

**UNIVERSIDAD DE GRANADA**  
**E.T.S.I INFORMÁTICA Y TELECOMUNICACIÓN**



**Recuperación de Información (RI)**

**Implementación de un Sistema de  
Recuperación de Información utilizando  
Lucene**

**Curso 2017-2018**

**Cuarto Curso del Grado en Ingeniería Informática**

María Camarero Granados

Javier Gómez Luzón

Francisco Porcel Molina

# Índice

Análisis previo de los requisitos

3

Interfaz de búsqueda

3

Analizador

4

Facetas

4

Diseño de la solución

5

Elección de facetas

7

Manual de usuario

8

Aplicación Lucene

8

Luke 11

# Análisis previo de los requisitos

Se nos pide implementar un programa que permita añadir documentos a un índice creado utilizando la biblioteca Lucene, que posteriormente formara parte de un Sistema de Recuperación de Información que servirá para buscar artículos científicos.

Además, se debe utilizar Luke para ver el índice y realizar distintas consultas sobre él para demostrar que se ha creado de forma adecuada.

## Interfaz de búsqueda

Se debe saber qué y cómo se va a buscar en nuestra interfaz de búsqueda para así saber qué campos de los documentos con los que se trabaja y de qué forma van a ser tenidos en cuenta para elaborar el índice.

Se deben identificar, al menos, los siguientes tipos de campos sobre los documentos de entrada:

- StringField: Texto simple que no se tokeniza. Útil para la búsqueda por facetas y para el filtrado de consultas.

- TextField: Términos que son procesados para la indexación.

- Numérico.

- Facetas (Categorías)

Además, en la aplicación se deben poder realizar:

- Consultas booleanas que involucre a los operadores lógicos OR, AND o NOT.

- Consultas avanzadas (consultas por proximidad, por ejemplo)

- Presentaciones de la información utilizando distintos criterios de ordenación.

## **Analizador**

Hay que identificar qué términos serán utilizados en la búsqueda y cuales no.

Para ello, se debe seleccionar entre los tipos de analizador que tiene implementados Lucene el más adecuado para la aplicación, justificando la decisión. Es posible que haya que utilizar un analizador distinto para cada uno de los campos a indexar.

## **Facetas**

Se deberá realizar la búsqueda por facetas. Para conseguirlo, se deben identificar los campos por los que se podrá clasificar los documentos.

Además, como resultado de la búsqueda podremos tener los resultados agrupados por categorías.

# Diseño de la solución

Ruta del directorio de Index

Ruta del directorio de CategoryIndex

Comenzar búsqueda

Campo de búsqueda general

Búsqueda por título

Búsqueda por autor

Búsqueda por resumen

☐ AND ☐ Ordenación por relevancia

☐ OR ☐ Ordenación por año

Año

Año

☐ Rango de fechas

☐ Fechas específicas

Nuestra interfaz tendrá un aspecto similar al de la imagen, y, en consecuencia, deberemos implementar la creación del índice de una determinada manera, que describimos a continuación.

En nuestra solución, incluimos todos los campos de los que constan los registros sobre artículos científicos de la base de datos Scopus que hemos utilizado.

De esta forma, podemos devolver toda la información y no nos vemos limitados a la hora de trabajar más adelante con los distintos tipos de consultas.

Para los campos en los que tenemos previsto realizar una búsqueda general o realizar una búsqueda por campo específico utilizamos el analizador EnglishAnalyzer, debido a que tanto el título, como el resumen o los distintos keywords se encuentran en inglés y deben ser tokenizados para facilitar que nuestras búsquedas obtengan buenos resultados (EnglishAnalyzer se ocupa de eliminar palabras vacías, hacer stemming, etc.).

Para el campo EID utilizamos el KeywordAnalyzer para que se indexe la cadena de texto sin dividirla de ningún modo. Esto se hace porque normalmente este campo sólo va a ser usado cuando se conoce el EID de un libro concreto y, por lo tanto, se introduce tal cual en el buscador.

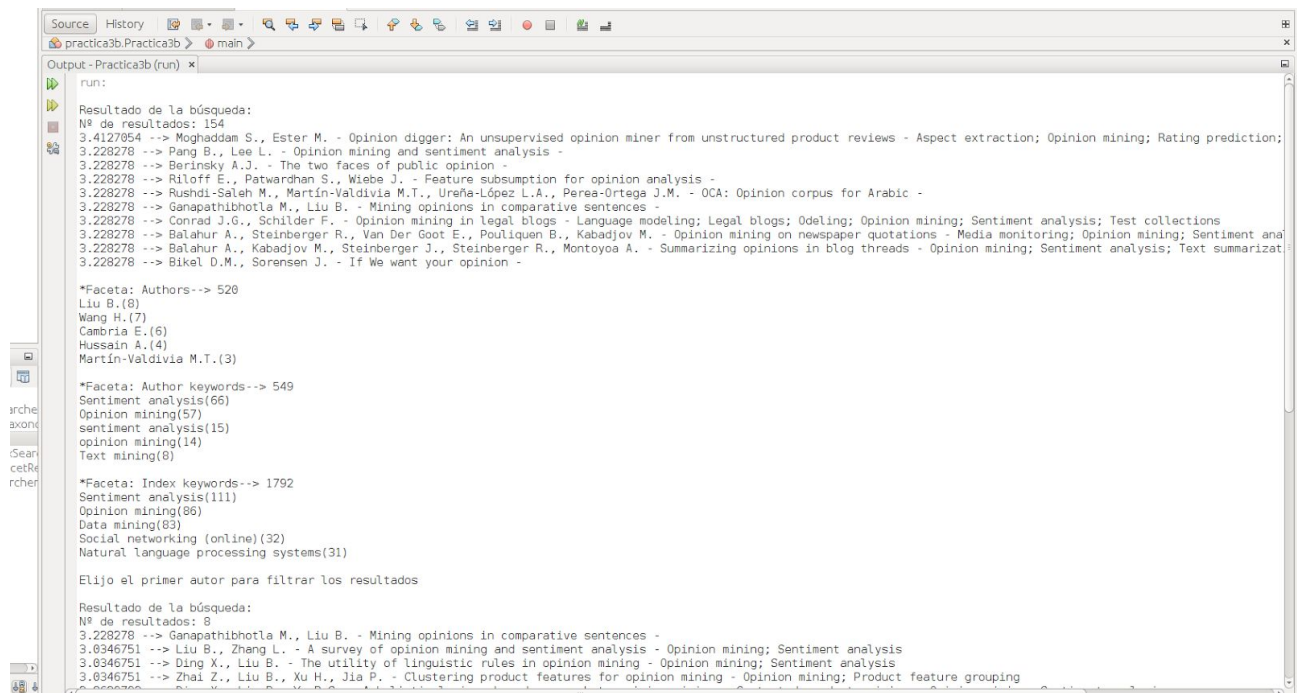
Por último, utilizamos el StandardAnalyzer para el resto de campos, por si es necesario buscar artículos a partir de ellos.

Esta asignación de analizadores se realiza mediante PerFieldanalyzerWrapper en la función cargaAnalyzer.

Hemos seleccionado como facetas los autores de los artículos, las keywords de los autores y las keywords para indexar. Estos valores nos serán muy útiles para clasificar la información que nos devuelvan las búsquedas, ya que tienen un gran poder de agrupación y, además, nos permitirán visualizar la información de forma interesante.

# Elección de facetas

Escogemos aquellos campos que nos permiten agrupar de la forma más adecuada los resultados de la búsqueda. El siguiente ejemplo de ejecución ilustra una posible serie de pasos cuando se pretende buscar un artículo mediante la búsqueda inicial de “opinion” en el título:



The screenshot shows a software window titled "practica3b.Practica3b" with a "main" tab. The "Output - Practica3b (run)" pane displays the following text:

```
run:
Resultado de la búsqueda:
Nº de resultados: 154
3.4127054 --> Moghaddam S., Ester M. - Opinion digger: An unsupervised opinion miner from unstructured product reviews - Aspect extraction; Opinion mining; Rating prediction;
3.228278 --> Pang B., Lee L. - Opinion mining and sentiment analysis -
3.228278 --> Berinsky A.J. - The two faces of public opinion -
3.228278 --> Riloff E., Patwardhan S., Wiebe J. - Feature subsumption for opinion analysis -
3.228278 --> Rushdi-Saleh M., Martín-Valdivia M.T., Ureña-López L.A., Perea-Ortega J.M. - OCA: Opinion corpus for Arabic -
3.228278 --> Ganapathibhotla M., Liu B. - Mining opinions in comparative sentences -
3.228278 --> Conrad J.G., Schilder F. - Opinion mining in legal blogs - Language modeling; Legal blogs; Odeling; Opinion mining; Sentiment analysis; Text collections
3.228278 --> Balahur A., Steinberger R., Van Der Goot E., Pouliquen B., Kabadjov M. - Opinion mining on newspaper quotations - Media monitoring; Opinion mining; Sentiment ana
3.228278 --> Balahur A., Kabadjov M., Steinberger J., Steinberger R., Montoya A. - Summarizing opinions in blog threads - Opinion mining; Sentiment analysis; Text summarizat
3.228278 --> Bikel D.M., Sorensen J. - If We want your opinion -

*Faceta: Authors--> 520
Liu B.(8)
Wang H.(7)
Cambria E.(6)
Hussain A.(4)
Martín-Valdivia M.T.(3)

*Faceta: Author keywords--> 549
Sentiment analysis(66)
Opinion mining(57)
sentiment analysis(15)
opinion mining(14)
Text mining(8)

*Faceta: Index keywords--> 1792
Sentiment analysis(111)
Opinion mining(86)
Data mining(83)
Social networking (online)(32)
Natural language processing systems(31)

Elijo el primer autor para filtrar los resultados

Resultado de la búsqueda:
Nº de resultados: 8
3.228278 --> Ganapathibhotla M., Liu B. - Mining opinions in comparative sentences -
3.0346751 --> Liu B., Zheng L. - A survey of opinion mining and sentiment analysis - Opinion mining; Sentiment analysis
3.0346751 --> Ding X., Liu B. - The utility of linguistic rules in opinion mining - Opinion mining; Sentiment analysis
3.0346751 --> Zhai Z., Liu B., Xu H., Jia P. - Clustering product features for opinion mining - Opinion mining; Product feature grouping
```

```
Source History
practica3b.Practica3b main
Output - Practica3b (run)
Elijo el primer autor para filtrar los resultados

Resultado de la búsqueda:
Nº de resultados: 8
3.228278 --> Ganapathibhotla M., Liu B. - Mining opinions in comparative sentences -
3.8346751 --> Liu B., Zhang L. - A survey of opinion mining and sentiment analysis - Opinion mining; Sentiment analysis
3.8346751 --> Ding X., Liu B. - The utility of linguistic rules in opinion mining - Opinion mining; Sentiment analysis
3.8346751 --> Zhai Z., Liu B., Xu H., Jia P. - Clustering product features for opinion mining - Opinion mining; Product feature grouping
2.8629792 --> Ding X., Liu B., Yu P.S. - A holistic lexicon-based approach to opinion mining - Context dependent opinions; Opinion mining; Sentiment analysis
2.8629792 --> Liu B. - Sentiment analysis: Mining opinions, sentiments, and emotions -
2.8629792 --> Ding X., Liu B., Zhang L. - Entity discovery and assignment for opinion mining applications - Entity discovery; Sentiment analysis
2.5719478 --> Qiu G., Liu B., Bu J., Chen C. - Opinion word expansion and target extraction through double propagation -

*Faceta: Author keywords--> 11
Opinion mining(4)
Sentiment analysis(4)
Context dependent opinions(1)
Product feature grouping(1)
Entity discovery(1)

*Faceta: Index keywords--> 55
Sentiment analysis(5)
Problem solving(5)
Opinion mining(4)
Data mining(3)
Semantics(2)

Elijo el primer autor keyword para filtrar los resultados

Resultado de la búsqueda:
Nº de resultados: 4
3.8346751 --> Liu B., Zhang L. - A survey of opinion mining and sentiment analysis - Opinion mining; Sentiment analysis
3.8346751 --> Ding X., Liu B. - The utility of linguistic rules in opinion mining - Opinion mining; Sentiment analysis
3.8346751 --> Zhai Z., Liu B., Xu H., Jia P. - Clustering product features for opinion mining - Opinion mining; Product feature grouping
2.8629792 --> Ding X., Liu B., Yu P.S. - A holistic lexicon-based approach to opinion mining - Context dependent opinions; Opinion mining; Sentiment analysis

*Faceta: Index keywords--> 36
Opinion mining(3)
Sentiment analysis(3)
Data mining(3)
Problem solving(3)
Semantics(2)
BUILD SUCCESSFUL (total time: 0 seconds)
```

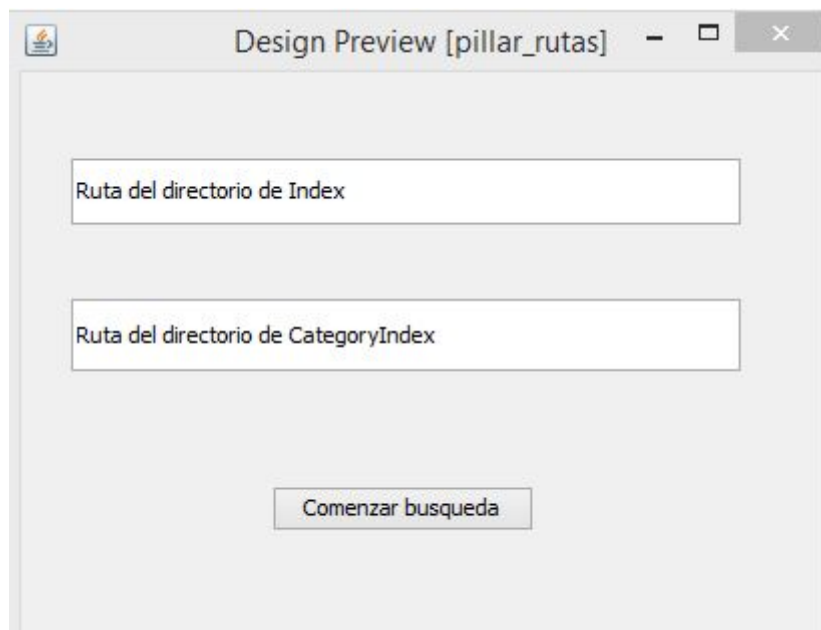


# Manual de usuario

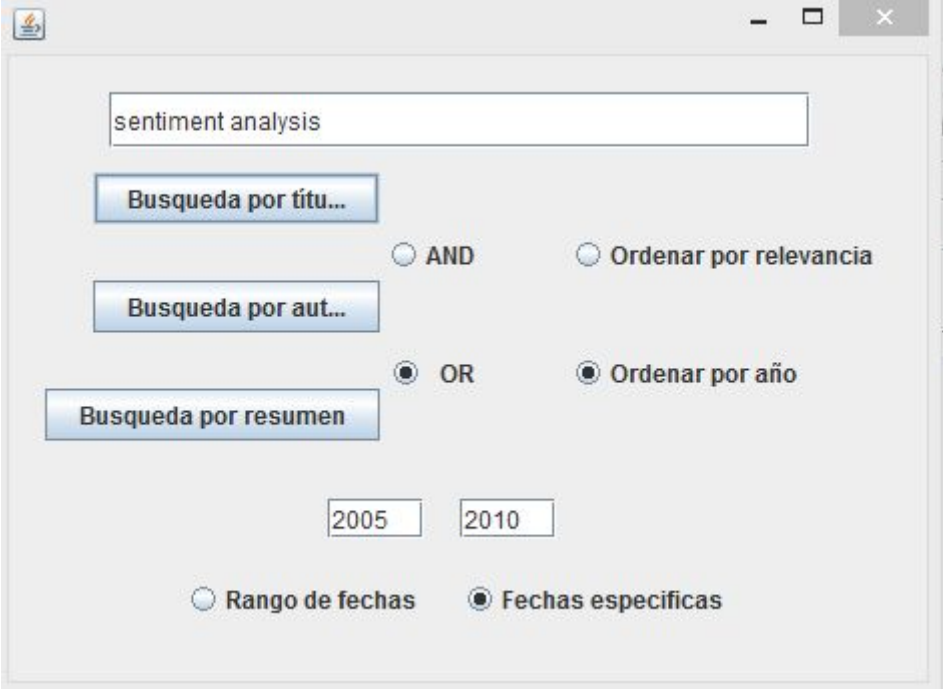
## Aplicación Lucene

Para la implementación del programa que se ocupa de crear el índice y de realizar las búsquedas se ha utilizado el IDE NetBeans.

El mismo se puede utilizar para su ejecución de forma sencilla, para lo cual hace falta ejecutar el programa y, en primer lugar, especificar como argumentos las rutas de los índices (index y categoryIndex) en la interfaz.



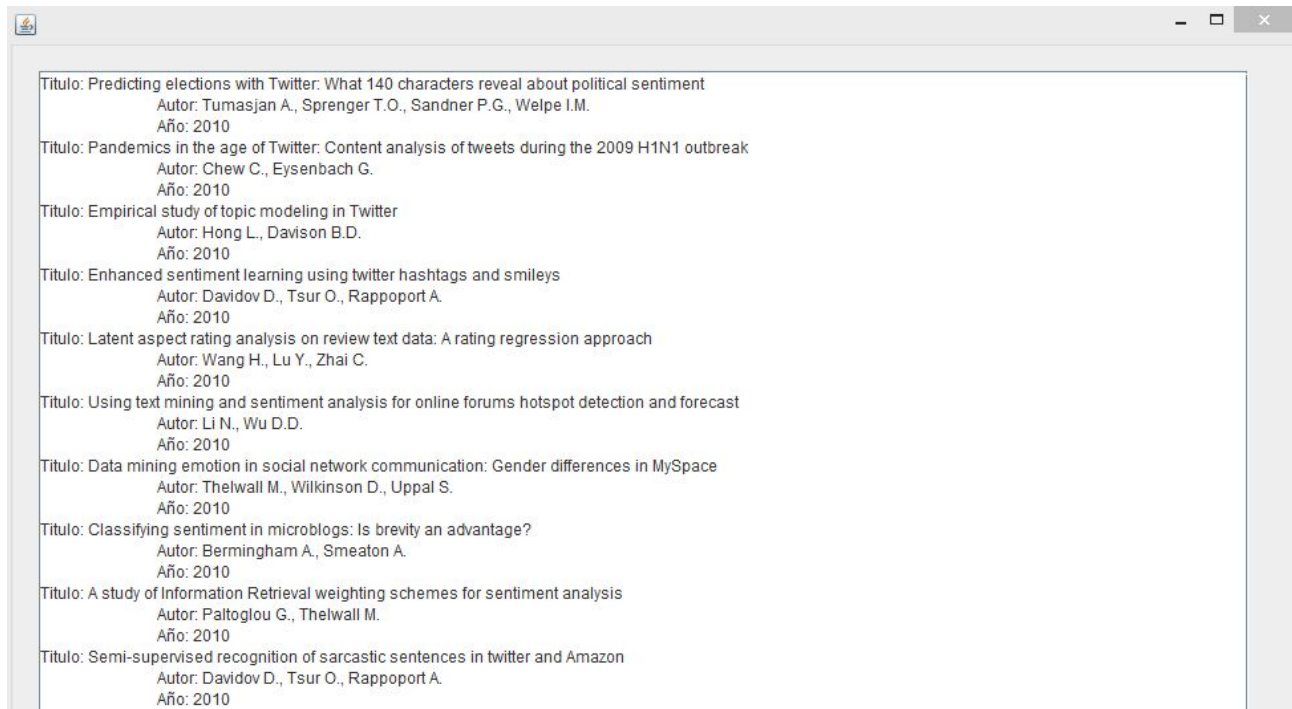
Tras esto, se pueden realizar búsquedas mediante las distintas opciones que ofrece la interfaz. Por ejemplo:



The screenshot shows a search interface window with a title bar containing a small icon and standard window controls (minimize, maximize, close). The main content area includes a text input field at the top containing the text "sentiment analysis". Below this field are three buttons: "Busqueda por título...", "Busqueda por autor...", and "Busqueda por resumen". To the right of these buttons are two columns of radio button options. The first column contains "AND" (unselected), "OR" (selected), and "Rango de fechas" (unselected). The second column contains "Ordenar por relevancia" (unselected), "Ordenar por año" (selected), and "Fechas específicas" (selected). Below the radio buttons are two text input fields for dates, containing "2005" and "2010".

La consulta se escribe en el campo de texto, tras lo cual se selecciona (si se desea) uno de los botones para realizar una búsqueda en un campo específico de los documentos. Se elige entre los operadores lógicos disponibles, el orden en el que van a aparecer los documentos relevantes a la consulta y, tras ingresar unas fechas opcionalmente, se decide entre realizar una búsqueda por rango o por fechas específicas

El resultado que se obtiene con la consulta anterior es el siguiente:



A screenshot of a window with a title bar containing a small icon on the left and standard minimize, maximize, and close buttons on the right. The window displays a list of ten research papers, each with its title, author(s), and year (2010). The papers are listed in a plain text format with line breaks separating the fields.

Titulo: Predicting elections with Twitter: What 140 characters reveal about political sentiment  
Autor: Tumasjan A., Sprenger T.O., Sandner P.G., Welpe I.M.  
Año: 2010

Titulo: Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak  
Autor: Chew C., Eysenbach G.  
Año: 2010

Titulo: Empirical study of topic modeling in Twitter  
Autor: Hong L., Davison B.D.  
Año: 2010

Titulo: Enhanced sentiment learning using twitter hashtags and smileys  
Autor: Davidov D., Tsur O., Rappoport A.  
Año: 2010

Titulo: Latent aspect rating analysis on review text data: A rating regression approach  
Autor: Wang H., Lu Y., Zhai C.  
Año: 2010

Titulo: Using text mining and sentiment analysis for online forums hotspot detection and forecast  
Autor: Li N., Wu D.D.  
Año: 2010

Titulo: Data mining emotion in social network communication: Gender differences in MySpace  
Autor: Thelwall M., Wilkinson D., Uppal S.  
Año: 2010

Titulo: Classifying sentiment in microblogs: Is brevity an advantage?  
Autor: Bermingham A., Smeaton A.  
Año: 2010

Titulo: A study of Information Retrieval weighting schemes for sentiment analysis  
Autor: Paltoglou G., Thelwall M.  
Año: 2010

Titulo: Semi-supervised recognition of sarcastic sentences in twitter and Amazon  
Autor: Davidov D., Tsur O., Rappoport A.  
Año: 2010

# Luke

Para el uso de Luke para la realización de consultas, basta con ejecutarlo y, a continuación, especificar la ruta del índice. Una vez hecho esto, accedemos a la pestaña Search y, una vez ahí, escribimos la expresión deseada además de especificar sobre qué campo queremos realizar la búsqueda y el analizador a utilizar.

En la siguiente imagen podemos observar como ejemplo la búsqueda de la palabra “twitter” en el campo Abstract utilizando el analizador EnglishAnalyzer:

The screenshot shows the Luke - Lucene Index Toolbox (7.1.0) interface. The 'Search' tab is active. The search expression is 'twitter'. The analyzer is set to 'org.apache.lucene.analysis.en.EnglishAnalyzer' and the default field is 'Abstract'. The search results are displayed in a table with columns: #, Score, Doc. Id, Abstract, Author keyw, Authors, Cited by, EID, Index keywor, Link, Source title, and Title. The results show 14 documents with scores ranging from 3,6988 to 3,3815. The index name is '/home/ppm/indice/index'.

#	Score	Doc. Id	Abstract	Author keyw	Authors	Cited by	EID	Index keywor	Link	Source title	Title
0	3,6988	1154	Sentiment	Microblogs	Giachanou	13	2-s2.0-8497	Data mining	https://www	ACM Compu	Like it
1	3,6449	112	Twitter mes	DAN2; Feat	Ghiassi M.,	101	2-s2.0-8487	DAN2; Feat	https://www	Expert Syst	Twitter
2	3,6002	1601	Microbloggi	Classifier; O	Shahheidar	9	2-s2.0-8488	Microbloggi	https://www	Proceeding	Twitter
3	3,5844	743	Social medi		Ranco G., A	22	2-s2.0-8494	finance; Int	https://www	PLoS ONE	The eff
4	3,4758	534	Twitter is a	Marijuana; f	Cavazos-Rel	31	2-s2.0-8490	cannabis; a	https://www	Journal of M	Charac
5	3,4529	138	Micro-blogs		Bifet A., Fra	85	2-s2.0-7865	Bloggng; D	https://www	Lecture Not	Sentim
6	3,4484	556	The dramat	Box office; F	Arias M., Ari	30	2-s2.0-8488	Box office; I	https://www	ACM Transa	Foreca
7	3,4324	1701	It was not u	Opinion min	Martínez-Cá	8	2-s2.0-8493	Artificial int	https://www	Journal of Ir	Polarit
8	3,4153	1045	To examine		Canvasser f	15	2-s2.0-8492	classificatio	https://www	Journal of E	The us
9	3,4134	2	Twitter is a		Turnasjan A	844	2-s2.0-8488	Content an	https://www	ICWSM 2010	Predict
10	3,4134	302	Twitter as a	Collective ir	Cheong M.,	52	2-s2.0-7404	Blogospher	https://www	Internation	Integra
11	3,4134	400	Twitter sent	Machine lea	Hassan A.,	41	2-s2.0-8488	Multi-class	https://www	Proceeding	Twitter
12	3,4101	965	Twitter is a		Bifet A., Hol	17	2-s2.0-8005	Data strear	https://www	Lecture Not	MOA-TV
13	3,3815	170	Microbloggi		Stieglitz S.,	76	2-s2.0-8485	Social netw	https://www	Proceeding	Politica
14	3,3815	220	The study o	Civilian rec	Cheong M.,	66	2-s2.0-8500	Data visuali	https://www	Information	A micro