UNIVERSIDAD AUTÓNOMA DE MADRID ESCUELA POLITÉCNICA SUPERIOR





TRABAJO FIN DE GRADO

RECOMENDACIÓN DE NOTICIAS EN LÍNEA BASADO EN TWITTER

ADRIÁN OTERO RODRÍGUEZ Tutor: ALEJANDRO BELLOGÍN KOUKI Ponente: PABLO CASTELLS AZPILICUETA

JULIO 2014

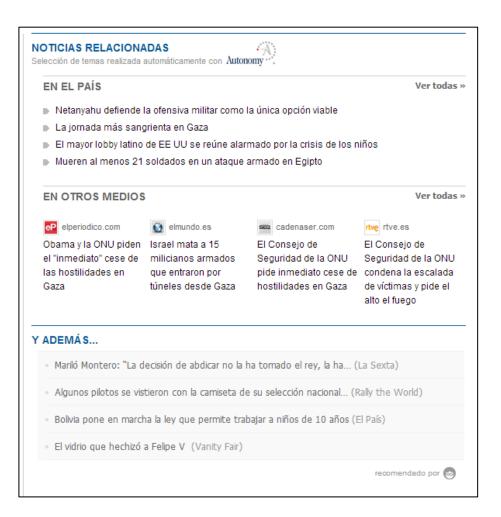
TABLA DE CONTENIDOS

- Introducción
- Objetivos
- Visión general
- Tecnologías utilizadas
- Análisis
- Diseño y desarrollo
- Pruebas y resultados
- Conclusiones
- Trabajo futuro
- Parte formativa

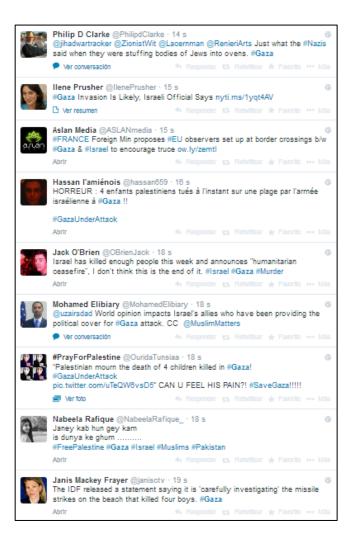
Diarios Digitales



Recomendación de noticias



Twitter



Twitter



- Una de las más importantes redes sociales.
- Mensajes de hasta 140 caracteres (tweets).
- Medidor de popularidad a través de retweets y favoritos.
- Genera un gran volumen de información útil.

Twitter



Propuesta

Un sistema de recomendación de

noticias basado en Twitter

Propuesta

Un sistema de recomendación de

noticias basado en Twitter

e integrarlo en un sistema real

Plista



- Entorno de recomendación real (integra usuarios, periódicos, desarrolladores).
- Envía las peticiones de recomendación a los sistemas registrados (como nosotros).
- Ofrece información de interés sobre el rendimiento del recomendador.

Plista

Interacción entre usuario final, web del publicador, plataforma Plista y desarrollador:

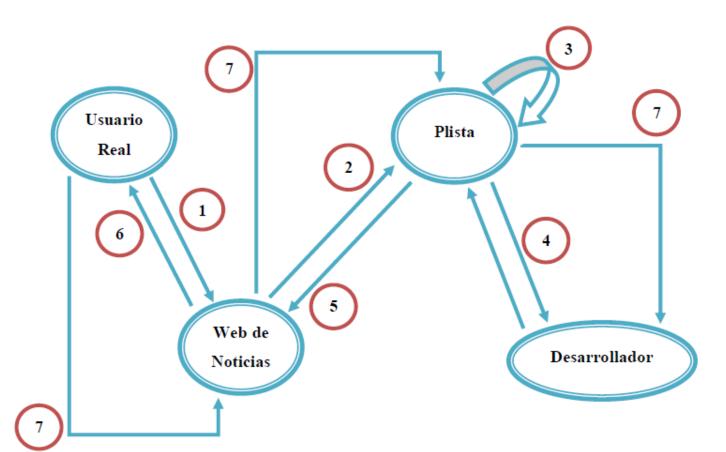


TABLA DE CONTENIDOS

- Introducción
- Objetivos
- Visión general
- Tecnologías utilizadas
- Análisis
- Diseño y desarrollo
- Pruebas y resultados
- Conclusiones
- Trabajo futuro
- Parte formativa

OBJETIVOS

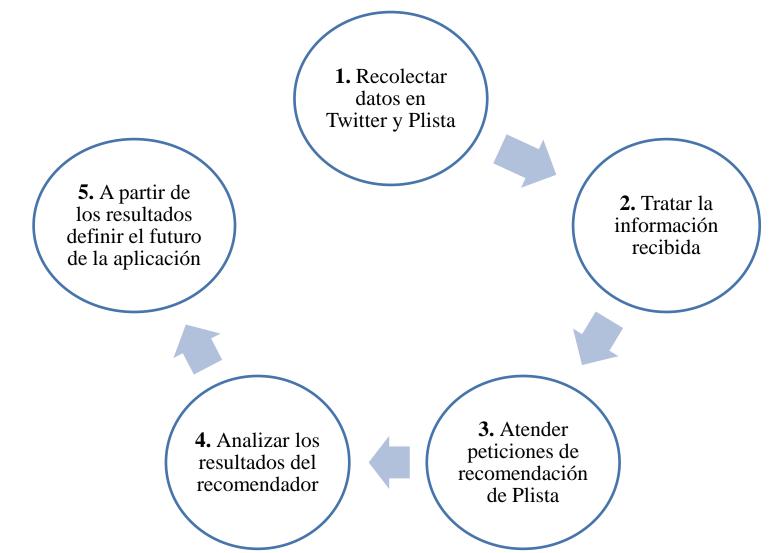


TABLA DE CONTENIDOS

- Introducción
- Objetivos
- Visión general
- Tecnologías utilizadas
- Análisis
- Diseño y desarrollo
- Pruebas y resultados
- Conclusiones
- Trabajo futuro
- Parte formativa

VISIÓN GENERAL

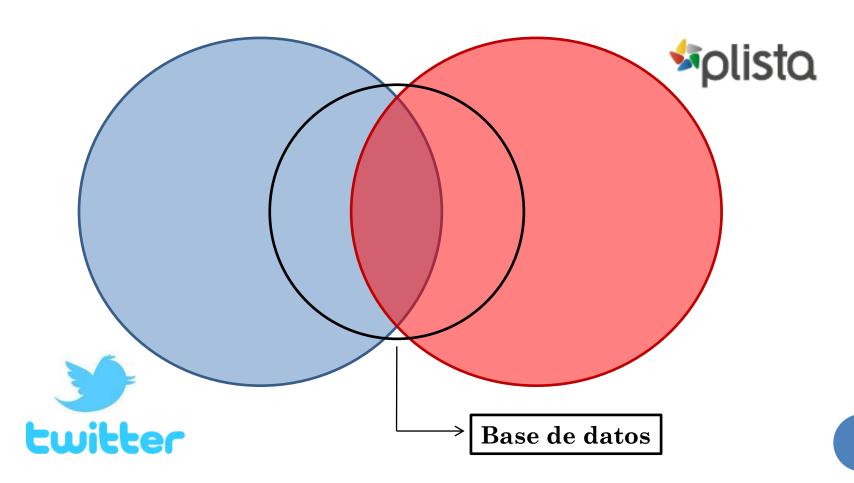
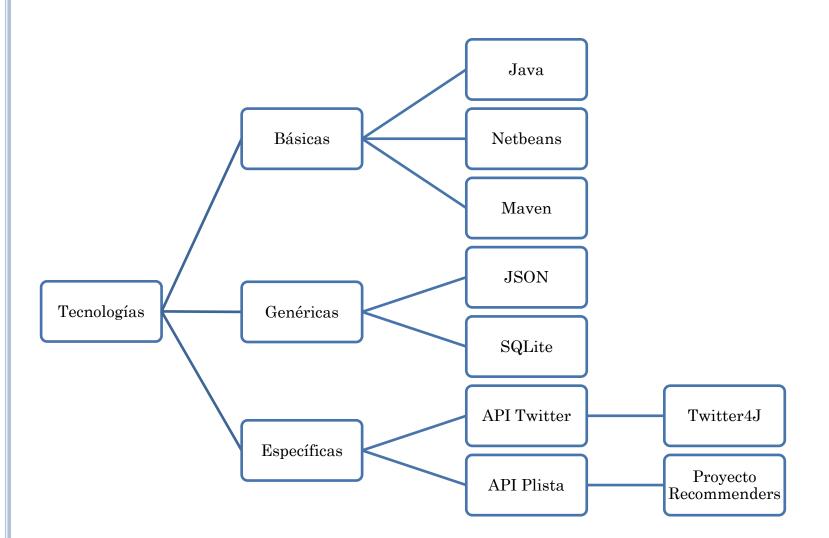


TABLA DE CONTENIDOS

- Introducción
- Objetivos
- Visión general
- <u>Tecnologías utilizadas</u>
- Análisis
- Diseño y desarrollo
- Pruebas y resultados
- Conclusiones
- Trabajo futuro
- Parte formativa



Tecnologías Básicas







Tecnologías Genéricas: JSON



```
<menu id="file" value="File">
     <popup>
        <menuitem value="New" onclick="CreateNewDoc()" />
        <menuitem value="Open" onclick="OpenDoc()" />
        <menuitem value="Close" onclick="CloseDoc()" />
        </popup>
        </menu>
```



Tecnologías Genéricas: SQLite

- Carece de un proceso servidor externo.
- Permite reducir los tiempos de acceso a la base de datos.
- No necesita ser instalado y configurado.
- Es un sistema de gestión de bases de datos transaccional.
- Garantiza la seguridad y consistencia de los datos involucrados.

Tecnologías específicas: API de Twitter y Twitter4J

Search API

Captura de información vía HTTP.

Búsqueda con filtros.

Limitación temporal de 7 días.

Rest API

Captura de información vía HTTP.

No está diseñada con la búsqueda a través de filtros como una prioridad.

Limitación temporal indefinida.

Streaming API

Captura de datos a través de un flujo constante de información.

Permite filtrar la información capturada.

Captura los tweets según son generados. Captura de datos

Filtros de búsqueda

Limitación temporal

Tecnologías específicas: API de Twitter y Twitter4J

Search API

Captura de información vía HTTP.

Búsqueda con filtros.

Limitación temporal de 7 días.



Tecnologías específicas: API de Plista y Proyecto Recommenders

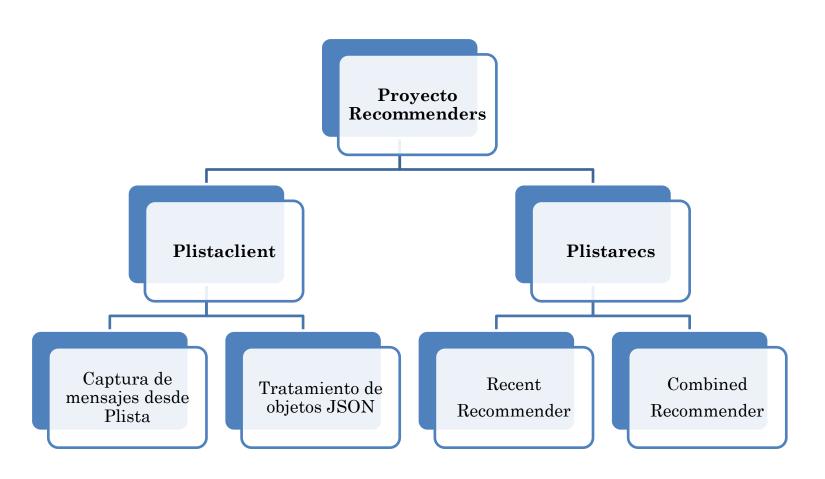


TABLA DE CONTENIDOS

- Introducción
- Objetivos
- Visión general
- Tecnologías utilizadas
- Análisis
- Diseño y desarrollo
- Pruebas y resultados
- Conclusiones
- Trabajo futuro
- Parte formativa

Análisis

Requisitos Funcionales

- ✓ RF01. Búsquedas en la API de Twitter.
- ✓ RF02. Capturar datos desde Plista.
- ✓ RF03. Extraer campos de un tweet para medir la popularidad.
- ✓ RF04. Extraer urls de un tweet.
- ✓ RF05. Tratar la información capturada desde Plista.
- ✓ RF06. Almacenar información extraída de Twitter.
- ✓ RF07. Almacenar ítems de Plista en la base de datos.
- ✓ RF08. Actualizar los ítems de Plista de la base de datos.
- ✓ RF09. Agrupar las urls de Twitter y ordenar por su popularidad.
- ✓ RF10. Asociar ítems de Plista y urls de Twitter.
- ✓ RF11. Atender peticiones de recomendación de Plista.

Análisis

Requisitos Funcionales

- ✓ RF01. Búsquedas en la API de Twitter.
- ✓ RF03. Extraer campos de un tweet para medir la popularidad.
- ✓ RF04. Extraer urls de un tweet.
- ✓ RF06. Almacenar información extraída de Twitter.
- ✓ RF09. Agrupar las urls de Twitter y ordenar por su popularidad.



- Olisto

- ✓ RF02. Capturar datos desde Plista.
- ✓ RF05. Tratar la información capturada desde Plista.
- ✓ RF07. Almacenar ítems de Plista en la base de datos.
- ✓ RF08. Actualizar los ítems de Plista de la base de datos.
- ✓ RF10. Asociar ítems de Plista y urls de Twitter.
- ✓ RF11. Atender peticiones de recomendación de Plista.

Análisis

Requisitos No Funcionales

- ✓ RNF01. Autentificación en la API de Twitter.
- ✓ RNF02. Limitación en el número de consultas contra la API de Twitter (180 consultas / 15 minutos).
- ✓ RNF03. Base de datos con suficiente capacidad para almacenar la información de Twitter.
- ✓ RNF04. Responder peticiones de Plista rápidamente (< 100ms).
- ✓ RNF05. Capaz de tratar miles de peticiones de Plista por segundo.
- ✓ RNF06. Suficiente capacidad de la base de datos para almacenar la información recibida desde Plista.
- ✓ RNF07. Fiabilidad, reducir errores al responder peticiones de Plista.

ANÁLISIS

Requisitos No Funcionales

- ✓ RNF01. Autentificación en la API de Twitter.
- ✓ RNF02. Limitación en el número de consultas contra la API de Twitter (180 consultas / 15 minutos).
- ✓ RNF03. Base de datos con suficiente capacidad para almacenar la información de Twitter.

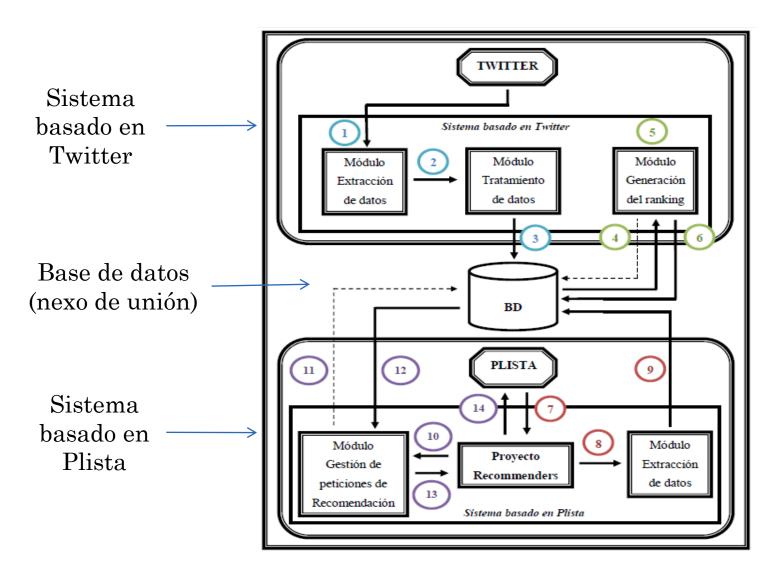


olisto

- ✓ RNF04. Responder peticiones de Plista rápidamente (< 100ms).
- ✓ RNF05. Capaz de tratar miles de peticiones de Plista por segundo.
- ✓ RNF06. Suficiente capacidad de la base de datos para almacenar la información recibida desde Plista.
- ✓ RNF07. Fiabilidad, reducir errores al responder peticiones de Plista.

TABLA DE CONTENIDOS

- Introducción
- Objetivos
- Visión general
- Tecnologías utilizadas
- Análisis
- Diseño y desarrollo
- Pruebas y resultados
- Conclusiones
- Trabajo futuro
- Parte formativa



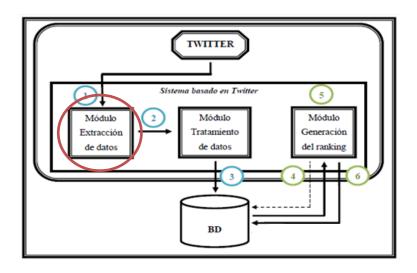
Twitter. Módulo de extracción de datos

Autentificación en la API de Twitter a través del protocolo OAuth.

Crear la consulta que se realizará sobre Twitter (estas consultas están almacenadas en la base de datos).

Ejecución de la consulta a utilizando la librería Twitter4J.

Almacenar los resultados de la consulta en un objeto Java.



En este módulo el uso de la librería Twitter4J es fundamental

Twitter. Módulo de extracción de datos

Resultado de una consulta a través de la aplicación

-		RETWEETS	FAVOURITES	TEXT	USERNAME
	18998	0	0	"@20m: El extesorero del #PP Álvaro #Lapuerta afirma que nunca se ha apropiado de	Moran_Fernandez
	18999	0	1	La Fiscalía Anticorrupción pide imputar al sobrino de Felipe González en el caso	Egarenca
	19000	3	0	El extesorero del #PP Álvaro #Lapuerta afirma que nunca se ha apropiado de dinero JAJAJA #ascodepeperos	MARIAJ_SALADO
	19001	0	0	@20m: El extesorero del #PP Álvaro #Lapuerta afirma que nunca se ha apropiado de	Antonio_314
1	19002	9	2	#ÚLTIMAHORA El extesorero del #PP Álvaro #Lapuerta afirma que nunca se ha	20m
	19003	0	0	Sólo 650.000 viviendas han obtenido el certificado energético en su primer año en vigor	gabrielavera08
	19004	0	0	Sólo 650.000 viviendas han obtenido el certificado energético en su primer año en vigor	josefina_silva9



Resultado de la misma consulta a través del sitio web de Twitter

Twitter. Módulo de tratamiento de datos

Utilizando como punto de partida la información previamente capturada desde Twitter



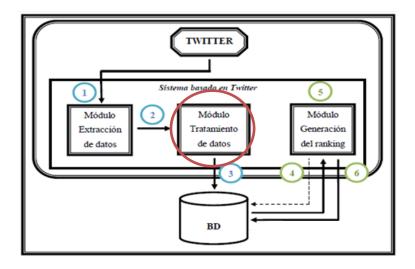
Tratar la información recibida y filtrar los atributos innecesarios.



Se conservarán los atributos: id. del tweet, url del tweet, nº retweets, nº favoritos, fecha de creación, texto del tweet, id. del usuario y nombre del usuario.



Introducir esta información en la base de datos.



Los distintos **tweets**capturados **se identifican mediante** su **id**, en caso
de que ya exista en la
base de datos se **actualizará** la
información

Twitter. Módulo de generación del ranking de popularidad

Se filtra la información de la base de datos conservando:

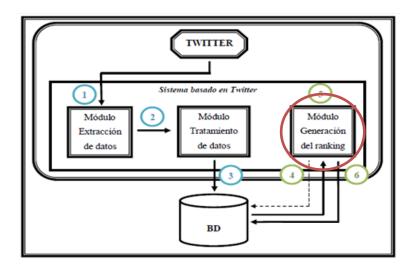
url + id. del dominio + nº retweets + nº favoritos



Se agrupan cada uno de los pares: [url + id. del dominio] sumando su número de retweets y favoritos



Se ordenan los pares en función de su popularidad (nº retweets + nº favoritos) y se almacena en la base de datos



Este proceso se inicia tras finalizar la búsqueda en Twitter de cada una de las consultas almacenadas en la base de datos.

Twitter. Módulo de generación del ranking de popularidad

Resultado de aplicar el método de actualización del ranking de popularidad sobre la base de datos



	URL	DOMAINID	RETWEETS	FAVOURITES	•
1	http://www.20minutos.es/noticia/1346662/0/marido/soraya-saenz-de-santamaria/telefonica/	1	324	57	
2	http://www.20minutos.es/noticia/2145622/0/bankia/rato/embargo/#xtor=AD-158xts=467263	1	187	31	
3	http://www.20minutos.es/noticia/2184521/0/alcaldes/sobresueldos/caso-mercurio/	1	38	9	
4	http://www.20minutos.es/noticia/2183255/0/juanjo-puigcorbe/boadilla-premio-cortos/veto-presiones-politicas/	1	15	6	
5	http://www.20minutos.es/noticia/2184605/0/pablo-iglesias/libro-entrevista/regular-medios-comunicacion/	1	8	9	
6	http://www.20minutos.es/noticia/2145622/0/bankia/rato/embargo/	1	10	5	
7	http://www.20minutos.es/deportes/noticia/seleccion-argelia-futbol-donacion-primas-gaza-mundial-brasil-2014-2184379/0/	1	10	4	+
4	III.				

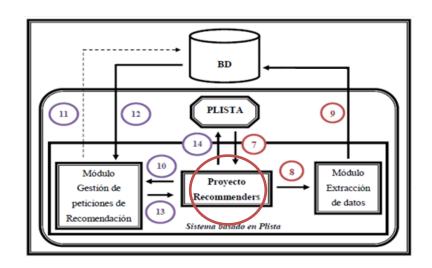
Plista. Proyecto Recommenders

Captura e identificación de los mensajes recibidos desde Plista

Tratamiento de los objetos JSON recibidos desde Plista

Implementación de los recomendadores:

- Recent Recommender
- · Combined Recommender



Plistarecs y Plistaclient son proyectos Open Source, se encuentran disponibles en:

https://github.com/recommenders/plistarecs https://github.com/recommenders/plistaclient

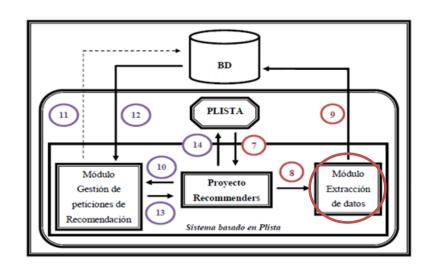
Plista. Módulo de extracción de datos

Plista envía un "Item Update"

> A través de Plistaclient se captura el ítem recibido

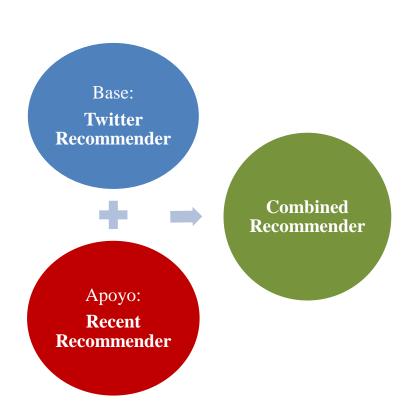
> > Se extrae la información útil del mensaje recibido

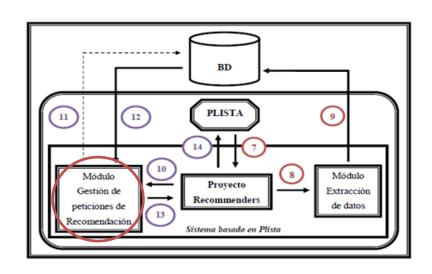
> > > Se inserta esta nueva información en la base de datos



En muchos casos no se tratará de nuevos ítems, si no actualizaciones de los ya existentes

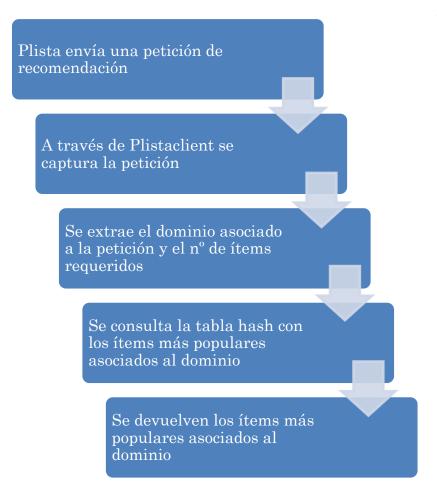
Plista. Módulo de gestión de peticiones de recomendación

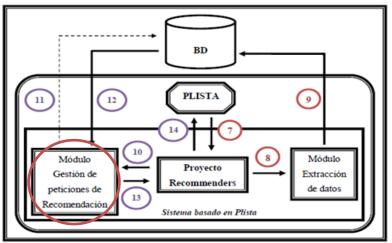




Inicialmente el Twitter
Recommender no dispone de
suficientes ítems para
realizar las recomendaciones
adecuadas

Plista. Módulo de gestión de peticiones de recomendación





La tabla hash será
actualizada cada 100
peticiones o cuando no
existan suficientes ítems
para el dominio asociado
(Para actualizarla se utiliza
la base de datos)

Base de datos SQL

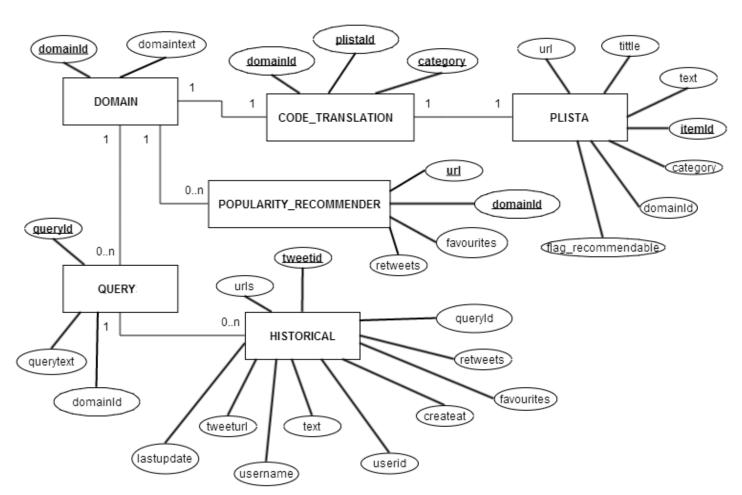


TABLA DE CONTENIDOS

- Introducción
- Objetivos
- Visión general
- Tecnologías utilizadas
- Análisis
- Diseño y desarrollo
- Pruebas y resultados
- Conclusiones
- Trabajo futuro
- Parte formativa

Pruebas

Pruebas realizadas sobre el sistema

Pruebas Unitarias Pruebas de Integración

Pruebas de Sistema y Aceptación

Sobre cada módulo del sistema Sobre el sistema basado en Twitter Sobre el sistema basado en Plista Prueba individual (entorno básico) Prueba con un grupo pequeño de usuarios (entorno medio)

Prueba en un entorno real (entorno grande)

Pruebas

Pruebas realizadas sobre el sistema

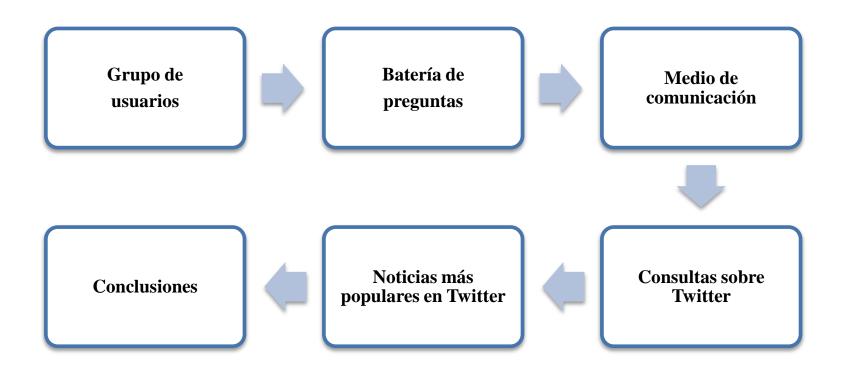
Pruebas Unitarias Pruebas de Integración

Pruebas de Sistema y Aceptación

Sobre cada módulo del sistema Sobre el sistema basado en Twitter Sobre el sistema basado en Plista Prueba individual (entorno básico) Prueba con un grupo pequeño de usuarios (entorno medio)

Prueba en un entorno real (entorno grande)

Resultados (grupo pequeño de usuarios)



Resultados (grupo pequeño de usuarios)

Grupo de usuarios

- · Población: 8 usuarios
- Genero: 5 hombres y 3 mujeres
- Edad: 18 50 años

Batería de preguntas

- ¿Te han resultado interesantes las noticias mostradas?
- Observando las noticias disponibles en la web del Publicador, ¿Te hubiera resultado más interesante otra selección de noticias distinta a la recomendada?

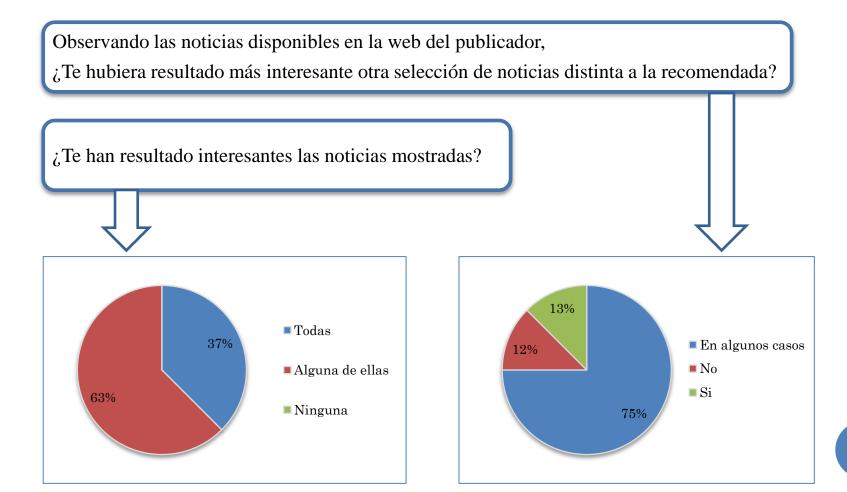
Medio de comunicación

• 20Minutos (http://www.20minutos.es/)

Consultas sobre Twitter

- http://www.20minutos.es/
- 20minutos.es
- 20minutos

Resultados (grupo pequeño de usuarios)



Resultados (entorno real)

- * 250.000 peticiones de recomendación.
- ❖ 3.200 clics (recomendaciones aceptadas por los usuarios).
- ❖ 1.28% de tasa de acierto.

	Peticiones	Clics	Tasa (%)
Día 1	2.145	29	1,35
Día 2	35.975	514	1,43
Día 3	58.574	807	1,38
Día 4	27.465	314	1,14
Día 5	1.877	23	1,23
Día 6	8.239	121	1,47
Día 7	52.868	540	0,95
Día 8	13.146	103	0,78
Día 9	45.663	703	1,54
TOTAL	245.952	3.154	1,28

Resultados (entorno real)



Evolución diaria del **TwitterRecommender** en Plista

Evolución diaria del

Recent Recommender

en Plista

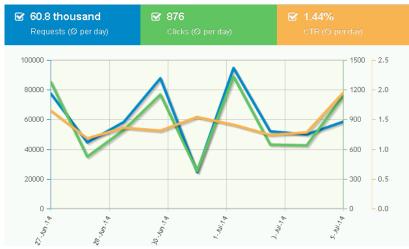


TABLA DE CONTENIDOS

- Introducción
- Objetivos
- Visión general
- Tecnologías utilizadas
- Análisis
- Diseño y desarrollo
- Pruebas y resultados
- Conclusiones
- Trabajo futuro
- Parte formativa

CONCLUSIONES

Objetivos

• Se han cumplido todos los objetivos marcados al inicio de este proyecto satisfactoriamente.

Restricciones

- Entorno real para las pruebas.
- Publicadores con poca presencia en Twitter.

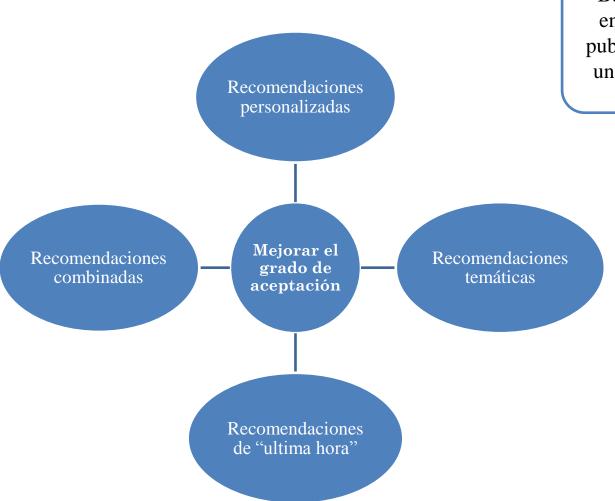
Resultados

- Positivos en las pruebas sobre un grupo pequeño de usuarios.
- Aceptables en un entorno real, a pesar de la falta de cobertura de algunos de los publicadores en Twitter.

TABLA DE CONTENIDOS

- Introducción
- Objetivos
- Visión general
- Tecnologías utilizadas
- Análisis
- Diseño y desarrollo
- Pruebas y resultados
- Conclusiones
- Trabajo futuro
- Parte formativa

TRABAJO FUTURO

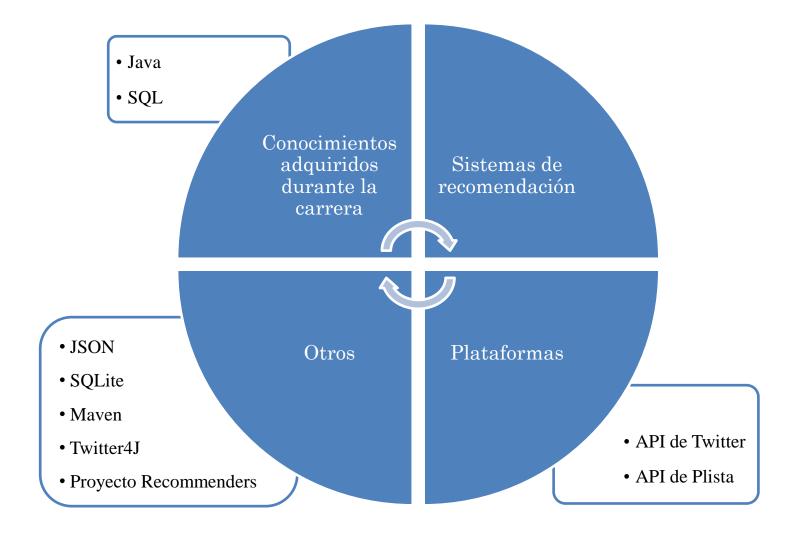


Búsqueda de un nuevo entorno real donde los publicadores cuenten con una mayor presencia en Twitter

TABLA DE CONTENIDOS

- Introducción
- Objetivos
- Visión general
- Tecnologías utilizadas
- Análisis
- Diseño y desarrollo
- Pruebas y resultados
- Conclusiones
- Trabajo futuro
- Parte formativa

PARTE FORMATIVA

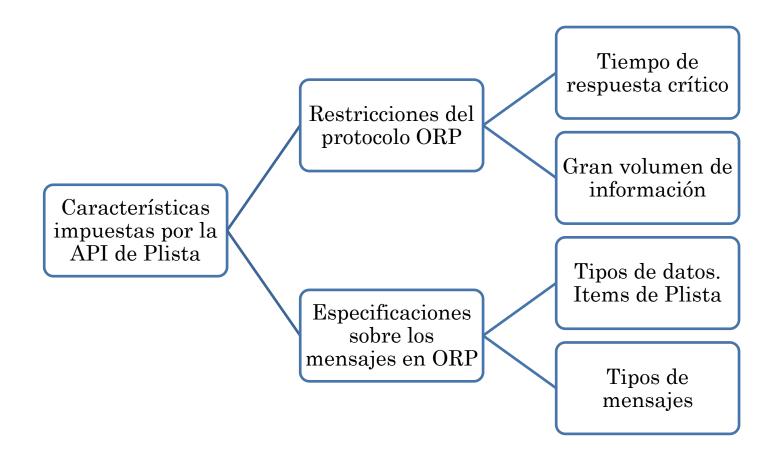




Muchas gracias

TECNOLOGÍAS UTILIZADAS

API DE PLISTA



Resultados (grupo pequeño de usuarios)

Noticia	Retweets	Favoritos
[1] Telefónica ficha al marido de Soraya Sáenz de Santamaría y a la esposa de Eduardo Madina	324	57
[2] Anticorrupción se opone al embargo de 24.000 millones de euros a Rodrigo Rato y su equipo	187	31
[3] Investigan a 44 alcaldes de toda Cataluña y casi todos los partidos por cobrar sobresueldos	68	19
[4] Culpan al alcalde de Boadilla de vetar a Puigcorbé y Tito Valverde sopesa devolver el premio	23	12
[5] Pablo Iglesias, a favor de regular los medios de comunicación	25	6

Módulos del sistema

- Extracción de datos desde Twitter y tratamiento de la información.
- Generación del ranking con las noticias más populares en Twitter.
- Extracción de datos desde Plista y tratamiento de la información.
- Gestión de las peticiones de recomendación de Plista.

Introducción

Diarios Digitales

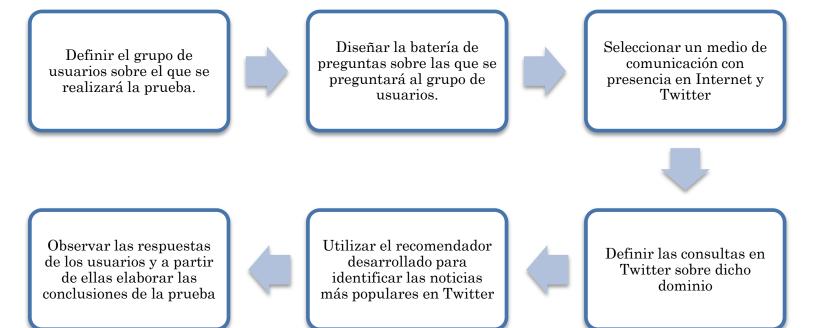


Introducción

Diarios Digitales

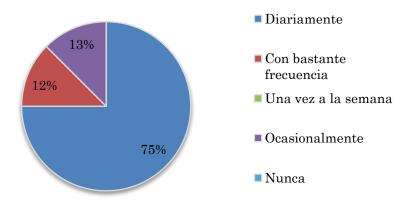


Resultados (grupo pequeño de usuarios)



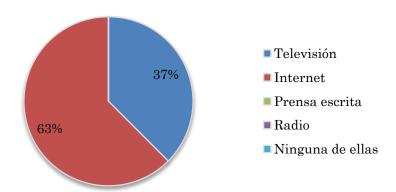
Resultados (grupo pequeño de usuarios)

Pregunta 1. ¿Con que frecuencia accedes a contenidos informativos de actualidad? Tanto prensa escrita como Internet, radio o televisión.



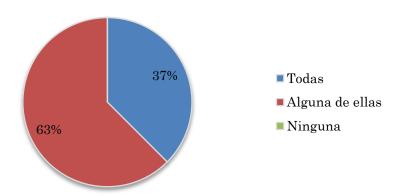
Resultados (grupo pequeño de usuarios)

Pregunta 2. ¿Cuál es la fuente que utilizas habitualmente como canal de comunicación?



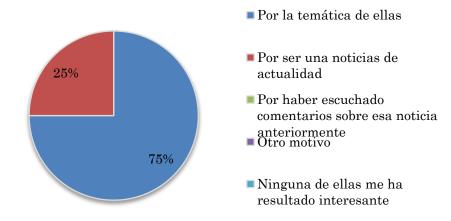
Resultados (grupo pequeño de usuarios)

Pregunta 3. ¿Te han resultado interesantes las noticias mostradas?



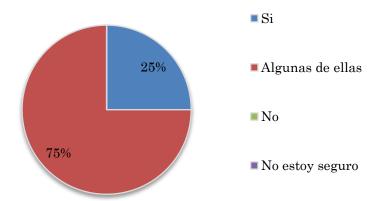
Resultados (grupo pequeño de usuarios)

Pregunta 4. ¿Por qué crees que esas noticias te han resultado interesantes?



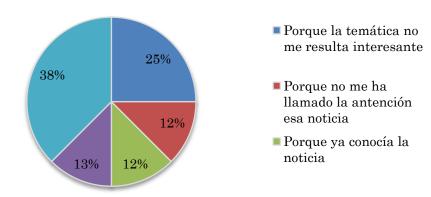
Resultados (grupo pequeño de usuarios)

Pregunta 5. Sobre las noticias que te han resultado interesantes, ¿Crees que hubieras accedido a ellas simplemente viendo su entradilla?



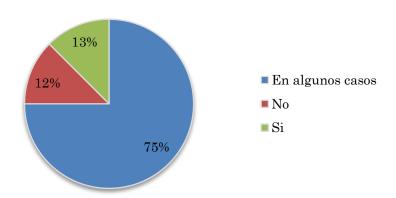
Resultados (grupo pequeño de usuarios)

Pregunta 6. ¿Por qué crees que esas noticias no te han resultado interesantes?



Resultados (grupo pequeño de usuarios)

Pregunta 7. Observando las noticias disponibles en la web del publicador, ¿Te hubiera resultado más interesante otra selección de noticias distinta a la recomendada?



Resultados (grupo pequeño de usuarios)

Pregunta 8. En cuanto a las noticias recomendadas, ¿Cuál crees que es el elemento que más condiciona para visitar alguna de esas noticias?

