Universidad Rey Juan Carlos

Análisis de género DBLP usando ElasticSearch y Kibana

Abril de 2019

Autores:

Marina Kurmanova

Rafael Durán Bautista

Elizabeth Jaimes Sequeda

Contenido

VISIÓN GENERAL

En este documento se describen los detalles del desarrollo de la segunda práctica para la asignatura de Recuperación de información y minería de texto usando Elasticsearch y Kibana.

El trabajo consiste en utilizar la fuente de datos DBLP (Digital Bibliography & Library Project, en inglés), en formato xml y realizar las transformaciones necesarias en el archivo original para hacerlo manejable y usable con Elasticsearch y Kibana, donde realizaremos una serie análisis de datos tendientes a realizar un estudio de género, para lo cual enriquecemos la información de la fuente cruzando los datos de los autores con el API genderize.io.

El sitio web de DBLP[1] posee un enorme repositorio bibliográfico de artículos relacionados con ciencias de la computación y se encuentra alojado en la Universidad de Trier, Alemania.

Esta base de datos está especializada en ciencias informáticas y recopila información desde 1980. Incluye más de 2,3 millones de artículos y simposios de conferencias de las publicaciones más prestigiosas. Además de sus citas, el buscador permite navegar fuera de línea para encontrar información desde diferentes puntos de vista: autores/editores, revistas, conferencias y monografías.

El objetivo principal de este estudio, es establecer el índice de participación de la mujer a lo largo de la historia, en el ámbito de las publicaciones científicas en diversos campos académicos, tomando como base los artículos publicados en el repositorio DBLP.

Con este análisis esperamos determinar la evolución y situación actual de la participación de la mujer en la comunidad científica, para identificar puntos de mejora y despertar el interés de la sociedad en la creación de estrategias de inclusión para reducir el sesgo de participación en estos aspectos.

Para cumplir con este objetivo, haremos uso del API genderize.io, para identificar realizar una revisión del acierto de la herramienta prediciendo el género a partir de un nombre dado. Para ello realizaremos un muestreo aleatorio con validación manual y se analizarán los resultados.

[1] https://dblp.uni-trier.de/xml/

Importancia del análisis de género

En la actualidad, los análisis y mediciones de fenómenos sociales relacionados con el género revisten especial interés para ayudar a determinar áreas de sesgo de inclusión de la mujer en la sociedad. Este tipo de estudios, permiten identificar puntos débiles de participación y pueden ser utilizados como herramienta para intentar equilibrar la balanza, fomentando la igualdad de participación de la mujer en diferentes aspectos del desarrollo global.

Para este tipo de estudios de género en distintas áreas, se hace uso de herramientas y/o algoritmos que pueden determinar con cierta precisión el género de una persona dado un nombre. Estas herramientas dan la posibilidad de enriquecer una fuente de datos, permitiendo inferir el género cuando se carece de dicha información.

Algunos ejemplos de este tipo de análisis incluyen perfiles de redes sociales, colaboradores de GitHub y autores de publicaciones científicas, cuyo análisis sobre género ha conducido a una mejor comprensión de la situación de las mujeres en dominios como la tecnología (Vasilescu, Serebrenik & Filkov, 2015), medios de comunicación (Matias, Szalavitz & Zuckerman, 2017; Macharia et al., 2015), y publicaciones académicas (Larivière et al., 2013a; West et al., 2013; Mihaljević-Brandt, Santamaría y Tullney, 2016; Bonham y Stefan, 2017b).

Teniendo en cuenta algunos estudios preliminares que evalúan diferentes herramientas de generificación, como *"Comparison and benchmark of name-togender inference services"*[2], hemos seleccionado Genderize.io para realizar la clasificación de género de nuestra fuente de datos DBLP, por su amplia difusión, y teniendo en cuenta que según el estudio de la referencia indica que esta herramienta logra una clasificación con una tasa de error inferior al 2%.

Genderize.io

Genderize.io utiliza grandes conjuntos de datos de información, desde perfiles de usuarios en las principales redes sociales y expone estos datos a través de su API. Es utilizada con frecuencia en el análisis, segmentación de anuncios, segmentación de usuarios, etc.

[2] Santamaría and Mihaljević (2018), Comparison and benchmark of name-to-gender inference services. PeerJ Comput. Sci. 4:e156; DOI 10.7717/peerj-cs.156. disponible en: https://peerj.com/articles/cs-156.pdf

En esta herramienta, la pieza clave para predecir el género es un string con el primer nombre, que se envía mediante GET. La respuesta del API es male (masculino), female (femenino) o ninguna (null), junto con dos Parámetros de confianza adicionales: probabilidad y conteo, que representan el número de entradas utilizadas para calcular la respuesta y la proporción de nombres con el género devuelto en la respuesta.

Los datos subyacentes se recopilan de las redes sociales en 79 países y 89 idiomas, en este momento, la base de datos contiene 216286 nombres distintos. Aunque el servicio no geo-localiza los nombres automáticamente, sí acepta dos parámetros opcionales, location_id y language_id, para obtener información más detallada.

Al utilizar filtros de localización para recuperar una asignación de género, la respuesta se basa únicamente en datos de un determinado país o idioma, lo cual mejora la precisión y acierto, ya que los nombres pueden depender mucho de la demografía.

La API es gratuita, pero limitada a 1000 nombres / día, teniendo en cuenta el tamaño del conjunto de datos DBLP, hacemos uso de las opciones disponibles en store.genderize.io. para hacer más de 120.000 peticiones al API, usando un script de python y almacenaremos las respuestas en un fichero Json para ser cruzada con todos los nombres de los autores con artículos publicados en DBLP.

OBJETIVOS

Utilizar Elasticsearch y Kibana para realizar el siguiente análisis:.

1. Analizar el acierto de la herramienta genderize.io para determinar el género de una persona dado un nombre.
2. Realizar un análisis de género en los artículos publicados en repositorio bibliográfico relacionados con ciencias de la computación dblp .

CARACTERÍSTICAS DE LA FUENTE DE DATOS

La recopilación completa de datos está actualizada al 26 de febrero de 2018 y es de libre acceso, se puede descargar como un único fichero XML desde http://dblp.uni-trier.de/xml/. Este fichero comprimido (gzipped) tiene un tamaño de 463 MB y expandido ocupa más de 2 GB.

La estructura de cada documento registrado en la base de datos contiene una pequeña descripción de los principales atributos de cada publicación.

En esta fuente de datos encontramos varios tipos diferentes de elementos como: article, inproceedings, proceedings, book, incollection, phdthesis, mastersthesis y www. En el documento DBLP - Some Lessons Learned[3], se describen los detalles de formato, en particular, la información relativa a los distintos campos para cada tipo de elemento.

Para el desarrollo de esta práctica solo se considerarán los documentos de tipo : Artículos de revista (article). Para cada uno de ellos, se indican al menos, los autores, el título, las páginas y la fecha, junto a otra información complementaria. A continuación, presentamos un ejemplo de la estructura de datos de cada documento en el xml:

```
<article mdate="2017-05-28" key="journals/acta/Saxena96">
    <author>Sanjeev Saxena</author>
    <title>Parallel Integer Sorting and Simulation Amongst CRCW Models.</title>
    <pages>607-619</pages>
    <year>1996</year>
    <volume>33</volume>
    <journal>Acta Inf.</journal>
    <number>7</number>
    <url>db/journals/acta/acta33.html#Saxena96</url>
    <ee>https://doi.org/10.1007/BF03036466</ee>
</article>
```

La línea del encabezado del documento xml, especifica la codificación ISO-8859-1 ("Latin-1")[4], pero en realidad el archivo solo contiene caracteres <128. Todos los caracteres que no son ASCII están representados por entidades simbólicas o numéricas que se declaran en el DTD.

Ejemplo de entidades simbólicas que pueden estar presentes en el documento[5]:

[3] https://dblp.uni-trier.de/xml/docu/dblpxml.pdf
[4] https://es.wikipedia.org/wiki/Ayuda:Caracteres_especiales
[5] Tabla tomada de wikipedia: https://es.wikipedia.org/wiki/Ayuda:Caracteres_especiales

Literal	Hex	Dec	Entidad	Carácter
	00A0	0160		espacio que no produce saltos de línea
¡	00A1	0161	¡	exclamación de apertura
¢	00A2	0162	¢	signo de centavo
£	00A3	0163	£	signo de libra
¤	00A4	0164	¤	signo internacional de moneda
¥	00A5	0165	¥	signo de yen
§	00A7	0167	§	signo de sección
¨	00A8	0168	¨	diéresis
©	00A9	0169	©	signo de copyright
ª	00AA	0170	ª	indicador ordinal femenino
«	00AB	0171	«	comillas anguladas de apertura
¬	00AC	0172	¬	signo de negación lógica
®	00AE	0174	®	signo de marca registrada

En total el documento xml completo ocupa más de 61 millones de líneas:

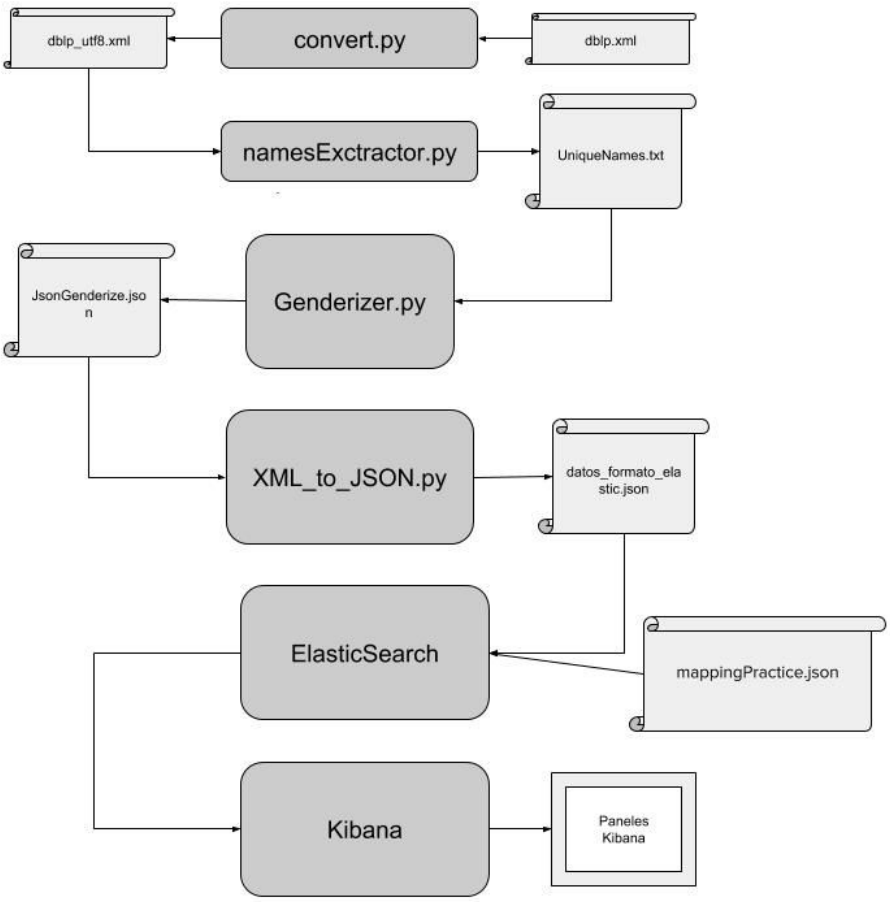
```
cat dblp.xml | wc -l
61765439
```

En esta práctica encontramos varios obstáculos para analizar el archivo xml por su tamaño, lo cual requiere una dedicación mayor de tiempo en tareas de limpieza y transformación de los datos no sólo a nivel de formato adecuado para los objetivos de la práctica, sino a nivel de codificación de los datos e incluso para la manipulación del fichero por su tamaño.

DISEÑO GENERAL DE LA SOLUCIÓN

Según los objetivos planteados, seguiremos unos pasos ordenados para la recolección de datos de la fuente, de transformación, enriquecimiento de los datos (usando genderize.io), nueva transformación y posterior carga en Elasticsearch y kibana.

El proceso del tratamiento de datos de la práctica con todos los ficheros/elementos de entrada y salida se representan en el siguiente esquema:



TRATAMIENTO DE LOS DATOS CON PYTHON

Elegimos Python como programa para el tratamiento de los datos teniendo en cuenta que durante el desarrollo del Máster hemos utilizado varias librerías adecuadas para la transformación y limpieza de datos.

Como primer paso en el proceso, nos enfrentamos al problema de: **la codificación del fichero**, que ocasiona errores en los registros con caracteres 'extraños', ejemplo:

```
Traceback (most recent call last):
  File "C:/Users/DORAFA/PycharmProjects/MongoDB/test1_1.py", line 26, in <module>
    for event, elem in context:
  File "C:\Users\DORAFA\Anaconda3\lib\xml\etree\ElementTree.py", line 1222, in iterator
    yield from pullparser.read_events()
  File "C:\Users\DORAFA\Anaconda3\lib\xml\etree\ElementTree.py", line 1297, in read_events
    raise event
  File "C:\Users\DORAFA\Anaconda3\lib\xml\etree\ElementTree.py", line 1269, in feed
    self._parser.feed(data)
xml.etree.ElementTree.ParseError: undefined entity &ouml;: line 43, column 18

Process finished with exit code 1
```

Este obstáculo, implica un pre-proceso del fichero original para cambiar su codificación de ISO-8859-1 a UTF-8, antes de ejecutar el script para cambiar el formato del archivo para subirlo a Elasticsearch (convert.py).

Una vez ajustada la codificación del fichero, realizamos un script sencillo para leer el XML y extraer el primer nombre de cada autor en las publicaciones. Esta operación carga el archivo completo a la memoria y esto ocasiona que las máquinas se queden sin recursos y no sea posible completar la ejecución, por lo que nos enfrentamos a un segundo obstáculo por resolver: **El tamaño del fichero**.

Para resolver este problema, investigamos a través de diferentes fuentes hasta encontrar una alternativa usando el paquete xml.etree.ElementTree, que se especializa en documentos XML y contiene diversas clases y funciones que se pueden utilizar para este propósito.

La clase Element de este paquete, nos permite inspeccionar el documento XML mediante el acceso a sus métodos y atributos, de esta manera analizamos cada elemento del XML por separado para seleccionar los atributos que serían necesarios en el análisis y para omitir aquellos tipos de documentos que no se quieren analizar, esto también reducirá el tamaño del fichero.

Una vez superado el obstáculo del tamaño del XML, obtenemos un nuevo fichero de texto llamado UniqueNames.txt, que contiene la lista de nombres únicos (no repetidos) de los autores de los artículos con 116383 registros.

Realizamos a continuación un nuevo script (Genderizer.py) para llamar al API genderize.io y enviar la petición de todos los nombres únicos de autores almacenados en UniqueNames.txt, con este proceso obtenemos como resultado un fichero en formato JSON que contiene entre otros datos, el género de todos los nombres únicos identificados en los autores de los artículos publicados en dblp. La respuesta del API genderize.io la guardamos en el fichero JsonGenderize.json cuyo formato es el siguiente:


```
{"R\u00e9mi": {"gender": "male", "probability": 0.98, "count": 142},
"Sanjeev": {"gender": "male", "probability": 1, "count": 125},
"Hans": {"gender": "male", "probability": 0.99, "count": 431},
```

```
"Nathan": {"gender": "male", "probability": 1, "count": 1632},
"Oded": {"gender": "male", "probability": 1, "count": 12},
```


A partir de este fichero, a continuación realizaremos el script definitivo (XML_to_JSON.py) para:

- Leer el XML original.
- Filtrar los datos para quedarnos solo con los elementos seleccionados.
- Convertir el xml a estructura de diccionario
- Cruzar para cada autor la información de género almacenada en JsonGenderize.json
- Crear una única estructura de datos en formato json que integre la información original más la información de género de autores.
- almacenar los resultados finales en un json

La estructura definitiva de la información a almacenar en la base de datos se describe a continuación.

PROCESO EN ELASTICSEARCH Y KIBANA

Diseño de la base de datos en Elasticsearch

Teniendo en cuenta el análisis planteado en los objetivos, y el proceso descrito anteriormente, hemos decidido quedarnos sólo con los atributos de interés para el análisis de cada artículo. La decisión de diseño que se detalla en este punto, fue tenida en cuenta en el script de tratamiento de datos en python (XML_to_JSON.py).

Según la información de nuestra fuente de datos original, hemos seleccionado por cada publicación los siguientes elementos:

- Autores: Lista de autores que han participado en la publicación.
- Título: Nombre de la publicación.
- Año: El año en que se publicó el artículo

y según el fichero de género de los nombres, tenemos una serie de información adicional para enriquecer la información:

- Nombre: Primer nombre de cada autor
- Género: Género asignado por Genderize.io
- Probabilidad: Un valor del porcentaje de acierto estimado por el API para asignar la predicción de género al nombre indicado.
- Contador: Indica cuántas veces fue encontrado el nombre en genderize.io.

Una vez realizada la verificación preliminar de los datos, decidimos como primer diseño de información integrada el siguiente:

```
{
"_index": "practice",
"_type": "article",
"_id": "907749525ebd11e983b75c5f672e1c64",
"_source":
      {
      "authors": [
            {"author_id": 6195954301384858266,
            "author": "Nathan Goodman",
            "name": "Nathan",
            "gender": "male",
            "probability": 1,
            "count": 1632},
            {"author_id": 9188571742520601015,
            "author": "Oded Shmueli",
            "name": "Oded",
            "gender": "male",
            "probability": 1,
            "count": 12}
            ],
      "IDtitle": 7563490448878735236,
      "title": "NP-complete Problems Simplified on Tree Schemas.",
      "year": 1983
      }
}
```

En este primer diseño, nos encontramos con problemas durante la carga de datos en Elasticsearch debido a que los autores se encontraban anidados dentro de la estructura de las publicaciones, lo cual ocasiona problemas durante la creación del mapping y limitaciones en el acceso a los atributos de los autores, dificultando las tareas analíticas.

Revisando las alternativas para resolver los problemas identificados con el primer diseño para los datos, decidimos que la mejor opción para analizarlos, era desnormalizar la información por autores, para evitar anidarlos dentro de las publicaciones. Es decir, aplanamos los datos para que en la estructura del json todos los atributos dentro del _source estén en el primer nivel. Ejemplo de la estructura para una publicación con dos autores que se desdobla en dos registros independientes:

```
{
"_index": "practice",
"_type": "article",
"_id": "907749525ebd11e983b75c5f672e1c64",
"_source":
      {"author_id": 6195954301384858266,
      "author": "Nathan Goodman",
```

```
      "name": "Nathan",
      "gender": "male",
      "probability": 1,
      "count": 1632,
      "IDtitle": 7563490448878735236,
      "title": "NP-complete Problems Simplified on Tree Schemas.",
      "year": 1983}
}

{
"_index": "practice",
"_type": "article",
"_id": "907749525ebd11e983b75c5f672e1c65",
"_source":
      {"author_id": 9188571742520601015,
       "author": "Oded Shmueli",
       "name": "Oded",
       "gender": "male",
       "probability": 1,
       "count": 12,
      "IDtitle": 7563490448878735236,
      "title": "NP-complete Problems Simplified on Tree Schemas.",
      "year": 1983}
}
```

Este nuevo diseño, aunque es una estructura más simple de información y facilita el mapping para Elasticsearch, representa un aumento significativo en la dimensión de los datos, lo que impacta en el consumo de recursos y tiempo de ejecución del proceso de transformación, ya que implica que por cada autor de una publicación se realizará la búsqueda de información de género y se creará un registro completo de información.

La ejecución del proceso completo (XML_TO_JSON.py) para el conjunto completo de datos cruzados con el resultado de genderize.io, genera el fichero definitivo datos_formato_elastic.json que subiremos a Elasticsearch para realizar los análisis.

Para cargar en Elasticsearch este diseño definitivo de información hemos creado el siguiente mapping (mappingPractice.json):

```
{"practice":
      {"mappings":
            {"article":
                  {
                  "properties":
                        {
                        "order":{"type":"integer"}
                        "IDtitle":{"type":"keyword"},
                        "title":{"type":"keyword"},
                        "year":{"type":"integer"},
                        "author_id":{"type":"keyword"},
                        "author_full_name":{"type":"keyword"},
                        "author_first_name":{"type":"keyword"},
                        "gender":{"type":"keyword"},
```

```
                        "probability":{"type":"integer"},
                        "count":{"type":"integer"}
                        }
                    }
                }
        }
}
```

Carga de datos en Elasticsearch y Kibana

Una vez tenemos listos los ficheros:

- datos_formato_elastic.json
- mappingPractice.json

Ejecutamos el siguiente proceso para la carga de datos a elasticsearch y kibana:

1. Desde la Terminal levantamos elasticsearch:
   `~/Programas/elasticsearch-6.6.1/bin$ ./elasticsearch`

2. En una terminal nueva, creamos un índice llamado practice:
   `curl -XPUT -k localhost:9200/practice`

3. Consultamos el índice creado:
   `curl -k localhost:9200/_cat/indices`

4. Creamos el mapping a partir del fichero mappingPractice.json:

```
elasticdump --input=mappingPractice.json
--output=http://localhost:9200 --type=mapping --output-index=practice
--headers='{"Content-Type": "application/json"}'
```

5. Consultamos la creación del mapping:
   `curl -k localhost:9200/practice/_mapping/`

6. Cargamos nuestros datos en el índice creado:

```
elasticdump --input=datos_formato_elastic.json
--output=http://localhost:9200 --type=data --output-index=practice
--headers='{"Content-Type": "application/json"}'
```

7. Una vez recibimos la confirmación de carga completa de nuestros datos, en una terminal nueva lanzamos Kibana:
   `~/Programas/kibana-6.6.1-linux-x86_64/bin$ ./kibana`

8. Abrimos el navegador y vamos a localhost:5601 y en la interfaz de Kibana creamos un nuevo Index pattern asociado a nuestro índice 'practice'. a partir de este momento ya tenemos disponibles todos los atributos para realizar el análisis objetivo de esta práctica.

ANÁLISIS DE LOS DATOS

En el apartado de Objetivos ya hemos trazado el principal propósito de esta práctica. Uno de ellos deriva del interés sobre el campo de trabajo de la detección de género a partir del nombre (given name) y en particular de la herramienta *Genderize.io* que hemos elegido. El otro tiene que ver con el dataset de trabajo.

En general, después de la transformación realizada, tenemos para el análisis la siguiente información:

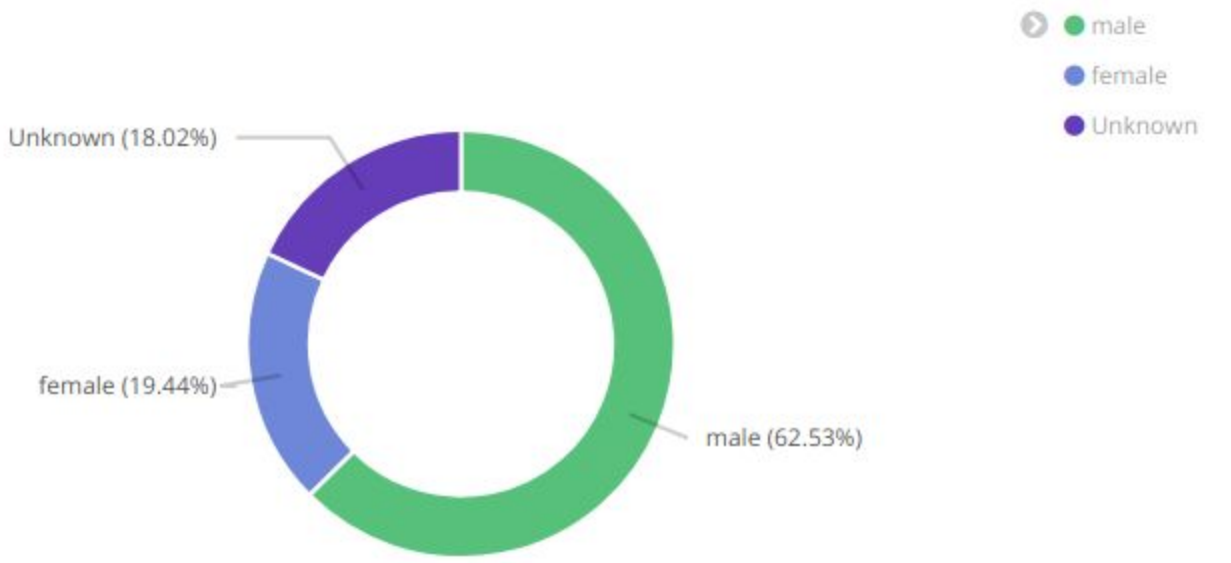
- Total líneas del fichero datos_formato_elastic.json después de los tratamientos: 85.971.420
- Total líneas de la estructura JSON por autor: 15
- Total autores y sus publicaciones, que serán analizados: 85.971.420/15= 5.731.428

A continuación describiremos los resultados de los principales paneles de Kibana desarrollados con el fin de responder a las preguntas planteadas.

Aunque se trata de un análisis muy interesante a nivel de representación gráfica, para optimizar espacio se incluirán solo algunas visualizaciones que ayudan a esclarecer los objetivos propuestos.

Análisis de la herramienta Genderize.io

Para analizar cómo de bueno es el acierto de la herramienta vamos a representar el porcentaje de géneros obtenidos a lo largo de toda la muestra de datos. De esta manera seremos capaces de ver el porcentaje de nombres que han sido etiquetados como *unknown.* A continuación se muestra el porcentaje de géneros masculino, femenino y géneros Unknown de nuestra muestra de datos:



Podemos identificar de primera mano, que el porcentaje de Unknown es bastante considerable (cercano al 20%). A continuación, para verificar algunos de los casos en los que la API no ha sido capaz de determinar el género, revisamos algunos ejemplos e dientificamos las características más comunes:

- Uso de caracteres extraños: Esto sucede por ejemplo en case de nombres que contienen caracteres extraños:

author_first_name: Descending	Count
þorgerður	1
Þröstur	1
Ün	2
Ümüt	1
Ümmüsan	1
Ülo	19
Ülle	52

- Acentos: Por otra parte encontramos varios registros donde los nombres llevan tildes. En este caso los nombres tampoco son tratados por el algoritmo y devuelven Unknown:

author_first_name: Descending	Count
Óscar	474
Òscar	11
Đurdica	1
Đura	1
Íñigo	10
Ínigo	1
Ígor	1
Í	2
Ìtalo	4
Éverton	3

author_first_name: Descending	Count
Émile	20
Émerson	1
Émeline	3
Éléonore	1
Éloi	30
Élodie	16
Élisabeth	14
Élisa	12
Élie	4
Élias	1

- Iniciales de autores: Pese a otros casos descritos, la mayoría de los registros en los que la API devuelve un Unknown, es debido a que los autores han usado solo sus iniciales y no el nombre completo, lo cual indica que estos fallos no son atribuibles al API. Por ejemplo:

author_first_name: Descending	Count
M	27,358
J	22,765
A	22,094
S	20,760
R	14,623
C	13,080
K	11,765
P	11,507
H	10,839
D	10,216

Como siguiente paso elegiremos al azar una serie de nombres y haremos un análisis intuitivo de géneros de la muestra elegida. De esta forma vamos a poder presentar una estimación de las métricas de *precisión* y *recall* que por ejemplo puede dar una idea sobre qué porcentaje de predicciones positivas fueron correctas o qué porcentaje de casos positivos fueron capturados. Para ello definiremos que si el género ha sido identificado correctamente es un 1, si ha sido identificado erróneamente es un negativo o 0. Entonces se definen los siguientes casos:

1. TN / True Negative: Caso cuando fue negativo y fue predecido negativo (nombres imposibles de predecir; estos nombres estarán en la muestra de Unknown).
2. TP / True Positive: Caso cuando fue positivo y fue predecido positivo (nombres predichos correctamente; estos nombres estarán en la muestra de male/female).
3. FN / False Negative: Caso cuando fue positivo y fue predecido negativo (nombres cuyo género es posible predecir pero la API no lo hizo; en la muestra de Unknown).
4. FP / False Positive: Caso cuando fue negativo y fue predecido positivo (nombres donde la API se equivocó, en la API de male/female).

La métrica de *precisión* se calcula como:

```
precisión = True Positive / (True Positive + False Positive)
```

La de *recall* o *recuperación*, se calcula como:

```
recuperación = True Positive / (True Positive + False Negative)
```

Hemos cogido una muestra de aproximadamente 100 autores y hemos analizado los géneros de nombres de forma empírica. Para ello, en casos de dudas, además de utilizar internet, hemos consultado a gente que conoce tanto la cultura oriental, como la occidental. Y la conclusión es que, salvo nombres que son unisex y nombres orientales que suelen en muchos casos ser unisex también, la API funciona muy bien y tiene un alto porcentaje de True Positives y muy bajo de False Positives. Es por ello porque es útil la asignación de una probabilidad a cada respuesta de la API.

Sin embargo, en esta revisión manual no habíamos considerado las muestras del tipo Unknown y constituyen un porcentaje muy considerable de los datos. Por ello hemos decidido analizar por separado una muestra de 100 unknown y obtener de allí los False y True Negatives.

A continuación se detallan los resultados del análisis de esta muestra:

- False negatives: 19 de 100 unknowns (algunos de estos casos son nombres no correctamente escritos, con comillas, o de las que conocemos su género con certeza.)
- True Negatives: 26 de 100 unknowns (algunos de estos casos son: iniciales, nombres chinos compuestos, es decir que claramente son apellido/nombre, apellidos occidentales, por ejemplo eslavos).
- Resto de casos: 100 - (19 + 26) = 55

Por último, podemos calcular las métricas, aunque por la proporción ya podemos adelantar que la API tiene en general bastante precisión:

precisión = True Positive / (True Positive + False Positive) ~ 99 / (99 + 1) = 0,99

recuperación = True Positive / (True Positive + False Negative) ~ 99 / (99 + 19) = 0,83

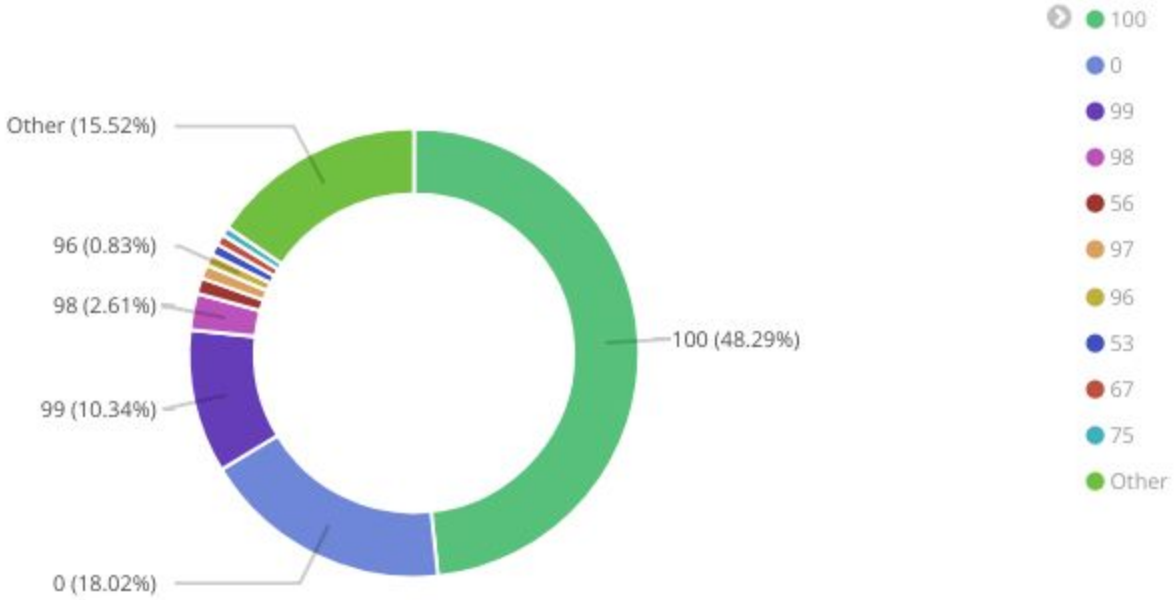
Como conclusión podemos indicar que el algoritmo tiene una precisión muy alta (TP alto y FP bajo) y una capacidad de recall o recuperación alta también (TP alto y FN bajo). El dato de la probabilidad le da una fiabilidad adicional indicando la probabilidad de acierto con un nombre dado. Por otra parte, podemos ver que el algoritmo depende mucho del preprocesado de los datos, como en cualquier problema de la Ciencia de datos, especialmente si se trata de procesamiento de texto.

Cabe destacar que estos datos están muy ligados a la probabilidad de acierto que devuelve el algoritmo, que revisamos a continuación.

Probabilidad de la predicción del género

Además de la 'etiqueta' de género que ofrece genderize.io, cada predicción viene acompañada de una probabilidad de acierto que el API asigna, según el nivel de certeza de la predicción. a continuación realizaremos un pequeño análisis de estas probabilidades asignadas a nuestro conjunto de datos:

Gráfico de probabilidades de asignación de género de Genderize.io:



Como podemos observar en este gráfica, la mayoría de las predicciones realizadas por genderize.io (descartando unknows), están respaldadas por una alta certeza en la predicción, ya que la mayoría de las asignaciones de género tienen probabilidades por encima del 95%.

Gender	Media de probabilidades	Probabilidad Minima	Probabilidad Máxima	50th percentile of Mediana de Probabilidad
male	95.62	50	100	100
female	86.182	50	100	99
Unknown	0	0	0	0

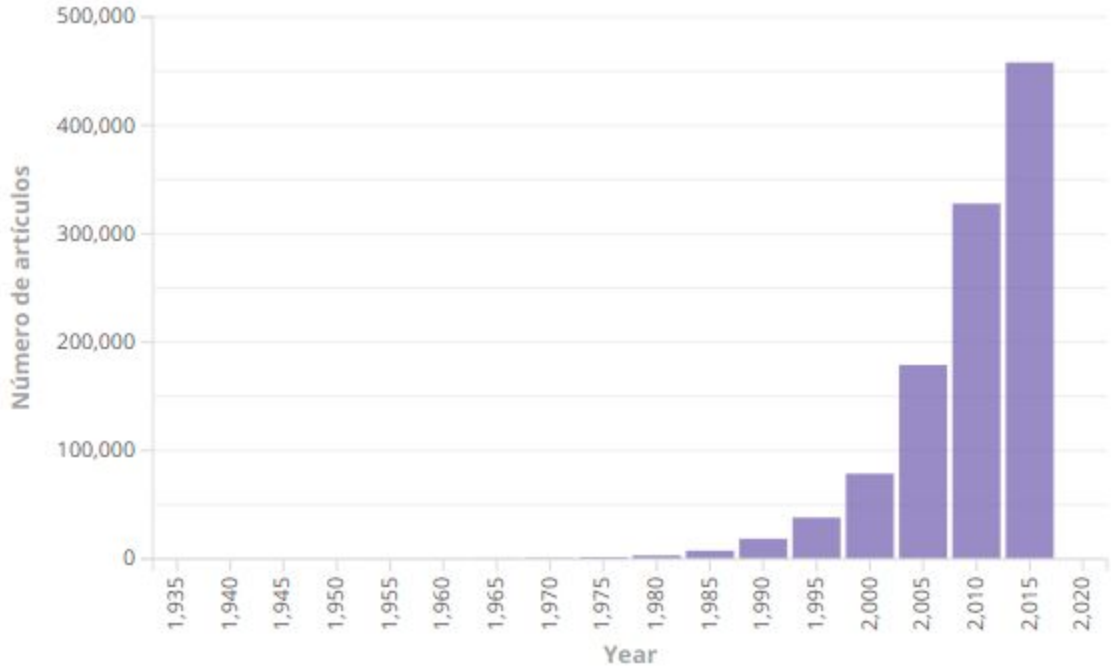
Estos resultados indican que esta herramienta, además de tener un alto recall (en el análisis manual realizado), la media de la probabilidad con la que decide el género es muy alta.

Un detalle interesante de revisar en este resumen estadístico, es que aunque las probabilidades con que asigna el género son bastante altas, para el género 'female' la media de las probabilidades que acompaña a las predicciones son significativamente menores que las del género 'male', con un 86,18% frente a un 95,62% respectivamente.

Análisis de género DBLP

Como ya hemos mencionado, una vez finalizado el proceso de tratamiento de datos, disponemos para nuestro análisis 5.731.428 de registros de datos que abarcan nombres de autores de publicaciones de tipo *article* desde el año 1936 hasta 2019.

Iniciamos el análisis, representamos la tendencia prácticamente exponencial de participación de la mujer en la publicación de artículos a través de los años:



Se ve claramente como la mujer, ha pasado de publicar menos de 100 artículos entre 1936 y 1965, a participar en más de 450.000 artículos publicados sólo entre 2015 y 2019.

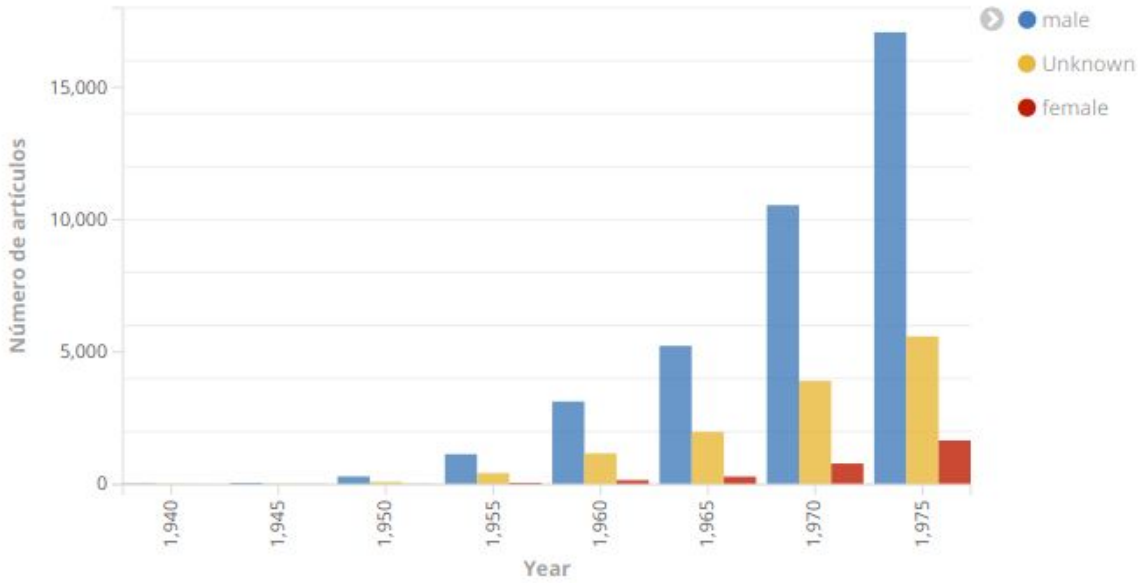
Hemos elegido como punto de inflexión en esta tendencia alrededor de los años 80. Es el momento cuando el número sobrepasa los 5000 y comienza su crecimiento exponencial, por lo que revisaremos las cifras en rangos de tiempo, para confirmar la información gráfica:

1,930 to 1,979: Year		1,980 to 1,999: Year		2,000 to 2,019: Year	
Gender	**Count**	**Gender**	**Count**	**Gender**	**Count**
male	33,680	male	325,490	male	3,106,926
female	2,571	female	58,115	female	1,015,732
Unknown	11,952	Unknown	75,788	Unknown	906,651

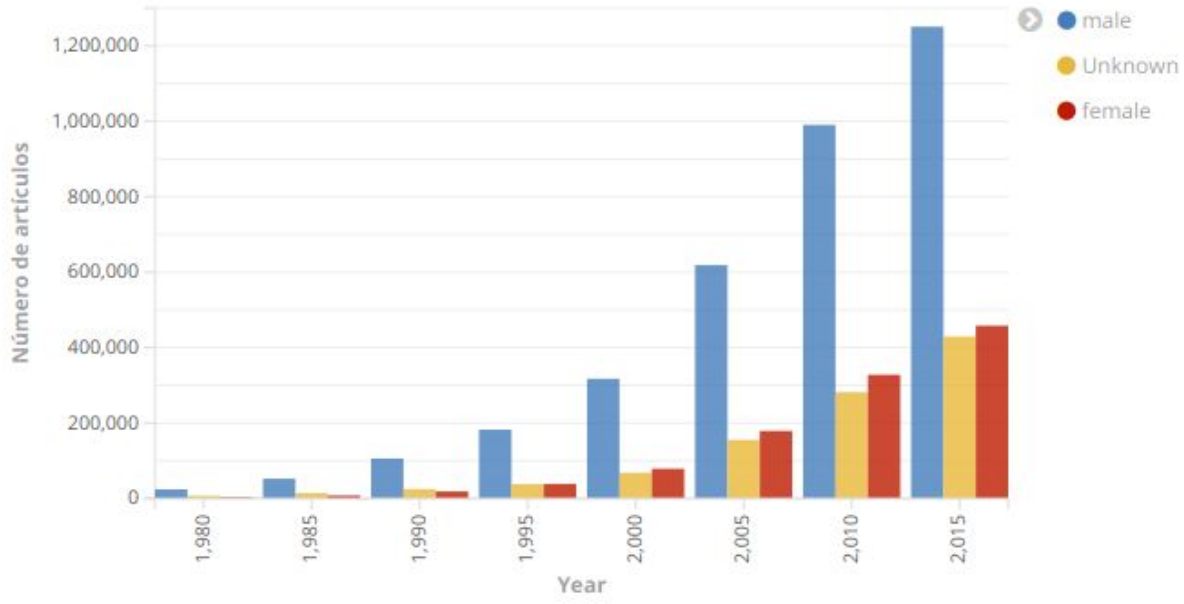
Las cifras del cuadro anterior, confirman el crecimiento de la participación de la mujer en la investigación, por rangos en el tiempo. sin embargo, al incluir en el resumen todos los géneros, empezamos a darnos cuenta, que este crecimiento obedece a una tendencia de crecimiento global en la publicación de artículos científicos, ya que los hombres tienen también una tendencia de crecimiento muy importante.

Para establecer gráficamente las diferencias de género, vamos a realizar algunas representaciones gráficas a continuación:

Evolución de números de géneros de autor hasta los años 80:



Números de autores por género desde los años 80 en adelante:



Podemos ver que el ratio de los Unknown/Male es casi siempre el mismo, sin embargo Female/Male crece significativamente a partir de los año 80.

Estas gráficas dejan en evidencia otra realidad importante, y es que a pesar del crecimiento en la participación de la mujer en el ámbito de las publicaciones científicas a través del tiempo, en comparación con la participación de los hombres, la brecha es muy amplia incluso en la actualidad.

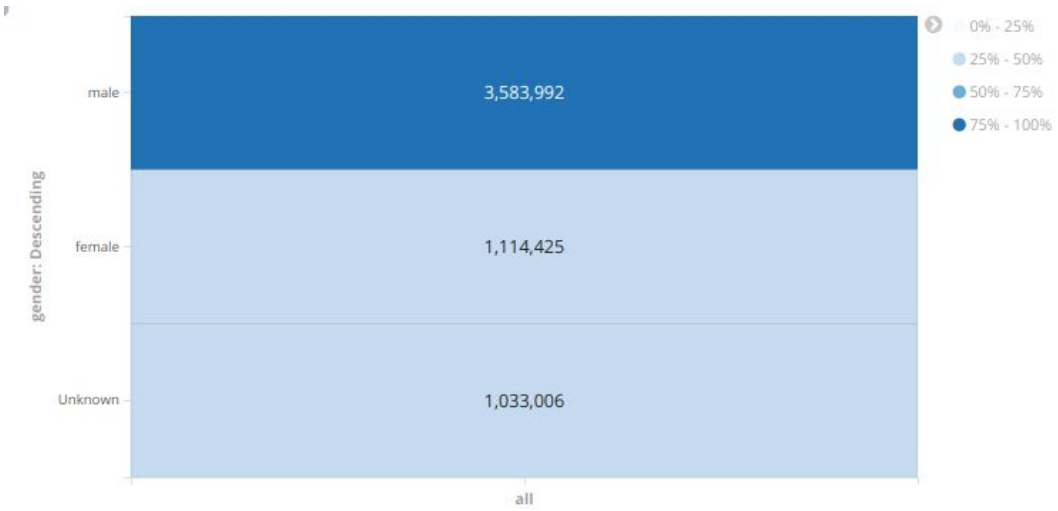
Si hacemos un zoom a los últimos 4 años (representados en la gráfica anterior como 2015), se ve claramente que la participación de las mujeres en la ciencia, es solo de ⅓ de la participación en publicaciones realizadas por hombres.

La siguiente consulta muestra los Top 10 de autores que más han publicado desde 1936 y podemos ver que solo 2 son mujeres:

Autor	Género	Probabilidad	Count
H. Vincent Poor	Unknown	0	1,139
Lajos Hanzo	male	100	753
Witold Pedrycz	male	100	749
Mohamed-Slim Alouini	male	99	699
Chin-Chen Chang 0001	male	53	604
Victor C. M. Leung	male	100	565
Dacheng Tao	Unknown	0	531
Wei Li	female	54	531
Wei Zhang	female	54	528
Licheng Jiao	male	100	500
		660	**6,599**

Si revisamos grandes totales, nos damos cuenta que el camino de la igualdad de participación entre mujeres y hombres en la comunidad científica, según el número de publicaciones, está aún por recorrer.

El siguiente mapa de calor, muestra las cifras totales de publicaciones por género entre 1936 y 2019, y el color representa el rango porcentual en el que se encuentra cada grupo:



CONCLUSIONES

Según los objetivos propuestos para el desarrollo de esta práctica, y los análisis realizados sobre los resultados, destacamos las siguientes conclusiones:

Según las pruebas realizadas y los resultados obtenidos, la herramienta de generificación Genderize.io ha logrado un desempeño destacado para asignar el género al conjunto de autores que participan en las publicaciones de artículos DBLP, que se respalda en las siguientes cifras:

- Ha logrado asignar el género al 80% de los autores de publicaciones de artículos de DBLP.
- Del 20% restante que no ha logrado etiquetar la herramienta, la mayoría de los casos, corresponden a autores que firman las publicaciones únicamente con una inicial, lo cual resulta insuficiente para predecir el género y no sería atribuible a la herramienta. Otros casos de fallos corresponde a nombres que incluyen caracteres extraños y/o acentos.
- Las predicciones de género realizadas como Male o Female, se han asignado con una probabilidad superior al 90% de media.
- En un muestreo aleatorio con revisión manual de las asignaciones de nombres de hombres y mujeres, el API ha obtenido un resultado de precisión del 99%.

A pesar de los buenos resultados generales obtenidos por el API genderize.io al realizar la tarea de asignar el género a los autores de artículos de DBLP, resaltamos que la media de la probabilidad con la que asignó género Female es inferior a la media de la probabilidad con la que asignó género Male con un con un 86,18% frente a un 95,62% respectivamente.

En términos generales podemos considerar el API Genderize.io como una herramienta adecuada y fiable para usar en este tipo de casos en que es necesario enriquecer una fuente de datos para incluir el género de un nombre dado.

En cuanto al análisis de género, hemos visto claramente como la participación de la mujer en la comunidad científica, a través de la publicación de artículos, ha ido aumentando a lo largo del tiempo de manera muy destacable, lo cual resulta positivo al comparar las cifras desde 1936 hasta 1980 frente al total de publicaciones a partir de los años 80.

Sin embargo, este crecimiento en la participación de las mujeres, puede estar obedeciendo en general a la tendencia global de aumento en el número de publicaciones de artículos con el paso del tiempo, que también es exponencial.

Los datos dejan de ser alentadores en cuanto realizamos un análisis comparativo de las publicaciones entre hombres y mujeres, ya que la tendencia de la brecha entre los dos géneros a lo largo del tiempo no ha cambiado apenas nada y la participación de la mujer en la comunidad científica es de ⅓ en comparación con la participación de los hombres tanto antes de los 80's como en el mismo 2019.

Este análisis no pretende identificar las razones de estas diferencias, pero busca evidenciar la brecha actual entre géneros y que según los resultados observados que al parecer se ha mantenido en los últimos 100 años.

Esperamos que este tipo de cifras permitan mostrar, que a pesar de los esfuerzos de la sociedad y los gobiernos, por favorecer la inclusión y participación de la mujer en el ámbito científico, y a pesar del crecimiento demostrado, la brecha continúa igual de abierta que hace 100 años, con lo que aún nos queda mucho trabajo por hacer, para lograr una participación más representativa de la mujer en este ámbito.

LECCIONES APRENDIDAS

Hemos realizado un trabajo de análisis de datos que requiere de prácticamente de todos los pasos de un proyecto de Data Science, incluyendo obtención, tratamiento, visualización y análisis de datos de publicaciones de artículos científicos del repositorio DBLP.

Los datos que se han usado son principalmente textuales y categóricos, de gran tamaño, por lo que han sido analizados usando Elasticsearch y kibana, lo cual encaja con los conceptos vistos durante este módulo de la Asignatura y ha favorecido los análisis para cumplir los objetivos propuestos en este estudio.

Durante el desarrollo de la parte de visualización hemos mejorado nuestras habilidades con Kibana, un plugin de visualización para ElasticSearch que permite hacer una visualización de forma fácil para datos grandes sobre ElasticSearch y sin olvidar de la eficiencia y estética.

Hemos hecho un estudio de género de autores de la DBLP, Digital Bibliography & Library Project y otro de precisión de la herramienta Genderize.io que hemos utilizado para conseguir los objetivos propuestos.