# Proyecto

Integrantes: María Paula Herrero, Rosana Vargas

Una organización de Derechos Humanos (DDHH) requiere identificar noticias relacionadas a violaciones de DDHH que sean publicadas en los sitios web de los periódicos nacionales de Venezuela, con el fin de usar dicha información como base para escribir sus informes sobre la situación de los DDHH en Venezuela (trimestrales, anuales).

La idea es automatizar este proceso y entregar via correo electrónico a los integrantes de la organización el grupo de noticias seleccionados en formato pdf.

**Objetivo**: Identificar si una noticia de un periódico publicado en línea (web) pertenece a una categoría de derechos humanos en particular

**Fuente de datos**: URL de noticias seleccionadas por un experto (Raúl Cubas) durante tres meses y tabla de clasificación de dichas noticas.

**Tarea de minería de datos**: clasificación

A) Categorización de noticias de DDHH

**Motivación principal** = posibilidad de realizar una clasificación automática de noticias según clasificación de DDHH, en función de los temas abordados en el texto.

Este es un trabajo que se hace a diarios e implica revisar muchos sitios web, es

Importante para llevar a cabo una asignación adecuada de clasificaciones.

**Objetivo**: Realizar, de manera automática, una categorización de noticias según la clasificación de DDHH.

**Fuente de datos**: URL de noticias seleccionadas por un experto (Raúl Cubas) durante tres meses y tabla de clasificación de dichas noticas.

**Tarea de minería de datos**: clasificación

(1) Recolección de datos:

El corpus está constituido por noticias categorizadas según clasificación de DDHH para un período de 4 meses en el año 2018 publicadas en los periódicos El Nacional y El Universal.

|  |  |  |
| --- | --- | --- |
| **CATEGORIA** | **DESCRIPCION** | **TOTAL** |
| 1.1 | Derecho a la vida | 15 |
| 1.2 | Derecho a la libertad personal | 39 |
| 1.3 | Derecho a la Integridad personal | 6 |
| 1.7 | Derecho a la libertad de expresión e información | 3 |
| 1.10 | Derecho a la justicia | 8 |
| 2.1 | Derecho a la Salud | 25 |
| 2.2 | Derecho a la Educación | 9 |
| 2.3 | Derecho al trabajo | 44 |
| 2.6 | Derecho a la seguridad social | 5 |
| 2.8 | Derecho a la Vivienda | 22 |
| 2.9 | Derecho de los Pueblos Indios | 3 |
| 2.10 | Derecho a la alimentación | 10 |
| 3.1 | Derecho a la tierra | 3 |
| 3.2 | Derecho al ambiente | 12 |
| 4.3 | Derecho a manifestar pacíficamente | 1 |
| 5 | Actividades de las ONG nacionales e internacionales de ddhh en Venezuela | 9 |
| 12 | Participación | 6 |
| 15 | Provea en prensa | 1 |
| 18 | Instancias internacionales de ddhh | 22 |
| CONTEXTO | CONTEXTO | 22 |
|  | Total: | 265 |

Un dato posee los siguientes atributos:

|  |  |
| --- | --- |
| **Atributo** | **Significado** |
| id | Identificador único |
| idPeriodico | Identificador único de la noticia en el periódico |
| periodico | Periódico (EN: El Nacional, EU: El Universidad) |
| url | Dirección URL donde se consigue la noticia en la Web |
| fecha | Fecha en formato dd/mm/AAAA |
| fechaPublicacion | Fecha de publicación del periódico |
| autor | Nombre del autor de existir, si no aparece NO\_TIENE |
| seccion | Sección del periódico a la que pertenece la noticia |
| palabrasClaves | Palabras claves con la que el periódico clasifica la noticia (EU no usa palabras claves) |
| titular | Titular de la noticia |
| subtitular | Sumario o subtitular de la notica (párrafo destacado) |
| imagen | URL de la imagen asociada a la noticia |
| contenido | Todos los párrafos del texto de la noticia |
| derecho | Identificador del DDHH principal asociado a esta noticia. Ver lista de identificadores |
| otrosDerechos | Otros identificadores de DDHH asociados a la misma noticia, separados por comas de existir, si no esta vacío |
| subDerechos | Identificadores de las subcategorías del DDHH principal asociado de existir, si no esta vacio |
| EP | SI: indica que la noticia ocurre En Protestas (EP)  No: indica que la noticia no ocurre en protestas |

De estas noticias se consideran el titular, el subtitular, el contenido y el derecho

También se considera para una segunda clasificación EP, dado que es necesario categorizar por derecho y EP.

No se consideran los atributos: id, idPeriodico, url, fecha, fechaPublicacion, imagen; por que funcionan como constantes

No se consideran los atributos otrosDerechos y subDerechos, por que solo vamos a clasificar según Derecho principal

No se consideran los atributos Periódico, autor, sección, las palabras claves; por que solo sirven para procesar la notica, no para clasificarla

(2) Preparación de datos:

* Se revisaron los documentos PDF, se extrajo los datos necesarios y se creó el archivo CSV a usar en el proceso de Minería.
  + Usando Python se creo un archivo csv con los siguientes datos de cada archivo PDF entregado:
    - No.
    - Nombre archivo
    - Fecha
    - Titular
    - Periodico
    - Fecha Periodico
    - URL
    - Derecho PPal
    - Otros Derechos
    - SubDerechos
    - EP

Este script de Python, toma cada uno de los PDFs del directorio “toda” y del nombre de cada archivo obtiene el derecho principal (que es el primero que aparece), los otros derechos y el subderecho. También extrae la fecha y el periódico. Si el primer derecho que aparece no tiene dos niveles, se extrae y se pone como subderecho el original. De este nombre también se extrae si la noticia tiene EP en si, si aparecen estas siglas en dicho nombre, si no aparece es no. Por otro lado, del cuerpo del pdf se obtiene el URL y el titular.

* + Usando Python, a partir del archivo CVS anterior, se buscó el contenido de la noticia a través del URL y se creo un segundo archivo CSV con la siguiente información:
    - id
    - idPeriodico
    - periodico
    - url
    - fecha
    - fechaPublicacion
    - autor
    - sección
    - palabrasClaves
    - titular
    - subtitular
    - imagen
    - contenido
    - derecho
    - otrosDerechos
    - subDerechos
    - EP

Usando el campo URL, el script se conecta al sitio web del periódico y se trae el contenido de la noticia. De ahí obtiene los datos del idPeriódico, el periódico como tal (ya no usamo el del nombre del pdf dado que puede estar incorrecto), la fecha de publicación, el autor, la sección donde el periódico lo clasifica, las palabras claves que el periódico le coloca de existir (NO\_EXISTE en caso contrario), el titular, subtitular, el url de la imagen, el contenido. Los últimos campos son los mismos que el csv anterior (derecho, otrosDerechos, subDrechos, EP)

* + Una vez generado el csv con todos los datos, se abre este archivo con un editor de texto, se cambian las dobles comillas ("") por slash comilla (\"), se abre en rapidMiner, indicando que los campos se separan con punto y coma (;) y que no tome comentario (#).
* Separación de objetos o tokens (tokenization)
  + Los tokens identifican partes en las cuales de puede dividir un texto.
  + Pueden ser de diferente granularidad, dependiendo de la aplicación (párrafos, frases, n-gramas, palabras, ...)
* Eliminación de palabras no informativas (stop words) y signos de puntuación:
  + Construcción un **diccionario de palabras frecuentes en el español** que no aportan información para la tarea de clasificación de textos (artículos, adjetivos, pronombres, entre otros)
  + Estas palabras fueron eliminadas de los documentos aplicando un proceso de comparación con el diccionario
* Lematización (stemming): aplicación del algoritmo **Porter Stemmer** para el español
  + Formas morfológicas de una misma palabra tienen el mismo significado.
  + El efecto de la lematización es una reducción de la dimensionalidad (al reducir un conjunto de palabras a una sola representación)
* Generación de vectores: se obtuvo un total de **5.063** términos (lexicón)
* Reducción de la dimensionalidad:
  + Se aplicaron diferentes técnicas de selección de atributos con el fin de identificar los términos más informativos para la tarea de categorización. Se estudiaron los resultados de Ganancia de Información, Chi-Cuadrado, Razón de Ganancia y Correlación, decidiendo que la mejor selección la hacía Ganancia de Información según el experto, seguido por el de Correlación.
  + Para definir el número de características finales se elaboró un sistema de votación eligiendo aquellos atributos que tenían un peso igual o mayor a 0,095 según el criterio de Ganancia de Información.
  + La cantidad de atributos obtenidos luego de evaluar el sistema de votación fueron 226.
  + Sin embargo, cuando se probó este grupo de datos con el algoritmo Bayes ingénuo, luego de varias pruebas se concluyó que el mejor peso era el 0,04
* Balancear las clases
  + Se eliminaron las categorías con muy poca muestra dejando solamente la siguiente data:

|  |  |  |
| --- | --- | --- |
| **CATEGORIA** | **DESCRIPCION** | **TOTAL** |
| 1.1 | Derecho a la vida | 15 |
| 1.2 | Derecho a la libertad personal | 39 |
| 2.1 | Derecho a la Salud | 25 |
| 2.3 | Derecho al trabajo | 44 |
| 2.8 | Derecho a la Vivienda | 22 |
| 2.10 | Derecho a la alimentación | 10 |
| 3.2 | Derecho al ambiente | 12 |
| 18 | Instancias internacionales de ddhh | 22 |
| CONTEXTO | CONTEXTO | 22 |
|  | Total: | 211 |

* + Luego se probó quitando las categorías con menos muestra (1.1, 2.10 y 3.2), pero esto no mejoró el modelo mucho.
  + También se probó quitando 18 y CONTEXTO, esto sí mejoró el modelo. Evaluando con el experto llegamos a la conclusión que ambas son categorías más generales, mientras que las demás son solo derechos, por eso la diferencia.
  + Además, para mejorar el balance de muestra se agregaron más casos a las clases con menos datos, quedando la muestra en:

|  |  |  |
| --- | --- | --- |
| **CATEGORIA** | **DESCRIPCION** | **TOTAL** |
| 1.1 | Derecho a la vida | 19 |
| 1.2 | Derecho a la libertad personal | 39 |
| 2.1 | Derecho a la Salud | 25 |
| 2.3 | Derecho al trabajo | 44 |
| 2.8 | Derecho a la Vivienda | 22 |
| 2.10 | Derecho a la alimentación | 22 |
| 3.2 | Derecho al ambiente | 15 |
| 18 | Instancias internacionales de ddhh | 21 |
| CONTEXTO | CONTEXTO | 23 |
|  | Total: | 230 |

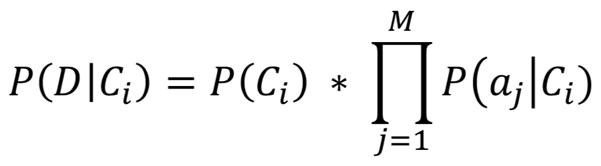
* + Esta nueva muestra, sin las categorías CONTEXTO y 18, más el peso en 0,04 según el criterio de Ganancia de Información, nos dieron un total de 186 registros y 966 atributos para trabajar el modelo.

**Resultado de la fase de preparación de datos**:

Vista minable = Matriz de datos (tabla atributo-valor) con **186** registros (filas) y **966** atributos, incluyendo la clase.

(3) Minería de datos:

* Algoritmo: Bayes Ingenuo con corrección laplace:



* Medida de rendimiento: exactitud predictiva, precisión y recall (sensibilidad), kappa.
* Técnica de evaluación: validación cruzada de 10 particiones
* Matriz de Confusión:

Exactitud = 89,30%

Kappa = 0,873

G-medias = 89,06%

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **true 2.10** | **true 2.8** | **true 2.3** | **true 2.1** | **true 1.1** | **true 1.2** | **true 3.2** | **class precision** |
| pred. 2.10 | 21 | 4 | 0 | 1 | 0 | 0 | 0 | 80,77% |
| pred. 2.8 | 0 | 15 | 0 | 0 | 0 | 0 | 0 | 100,00% |
| pred. 2.3 | 0 | 0 | 37 | 2 | 0 | 0 | 0 | 94,87% |
| pred. 2.1 | 1 | 0 | 5 | 22 | 0 | 1 | 0 | 75,86% |
| pred. 1.1 | 0 | 0 | 0 | 0 | 18 | 0 | 0 | 100,00% |
| pred. 1.2 | 0 | 0 | 1 | 0 | 1 | 38 | 0 | 95,00% |
| pred. 3.2 | 0 | 3 | 1 | 0 | 0 | 0 | 15 | 78,95% |
| class recall | 95,45% | 68,18% | 84,09% | 88,00% | 94,74% | 97,44% | 100,00% |  |

(4) Evaluación e interpretación:

Exactitud = 62.79%

Kappa: 0,559

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CATEGORIA** | **DESCRIPCION** | **TOTAL** | **Precision** | **Recall** |
| 1.1 | Derecho a la vida | 3 | 50% | 100% |
| 1.2 | Derecho a la libertad personal | 24 | 100% | 66,67% |
| 2.1 | Derecho a la Salud | 8 | 36,36% | 50% |
| 2.3 | Derecho al trabajo | 22 | 87,5% | 63,64% |
| 2.8 | Derecho a la Vivienda | 17 | 100% | 47,06% |
| 2.10 | Derecho a la alimentación | 9 | 40% | 66,67% |
| 3.2 | Derecho al ambiente | 3 | 21,43% | 100% |
|  | **Total**: | 86 | **G-media**: | 67,90% |
|  |  |  |  |  |

Datos para la prueba:

|  |  |  |
| --- | --- | --- |
| **CATEGORIA** | **DESCRIPCION** | **TOTAL** |
| 1.1 | Derecho a la vida | 3 |
| 1.2 | Derecho a la libertad personal | 24 |
| 2.1 | Derecho a la Salud | 8 |
| 2.3 | Derecho al trabajo | 22 |
| 2.8 | Derecho a la Vivienda | 17 |
| 2.10 | Derecho a la alimentación | 9 |
| 3.2 | Derecho al ambiente | 3 |
| 18 | Instancias internacionales de ddhh | 5 |
| CONTEXTO | CONTEXTO | 29 |
|  | Total: | 120 |

(5) Difusión y uso:

Datos prueba original

|  |  |  |
| --- | --- | --- |
| **CATEGORIA** | **DESCRIPCION** | **TOTAL** |
| 1.1 | Derecho a la vida | 8 |
| 1.2 | Derecho a la libertad personal | 24 |
| 1.3 | Derecho a la Integridad personal | 5 |
| 1.7 | Derecho a la libertad de expresión e información | 1 |
| 1.10 | Derecho a la justicia | 18 |
| 2.1 | Derecho a la Salud | 8 |
| 2.2 | Derecho a la Educación | 3 |
| 2.3 | Derecho al trabajo | 22 |
| 2.6 | Derecho a la seguridad social | 5 |
| 2.8 | Derecho a la Vivienda | 17 |
| 2.9 | Derecho de los Pueblos Indios | 1 |
| 2.10 | Derecho a la alimentación | 21 |
| 3.1 | Derecho a la tierra | 1 |
| 3.2 | Derecho al ambiente | 6 |
| 4.3 | Derecho a manifestar pacíficamente | 4 |
| 5 | Actividades de las ONG nacionales e internacionales de ddhh en Venezuela | 10 |
| 12 | Participación | 41 |
| 18 | Instancias internacionales de ddhh | 5 |
| CONTEXTO | CONTEXTO | 30 |
|  | Total: | 230 |

Matriz confusión pruebas:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **true 2.1** | **true 1.2** | **true 2.3** | **true 2.8** | **true 3.2** | **true 1.1** | **true 2.10** | **class precision** |
| **pred. 2.1** | 4 | 2 | 3 | 0 | 0 | 0 | 2 | 36,36% |
| **pred. 1.2** | 0 | 16 | 0 | 0 | 0 | 0 | 0 | 100,00% |
| **pred. 2.3** | 0 | 0 | 14 | 2 | 0 | 0 | 0 | 87,50% |
| **pred. 2.8** | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 100,00% |
| **pred. 3.2** | 0 | 5 | 2 | 3 | 3 | 0 | 1 | 21,43% |
| **pred. 1.1** | 1 | 1 | 1 | 0 | 0 | 3 | 0 | 50,00% |
| **pred. 2.10** | 3 | 0 | 2 | 4 | 0 | 0 | 6 | 40,00% |
| **class recall** | 50,00% | 66,67% | 63,64% | 47,06% | 100,00% | 100,00% | 66,67% |  |