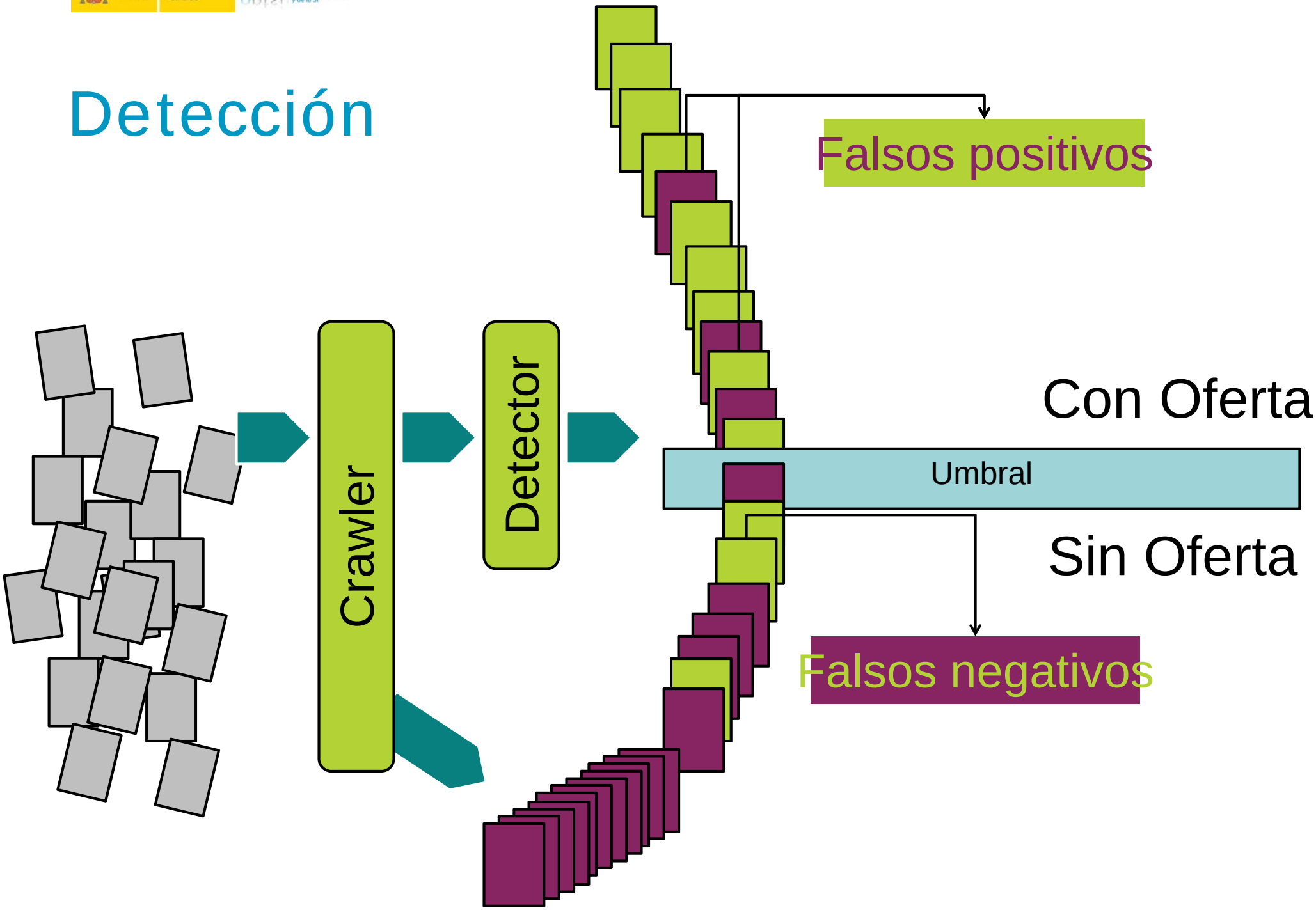
webs **SIN OFERTAS DE EMPLEO** perfilables

25

webs **CON OFERTAS de EMPLEO** perfilables, en castellano



Detección


Crawler
Detector
Falsos positivos
Con Oferta
Umbral
Sin Oferta
Falsos negativos

Perfilado de la oferta de empleo

Topic 0 (34,45%)	Topic 1 (25,25%)	Topic 2 (23,3%)	Topic 3 (10,8%)	Topic 4 (6,27%)
programacion	servicios	partner	csb	adjuntar
java	servidores	atsistemas	system	carreras
atsistemas	sql	consultar	aleman	multiplica
javascript	administracion	login	comercial	puntos
j2ee	unix	directions	ofertas	vista
php	linux	signup	marketing	disciplinas
drupal	vacantes	company	vitae	oral
sql	sap	position	progresar	significativos
analisis	informatica	summary	envienos	fueras

Resultados

http://laD.tsc.uc3m.es/

06.

Conclusiones

Conclusiones particulares

- El ML constituye una potente herramienta de análisis masivo de información a bajo coste, que no sustituye a otras fuentes de análisis, pero que las complementan
  - Permite la detección automática de B2C.
  - La combinación de rastreo inteligente + ML permite detectar ofertas de empleo en la web de las empresas.
  - Los algoritmos de perfilado automático identifican patrones generales en el conjunto de las oferta de empleo de las empresas, o de la oferta formativa.
- A pesar de las diferencias (en formato, vocabulario, estructura) de los corpus de datos, los algoritmos de matching pueden constituir una potente herramienta de análisis comparativo del ajuste global entre la oferta y la demanda de profesionales en España.

Lecciones

- Combinar:
  - El desarrollo de un buen sistema para escenarios de IaD requiere de la combinación eficiente de varias tecnologías (crawling, NLP, ML)
- Es importante optimizar cada paso:
  - Rastreo… selectivo
  - Procesamiento del lenguaje natural y de la estructura del documento
  - Clasificación… escalable.

Lecciones

- La intervención manual es importante.
  - Económica, sin exigencias tecnológicas
  - Criterios y metodología de etiquetado sistemático.
  - Metodología de supervisión de perfiles.
- El proceso de evaluación es importante
  - Evaluación objetiva.
  - Evaluación de los procesos manuales.

Potencial de mejoras

- Esto es versión 1.
  - Mejoras en el etiquetado
  - Procesamiento más eficiente del contenido web (javascript, segmentación de ofertas de empleo en la página web, mejoras en los modelos jerárquicos, aplicación de modelos dinámicos, etc)
  - Tratamiento avanzado de textos.

GOBIERNO DE ESPAÑA
MINISTERIO DE INDUSTRIA, TURISMO Y COMERCIO
ntsi

Muchas Gracias