CLASIFICACIÓN DE ARTÍCULOS INVESTIGATIVOS POR CATEGORÍAS SEGÚN EL CONTENIDO DE SUS ABSTRACT

A PREPRINT

Jairo Castrellón Torres
Estudiante maestría en Estadística
Universidad Nacional de Colombia
jcastrellont@unal.edu.co

6 de noviembre de 2019

1. Comprensión del negocio

Hoy en día, existen muchas plataformas que tienen bases de datos de artículos de investigación y facilitan su busqueda (Scopus, Google académico, Science Direct, etc), en estas plataformas se pueden hacer filtros para encontrar los articulos de interés. Estos filtros se pueden realizar por palabras clave, título del artículo, nombre del autor, entre otros.

Sin embargo, en la mayoría de ocaciones, cuando el investigador está interesado en un tema en específico, la busqueda por filtros de palabras clave no es suficiente y se tienen que leer muchos artículos (o por lo menos sus abstract) para encontrar cuales son aquellos que realmente sirven para la investigación, lo que puede resultar una tarea tediosa, y no siempre muy útil.

Dada esta situación, el proyecto que se pretende realizar haciendo uso de las herramientas de la minería de datos, consiste en, teniendo un conjunto de datos que consista de información fundamental (nombre del artículo, palabras clave, abstract, autor, etc) sobre una cantidad considerable de artículos de investigación acerca de un tema en particular como por ejemplo Inteligencia Artificial, se pueda realizar posteriormente una clasificación por los temas que trata cada uno de estos artículos haciendo un estudio de las frases más mencionadas en sus abstract y si efectivamente están dentro de sus palabras clave, de tal forma que se puedan hacer agrupaciones de aquellos que hablen por ejemplo de aprendizaje de máquina, minería de datos, etc, y de esta manera, facilitar más la búsqueda de los temas de interés del investigador.

1.1. Objetivos del negocio

1.1.1. Antecedentes

Las bases de datos actuales de artículos de investigación, cuentan con un sistema de búsqueda tradicional, en la que se puede filtrar los artículos deseados por palabras clave, nombre del autor, revista donde se públicó, etc. Sin embargo, los resultados de esta búsqueda se ordenan de acuerdo a la relevancia de los artículos, es decir, en primera instancia pueden aparecer aquellos más visitados o los cuales presenten más puntaje en alguno de los item de calificación.

Todo lo anterior resulta poco eficiente si el usuario (investigador) está en la necesidad de encontrar aquellos artículos de su interés, y aunque su búsqueda sea muy específica, tiene que leer, por lo menos, los abstract de todos esos artículos que parecieran coincidir con lo deseado.

Esta actividad de busqueda se ha convertido en algo usual, no solo para artículos de investigación, sino en información en general, esta actividad se realiza usualmente mediante motores de busqueda, que están basados en lo que se conoce como **Information retrieval** o recuperación de la información, lo cuál se podría definir como:.[Es] *la búsqueda de material (usualmente documentos) de una naturaleza sin estructura (usualmente textual) que satisface una información dentro de grandes colecciones"*(**Introduction to Informa- tion Retrieval.** P.R. Christopher)

1.1.2. Objetivos del negocio

Objetivo general: Clasificar textos investigativos por categorías, teniendo información básica sobre éstos como el título y el abstract, de tal manera que la temática del texto sea mejor interpretada mas allá de sus palabras claves.

Objetivos específicos:

- Organizar los datos textuales de tal manera que se pueda prescindir de aquella información no necesaria a la hora de clasificar un texto por su tema.
- Encontrar palabras claves de artículos investigativos, únicamente con la información que proporciona el abstract y el título de este.
- Clasificar los textos investigativos teniendo en cuenta ïdeas claves"que proporcione este.

1.1.3. Criterío de éxito

El éxito de este proyecto, estará en la clasificación optima de los articulos a estudiar. Para esto, se cuenta con un dataset que contiene información básica (título, autor, abstract) de 448 artículos de investigación sobre inteligencia artificial, estos artículos ya estan clasificados por sus lineas temáticas, teniendo en total 83 categorías de clasificación. Con base en esto, se pretende clasificar otros artículos sobre inteligencia artificial que no estan categorizados, utilizando la extracción de palabras clave e ideas clave que contengan en sus abstract.

1.2. Evaluación de la situación

En la actualidad, existen muchos motores de búsqueda que filtran los artículos investigativos por sus palabras claves y relevancia (Scopus, Google Academics, ScienceDirect, etc), y adicionalmente, hay multiples trabajos que consisten en la clasificación de textos para evitar la tediosa tarea de leer cada artículo para encontrar el de interés, sin embargo, ninguna de las plataformas de búsqueda hace uso de alguno de estos métodos. Lo ideal es encontrar un método que realice esta clasificación de manera más optima, y que despues se pueda llevar a cabo en otros temas diferentes a los que se refieren a la inteligencia artificial.

1.2.1. Inventario de requerimientos de Recursos, Hipótesis y Limitaciones

Inicialmente se requiere un conjunto de datos que contenga información relevante sobre determinado número de artículos investigativos, de los cuales, algunos de ellos puedan estar ya clasificados y que sirvan como conjunto de datatest, posteriormente, se requieren herramientas básicas de análisis textual, de tal manera que con la información que puedan arrojar esto conjuntos de datos se pueda obtener resultados interesantes. Por último, se requiere un conocimiento de minería de datos y aprendizaje de máquina, de tal manera que se pueda hacer la mejor clasificación sobre los artículos.

Las hiótesis a tener en cuenta es que este trabajo estará restringido al estudio de datos sobre artículos que tratan únicamente temas de inteligencia artificial, lo que sugiere una solución parcial a un problema que se da en textos de todas las temáticas.

Lo anterior conlleva a una limitación, debido a que lo ideal seria tener información de todo tipos de textos investigativos, sin embargo, esto requiere máquinas más robustas que puedan dar tratamiento a la gran cantidad de datos con las que se podría contar.

1.2.2. Riesgos y contingencias

Se puede tener el riesgo de que los datos que se tienen sobre los artículos de investigación no sean suficientes para poder categorizarlos de una mejor manera, y que haga falta tener disponible el artículo completo, pero esto podría llevar a una dificultad en el tiempo y la dificultad de conseguir grandes cantidades de artículos completos, es por esto que se cuenta con un conjunto de datos ya clasificados, para así buscar acercarse lo más posible a lo deseado.

1.2.3. Costos y beneficios

Costos: Debido a la cantidad de información que se tiene de documentos sin clasificar, esto podría llevar a un alto costo en cuanto al tiempo de clasificación de cada uno de los textos. También, se requiere garantizar la óptima clasificación de los artículos con los datos disponibles, lo que requiere un estudio más específico de los artículos con los que se trabajaron, esto implica, leer el artículo y a criterio de un experto, garantizar si este artículo esta bien clasificado, o si por el contrario, cala mejor en otra clase o categoría.

Beneficios: Se contará con una herramienta que pueda extenderse a todos los campos del conocimiento y que pueda ayudar a la clasificación más rigurosa de textos investigativos, con el fin de que se reduzca el proceso de revisión del estado del arte en temas puntuales.

1.3. Determinar el objetivo de Minería de Datos

Desde el concepto como tal, lo que busca la minería de datos es completar el ciclo de extracción de datos, manipulación de los mismos, de tal manera que sea más comprensible, para cualquier persona, la información con la que se cuenta, encontrar patrones en los datos con los que se dispone y por último realizar clasificaciones o predicciones en futuros datos, todo esto es lo que se pretende realizar en el proyecto que se quiere llevar a cabo.

1.3.1. Objetivos de Minería de Datos

Extraer, transformar y identificar patrones obtenidos de el conjunto de datos con los que se dispone, en particular, se quiere manipular un conjunto de datos textuales de tal manera que se pueda encontrar una idea principal en lo dicho allí y clasificarlo en una categoría en particular.

1.3.2. Criterio de Éxito de Minería de Datos

El criterio de éxito será medido principalmente en si, usando los métodos de minería de datos, se pueden extraer las palabres claves que realmente identifiquen al texto, y consecuentemente, encontrar ideas principales que permitan clasificar los artículos de manera conveniente.

1.4. Desarrollar el Plan de Proyecto

El plan de proyecto, basicamente tendrá como hilo conductor las herramientas que se aprendan en la asignatura de minería de datos, pero para ser mas generales, será llevado a cabo teniendo en cuenta los objetivos del proyecto, en una primera etapa se pretendera explorar y procesar los datos, en una segunda etapa se querrá hacer asociaciones y por último predicciones mediante clasificación.

1.4.1. Plan de proyecto

El plan de proyecto se llevará a cabo teniendo en cuenta las diferentes etapas de la minería de datos, sin embargo, en terminos generales, se pretenderá, obtener los datos adecuados, hacer la limpiezza eficiente de estos, obtener palabras claves según sus abstract, hacer asociaciones con aquellos artículos que ya están clasificados y por último clasificar los textos.

2. Comprensión de Datos

La idea principal en la búsqueda de los datos es encontrar aquellos que brinden la información necesaria y suficiente con la que se pueda categorizar un artículo, como se mencionó anteriormente, lo deseable es contar con un conjunto de datos que contenga los artículos completos para hacer una clasificación más exitosa, sin embargo, la obtención de este tipo de datos tiene grandes dificultades, tanto para la obtención como para la manipulación, es por esto que se cuenta unicmente con la información más relevante de los artículos investigativos.

2.1. Obtener los datos iniciales

Los datos iniciales se obtuvieron vía internet, y son datos disponibles al público, por lo que su extracción no infirió en ningún inconveniente o en el requerimiento de obtener algún permiso especial para su obtención.

2.1.1. Reporte de la obtención de los datos iniciales

El conjunto de datos con el que se va a trabajar proviene de dos fuentes distintas. La primera de ellas consiste de un conjunto de artículos que fueron expuestos en un conjunto de conferencias acerca de Machine Learning, y que por lo tanto, sus temáticas fueron clasificadas, artículo por artículos, este dataset fue extraído del *IEEE Xplore Digital Library* y de la página web oficial del *International Conference on Machine Learning Applications (ICMLA)*, se extrayeron estos datos de cuatro ediciones diferentes (del 2014 al 2017). Este conjunto de datos contiene 448 artículos presentados en las conferencias y contiene 6 atributos incluyendo información temática de la sesión en que se presentó el artículo. Este dataset está disponible en `https://data.mendeley.com/datasets/wj5vb6h9jy/2` en formato CSV.

El segundo Dataset contiene información relacionada a 41000 artículos de investigación en Machine Learning, Inteligencia Artificial, entre otros, publicados entre 1992 y 2018. Este dataset contiene 9 atributos de los artículos. Este dataset está disponible en `https://www.kaggle.com/neelshah18/arxivdataset` en formato json.

2.2. Descripción de los datos

El primer conjunto de datos, tomado de la ICMLA se encuentra en formato CSV, este contiene 448 objetos de los cuales a cada uno se le obtuvo 6 atributos tales como abstract, palabras clave, codigo del paper, sesión (categoría en la cual se clasifica el artículo), título y año:

	abstract	keywords	paper_id	session	title	year
0	Statistical word alignment models need large a...	statistical word alignment, ensemble learning,...	1	Ensemble Methods	Ensemble Statistical and Heuristic Models for ...	2014
1	Spectral learning algorithms learn an unknown ...	representation, spectral learning, discrete fo...	2	Ensemble Methods	Improving Spectral Learning by Using Multiple ...	2014
2	Number of defects remaining in a system provid...	software defect prediction, particle swarm opt...	3	Ensemble Methods	Applying Swarm Ensemble Clustering Technique f...	2014
3	Not all instances in a data set are equally be...	filtering, label noise, instance weighting	4	Ensemble Methods	Reducing the Effects of Detrimental Instances	2014
4	Learning in non-stationary environments is not...	twitter, adaptation models, time-frequency ana...	5	Ensemble Methods	Concept Drift Awareness in Twitter Streams	2014

Figura 1: Tabla con los datos recolectados en el primer dataset

El segundo conjunto de tados se encuentra en formato JSON y contiene 41000 objetos, cada uno con 9 atributos que son, autor, dia, mes, año de publicación, abstract, título, código del artículo y tag. De este podemos observar que los datos se presentan con algunos inconvenientes en su formato y visualización, por lo que es necesario algunas transformaciones para su mejor entendimiento.

	author	day	paper_id	link	month	abstract	tag	title	year
40995	[{'name': 'Vitaly Feldman'}, {'name': 'Pravesh...	18	1404.4702v2	[{'rel': 'alternate', 'href': 'http://arxiv.or...	4	We study the complexity of learning and approx...	[{'term': 'cs.LG', 'scheme': 'http://arxiv.org...	Nearly Tight Bounds on ℓ_1 Approximation ...	2014
40996	[{'name': 'Orly Avner'}, {'name': 'Shie Mannor'}]	22	1404.5421v1	[{'rel': 'alternate', 'href': 'http://arxiv.or...	4	We consider the problem of multiple users targ...	[{'term': 'cs.LG', 'scheme': 'http://arxiv.org...	Concurrent bandits and cognitive radio networks	2014
40997	[{'name': 'Ran Zhao'}, {'name': 'Deanna Needel...	22	1404.5899v1	[{'rel': 'alternate', 'href': 'http://arxiv.or...	4	In this paper, we compare and analyze clusteri...	[{'term': 'math.NA', 'scheme': 'http://arxiv.o...	A Comparison of Clustering and Missing Data Me...	2014
40998	[{'name': 'Zongyan Huang'}, {'name': 'Matthew ...	25	1404.6369v1	[{'rel': 'related', 'href': 'http://dx.doi.org...	4	Cylindrical algebraic decomposition(CAD) is a ...	[{'term': 'cs.SC', 'scheme': 'http://arxiv.org...	Applying machine learning to the problem of ch...	2014
40999	[{'name': 'Imen Trabelsi'}, {'name': 'Dorra Be...	27	1407.0380v1	[{'rel': 'alternate', 'href': 'http://arxiv.or...	6	Several speaker identification systems are giv...	[{'term': 'cs.SD', 'scheme': 'http://arxiv.org...	A Multi Level Data Fusion Approach for Speaker...	2014

Figura 2: Tabla con los datos recolectados en el segundo dataset

2.3. Exploración de los datos

En ambos conjuntos de datos se realizó una exploración detallada de tal manera que pudieramos observar la cantidad de datos que tenemos y el tipo de datos con el que se trabaja, inicialmente, al realizar esto, obtuvimos lo siguiente:

```
df2.shape
```

```
(448, 6)
```

```
## porcentaje de datos faltantes por columna en df2
df2.isnull().mean().sort_values(ascending=False)
```

```
year        0.0
title       0.0
session     0.0
paper_id    0.0
keywords    0.0
abstract    0.0
dtype: float64
```

```
## Tipos de datos en cada columna de df
df2.dtypes
```

```
abstract    object
keywords    object
paper_id     int64
session     object
title       object
year         int64
dtype: object
```

Figura 3: Descripción de los datos obtenidos

Donde podemos observar la cantidad de los datos en el primer dataset, tenemos una matríz de 448 filas y 6 columnas, no tenemos datos perdidos, es decir, para todos los artículos tenemos información guardada, y podemos observar el tipo de dato que tenemos. Adicionalmente, en este dataset, pudimos obtener el número de clases y frecuencia de cada una de ellas, donde se obtuvieron 83 clases, y a continuación algunas de sus frecuencias:

```
# listar los objetos en Session de df2
clase = pd.DataFrame(df2['session'].value_counts())
clase.head(5)
```

	session
Machine Learning in Energy Applications	20
Machine Learning for Predictive Models in Engineering Applications	19
Machine Learning Algorithms Systems and Applications	16
Machine Learning in Information and System Security Issues	15
Machine Learning Algorithms, Systems and Applications Workshop	11

Figura 4: Clases del conjunto de datos clasificado

donde observamos que la clase que mas se repite lo hace en 20 ocasiones.

2.4. Calidad de los datos

En ambos dataset, se observo que no se tienen datos perdidos, sin embargo, como se visualizó en el segundo conjunto de datos, hay atributos que no son claramente comprensibles como el que se etiqueta como Tag, o que su formato dificulta su comprensión, además que no son necesarios para el estudio que se pretende hacer. En el primer conjunto de datos, se eliminaron las dimensiones de "paper_idz zear", mientras que en el segundo conjunto se eliminaron las columnas correspondientes a .ªuthor", "day", "month", "paper_id", "tagz zear", puesto que no son relevantes para nuestro estudio, en todo lo demás, podemos observar que contamos con un buen conjunto de datos.

	paper_id	abstract	title
0	1802.00209v1	We propose an architecture for VQA which utili...	Dual Recurrent Attention Units for Visual Ques...
1	1603.03827v1	Recent approaches based on artificial neural n...	Sequential Short-Text Classification with Recu...
2	1606.00776v2	We introduce the multiresolution recurrent neu...	Multiresolution Recurrent Neural Networks: An ...
3	1705.08142v2	Multi-task learning is motivated by the observ...	Learning what to share between loosely related...
4	1709.02349v2	We present MILABOT: a deep reinforcement learn...	A Deep Reinforcement Learning Chatbot

Figura 5: tabla de los datos sin clasificar eliminando atributos que no serán relevantes en nuestro estudio

3. Preparación de los datos

Como se puede observar en nuestros datos, hay atributos con los que no se quieren trabajar como el año de publicación, el código del artículo, el autor, etc. Por el contrario, hay otros atributos que son más de nuestro interés como el título, abstract, palabras claves y la clase a la cuál pertenencen los artículos, por esta razón, se realiza un proceso riguroso en el que se prescinde de aquellas dimensiones en las que no estamos interesados y se transforman aquellas en que sí, de tal manera que se pueda obtener más información de los datos mismos.

3.1. Selección de Datos

Como nuestra idea es clasificar los artículos por su temática, hay datos que no nos aportan mucha información al respecto, aunque, si bien, el dato del autor(es) de el artículo es un dato relevante a la hora de filtrar la información, esta no nos dice nada acerca de la temática del artículo, a menos que se tenga más información del autor, pero esto implicaría buscar mas datos y sería muy costoso a la hora de estructurar y procesar los datos.

Otras dimensiones poco importantes para nuestro estudio son aquellas que se refieren a la fecha de publicación del artículo, debido a que esta no nos brinda ninguna información acerca del tema del artículo, esto ocurre en el mismo sentido con el código del artículo.

Por otro lado, el abstract, título y palabras clave, son fundamentales a la hora de clasificar nuestros artículos, porque estos nos pueden dar más palabras e ideas clave de lo que se refiere el artículo.

3.2. Limpiar datos

Aunque en nuestro caso no tenemos datos perdidos ni datos que no estén en el formato de los demás, si debemos estructurarlos de tal manera que se puedan presentar y trabajar con mayor comodidad, en particular, ya que nuestra fuente de estudio en un mayor porcentaje será el trabajo que se realice al abtract, es necesario dejar este de forma que sea más sencillo obtener información relevante.

Por ejemplo, no nos será de gran ayuda los signos de puntuación a la hora de presentar una palabra o idea clave, de igual manera, hay palabras que ocurren en mayor proporción y que no nos aportan mucha información (Stop_words), un ejemplo de estas son “the”, “a”, “an”, “is”, etc, por lo cual debemos removerlas para ver la información realmente importante.

```
0    Statistical word alignment models need large a...
1    Spectral learning algorithms learn an unknown ...
2    Number of defects remaining in a system provid...
3    Not all instances in a data set are equally be...
4    Learning in nonstationary environments is not ...
```

Una vez realizado esto, se hace un proceso de *tokenzación* que se utiliza para dividir cadenas de texto más largas de tal manera que se conviertan en cadenas más pequeñas o tokens. Así, los segmentos de texto de mayor longitud pueden transformarse en oraciones y las oraciones pueden ser tokenizadas en palabras. De esta forma, después del proceso de Tokenización se obtuvo:

```
0    [statistical, word, alignment, models, need, l...
1    [spectral, learning, algorithms, learn, unknow...
2    [number, defects, remaining, system, provides,...
```

```
3    [instances, data, set, equally, beneficial, in...
4    [learning, nonstationary, environments, easy, ...
5    [work, presents, new, method, classifying, pre...
6    [captchas, challengeresponse, tests, widely, u...
7    [reinforcement, learning, techniques, become, ...
8    [signatures, single, widely, used, method, ide...
9    [automatic, detection, abnormal, events, one, ...
```

Nótese que en esencia, son los mismos objetos, pero esta vez separados en sus elementos atómicos que son las palabras. Una vez realizado esto, se realiza el proceso de *Lematizar y stemizar*. Estos procesos consisten en dejar en la manera más básica cada elemento atómico, de tal manera que palabras como "learning" y "learn" puedan ser interpretadas como del mismo núcleo y ver que tan frecuentes son estos núcleos. En el proceso de lematización se obtuvo lo siguiente:

```
0    [statistical, word, alignment, model, need, la...
1    [spectral, learning, algorithm, learn, unknown...
2    [number, defect, remaining, system, provides, ...
3    [instance, data, set, equally, beneficial, ind...
4    [learning, nonstationary, environment, easy, t...
5    [work, present, new, method, classifying, prev...
6    [captchas, challengeresponse, test, widely, us...
7    [reinforcement, learning, technique, become, p...
8    [signature, single, widely, used, method, iden...
9    [automatic, detection, abnormal, event, one, c...
```

Nótese que a diferencia del proceso anterior, en este las palabras con plurales se contraen trabajando con su núcleo en singular, y ahora, durante el proceso de stemming se obtiene:

```
0    [statistical, word, alignment, model, need, la...
1    [spectral, learning, algorithm, learn, unknown...
2    [number, defect, remaining, system, provides, ...
3    [instance, data, set, equally, beneficial, ind...
4    [learning, nonstationary, environment, easy, t...
5    [work, present, new, method, classifying, prev...
6    [captchas, challengeresponse, test, widely, us...
7    [reinforcement, learning, technique, become, p...
8    [signature, single, widely, used, method, iden...
9    [automatic, detection, abnormal, event, one, c...
```

De la misma manera se puede proceder a organizar los datos de tal manera que sean más comprensibles, por ejemplo, en nuestro conjunto de datos clasificados, podemos asignar un número a cada clase, de tal manera que podamos identificarlo con mayor facilidad y podamos hacer algunos análisis más allá de la categoría misma.

3.2.1. Construir datos

Como se mencionó anteriormente, una fuente importante para poder obtener nuestras ideas claves es el abstract, y con ella sus palabras claves, en nuestro conjunto de datos que es objeto de estudio, podemos ver que no tenemos algunos atributos que en el otro conjunto si tenemos, por ejemplo, las palabras claves y la clase.

Las palabras claves deben ser un suministro fuerte a la hora de ayudarnos a la clasificación de los textos, es por eso que para el dataset de estudio se deben construir estas palabras, por lo que se trendrá en cuenta todas aquellas palabras que son más frecuentes (sin incluir las Stop_words mencionadas antes) y aquellas que son menos frecuentes, de esta manera podemos darnos una idea de las palabras importantes en los textos y que nos puedan ayudar a identificar el tema central de este.

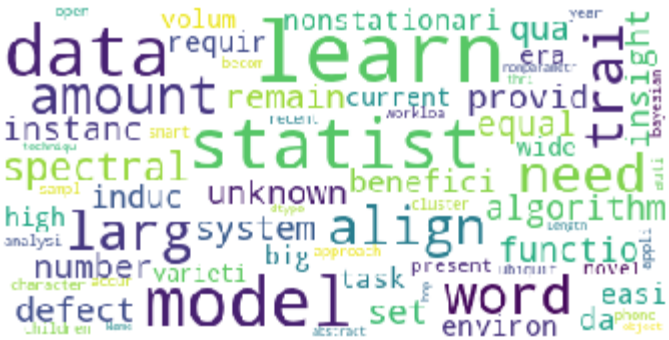

Figura 6: Nube de palabras más frecuentes

Además de lo anterior, se puede observar, que al realizar los procesos descritos en 3.2, las palabras que se encontraron más frecuentes fueron:

```
use          573
data         462
learn        425
algorithm    350
model        347
propos       333
method       318
result       258
paper        256
```

Tal como se mostraba en la imagen anterior. En total, se pudieron encontrar 3944 palabras diferentes en todos los abstract analizados.

3.2.2. Integrar Datos

Una vez hecho el proceso anterior, se deben integrar los datos nuevamente dentro de nuestra matríz en su nuevo formato, de tal manera que sea más sencilla su interpretación:

paper_id	title	year	abstract	keywords	classification
1	Ensemble Statistical and Heuristic Models for ...	2014	Statistical word alignment models need large a...	statistical word alignment, ensemble learning,...	10
2	Improving Spectral Learning by Using Multiple ...	2014	Spectral learning algorithms learn an unknown ...	representation, spectral learning, discrete fo...	10
3	Applying Swarm Ensemble Clustering Technique f...	2014	Number of defects remaining in a system provid...	software defect prediction, particle swarm opt...	10
4	Reducing the Effects of Detrimental Instances	2014	Not all instances in a data set are equally be...	filtering, label noise, instance weighting	10
5	Concept Drift Awareness in Twitter Streams	2014	Learning in non-stationary environments is not...	twitter, adaptation models, time-frequency ana...	10

Figura 7: tabla con los datos arreglados

De esta manera se le da el formato deseado a los datos, ara así poder iniciar con los análisis correspondientes. Otra manera de integrar los datos, por lo menos los de los abstract, es por medio de una matríz que nos de información por artículo, si contiene o no una palabra en específico como se muestra a continuación:

women	word	wordlist	work	worker	workfre	workload	world	worldwid	wors	worst	worth	worthwhil	worthwhilein	would	wrap	wrapper	w
False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	
False	False	False	False	False	False	False	False	False	True	False	False	False	False	False	False	False	
False	False	False	True	False	False	False	True	False	False	False	False	False	False	False	False	False	
False	False	False	True	False	False	True	False	False	False	False	False	False	False	False	False	False	
False	False	False	True	False	False	False	False	False	False	False	False	False	False	False	False	False	
False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	
False	False	False	True	False	False	False	False	False	False	False	False	False	False	False	False	False	

Figura 8: Vectorización de los abstract

4. Asociación

Se aplicó asociación a nuestro conjunto de datos. Para esto, se usaron los algorítmos apriori y FPgrowth para comparar los resultados, de tal manera que se pueda especificar cuál de estos nos dá mejores resultados.

4.1. Algoritmo apriori

Para el algorítmo apriori, se modificaron los datos de tal manera que se obtuviera una matríz con cada uno de los núcleos de las palabras obtenidas en el proceso de lemmatization y stemming. La matríz obtenida al final fue la siguiente:

	0	1	2	3	4	5	6	7	8	9	10	11
session												
8	statist	word	align	model	need	larg	amount	train	data	weak	smallsiz	corpora
8	spectral	learn	algorithm	learn	unknown	function	learn	spectral	eg	fourier	represent	function
8	number	defect	remain	system	provid	insight	qualiti	system	defect	detect	system	predict
8	instanc	data	set	equal	benefici	induc	model	data	instanc	outlier	nois	detriment
8	learn	nonstationari	environ	easi	task	requir	distinct	approach	learn	model	must	abil

5 rows × 139 columns

Figura 9: Matríz con las palabras de los abstract

Nótese que cada fila representa el abstract de un artículo, y cada casilla tiene cada palabra del abstract. En este punto, fue necesario eliminar filas (abstracts) de la tabla debido a la longitud de su abstract, es decir, al número de columnas tan extenso que éstas producian, todo esto para no generar ruido a la hora de realizar la asociación.

Al hacer uso del algorítmo de asociacón apriori, con un soporte mínimo de 4 % y una confianza del 70 % se obtuvieron las siguientes reglas:

```
Rule: machin -> learn
Support: 0.0968586387434555
Confidence: 0.9024390243902439
Lift: 4.996111700247438
====================================
Rule: neural -> network
Support: 0.06282722513089005
Confidence: 1.0
Lift: 6.821428571428572
====================================
```

Lo que nos ratifica lo que se veía anteriormente cuando se explicaba la frecuencia de las palabrs y la temática de los artículos escogidos. Es decir, esto nos confirma que la temática principal en la mayoría de artículos es acerca del aprendizaje de máquina (machine learning) pues alcanzando el soporte mínimo, cada vez que apareció la palabra

machin tambien apareció learn, de igual manera con las palabras neural y network, haciendo referencia a la relevancia de el tema de redes neuronales en los artículos trabajados.

Ahora, modificando el soporte, llevandolo al 2 % y aumentando la confianza al 90 % se obtienen diferentes reglas, aparte de las que ya obtuvimos anteriormente, como las siguientes:

```
Rule: method -> machin
Support: 0.020942408376963352
Confidence: 1.0
Lift: 5.536231884057971
====================================
Rule: neural -> paper
Support: 0.02356020942408377
Confidence: 1.0
Lift: 6.821428571428572
====================================
Rule: use -> neural
Support: 0.02617801047120419
Confidence: 1.0
Lift: 6.821428571428572
====================================
```

explicando la importancia de las palabras aquí mencionadas. Por último, aumentando el soporte mínimo al % y disminuyendo la confianza al % se obtiene:

```
Rule: machin -> learn
Support: 0.0968586387434555
Confidence: 0.5362318840579711
Lift: 4.996111700247438
====================================
```

Lo que confirma de nuevo la relevancia de los temas relacionados a machine learning en los artículos seleccionados.

4.2. Algorítmo FPgrowth

En este caso se hace uso de la tabla con los datos como se muestran en la Figura 8, se establece un soporte mínimo del 30 % y una confianza del 90 % y se obtiene lo siguiente:

```
 support itemsets
0 0.746073 (use)
1 0.617801 (paper)
2 0.536649 (result)
3 0.526178 (data)
4 0.505236 (propos)
5 0.494764 (learn)
6 0.421466 (model)
7 0.421466 (algorithm)
8 0.408377 (method)
9 0.397906 (base)
10 0.358639 (show)
11 0.340314 (approach)
12 0.400524 (perform)
13 0.471204 (use, paper)
14 0.416230 (use, result)
15 0.332461 (result, paper)
16 0.410995 (machin, learn)
17 0.319372 (paper, data)
18 0.369110 (use, propos)
19 0.340314 (propos, paper)
20 0.413613 (use, learn)
```

```
21 0.316754 (use, model)
22 0.321990 (use, algorithm)
23 0.311518 (method, use)
24 0.316754 (use, base)
25 0.303665 (use, perform)
```

Se puede observar que FPgrowth, aunque saca más reglas, éstas no tinene la misma presición que las obtenidas con el algorítmo apriori. Nuevamente aparece la regla machin learn, lo que nuevamente refleja la importancia del tema en estos artículos.

5. Agrupamiento

En nuestro conjunto de datos, se intentó realizar clustering de la manera tradicional haciendo uso de los algorítmos KMeans, Hierarchical clusterer y Densidad basada en clusters, sin embargo, al ser datos textuales su interpretación no fue muy sencilla, ni la agupación muy eficaz.

Para todos los casos se utilizaron los datos como se muestra en la Figura 8.

5.1. Simple KMeans

Se utilizó Weka para hacer KMeans, en este caso, se eliminaron algunas palabras que no tenían mucha trascendencia en los artículos, es decir, aquellas que no se repetían más de 20 veces en todos los artículos. Para el conjunto de datos clasificado, se sabe que hay 86 clases, sin embargo, tomamos 9 clusters obteniendo la siguiente información:

```
kMeans
======

Number of iterations: 11
Within cluster sum of squared errors: 12940.0


=== Model and evaluation on training set ===

Clustered Instances

0       94 ( 25%)
1       42 ( 11%)
2       72 ( 19%)
3      128 ( 34%)
4        3 (  1%)
5       38 ( 10%)
6        2 (  1%)
7        1 (  0%)
8        2 (  1%)
```

Sin embargo, como se mencionó anteriormente, la interpretación de estos clusters es compleja, pero lo que se puede decir es que si se pueden establecer agrupaciones entre artículos dadas las palabras mas relevantes que estos tengan.

5.2. Hierarchical clusterer

En este caso, haciendo uso de este algorítmo de agrupación, se puede ver que no se hace una verdadera distinción entre los clusters, es decir, que no se agrupa de manera eficiente, con los mismos datos usados en KMeans, se obtuvo lo siguiente:

```
=== Model and evaluation on training set ===

Clustered Instances

0      374 ( 98%)
```

```
1         1 (  0%)
2         1 (  0%)
3         1 (  0%)
4         1 (  0%)
5         1 (  0%)
6         1 (  0%)
7         1 (  0%)
8         1 (  0%)
```

Que no muestra nada interesante.

5.3. Densidad basada en clusters

De igual manera, se utilizaron los datos como en antes, pero en esta ocación, el algorítmo utilizado, definió dos clusters como los mas importantes, como se muestra a continuación:

```
=== Model and evaluation on training set ===

Clustered Instances

0      264 ( 69%)
1      118 ( 31%)


Log likelihood: -107.50134
```