

DESAFÍO FINAL

ASIGNACIÓN DE PROYECTOS DE LEY A COMISIONES
PERMANENTES

CONTENIDO

- CONTEXTO DE NEGOCIO
- DESCRIPCIÓN DEL PROBLEMA
- PREPARACIÓN DE LOS DATOS
- MODELOS A UTILIZAR
- MEDICIÓN DE RESULTADOS
- API WEB
- APLICACIÓN CLIENTE

CONTEXTO DE NEGOCIO

CONTEXTO DE NEGOCIO

- A LA CÁMARA DE DIPUTADOS INGRESAN PROYECTOS DE LEY SOBRE DISTINTOS ASUNTOS.
- AL INGRESAR SE REDACTA UN BREVE SUMARIO (DENOMINADO “TÍTULO”) PARA EL PROYECTO, Y SE LO ENVÍA (O "GIRA") A UNA O MÁS COMISIONES DE ESTUDIO DE ACUERDO AL TEMA DEL PROYECTO.
- LAS **COMISIONES DE ESTUDIO SON 45**; ESTÁN FORMADAS POR GRUPOS DE VEINTE A CUARENTA Y CINCO DIPUTADOS Y CADA UNA SE ESPECIALIZA EN UN TEMA DETERMINADO (EDUCACIÓN, SALUD, MEDIO AMBIENTE, ECONOMÍA, ETC.).
- **EN PROMEDIO CADA PROYECTO ES GIRADO A 2 O 3 COMISIONES.**
- UNA VEZ QUE LAS COMISIONES A LAS QUE FUE GIRADO UN PROYECTO EMITEN UN DICTAMEN SOBRE EL MISMO, ÉSTE ESTÁ EN CONDICIONES DE SER TRATADO EN EL RECINTO DE LA CÁMARA.

JUICIOS POR JURADOS
POPULARES. CREACIÓN

SISTEMA DE INCLUSION
Y ACCESIBILIDAD A LOS
CAJEROS AUTOMATICOS
PARA DISCAPACITADOS
MOTRICES. REGIMEN

ASUNTOS CONSTITUCIONALES

LEGISLACION PENAL

PRESUPUESTO Y HACIENDA

FINANZAS

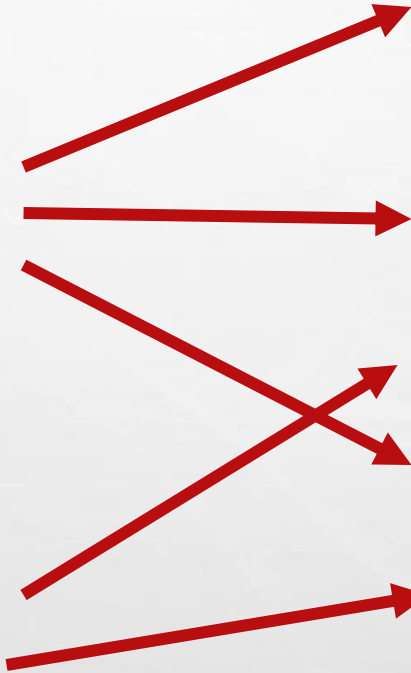
JUSTICIA

DISCAPACIDAD

FAMILIA

MEDIO AMBIENTE

INDUSTRIA



DESCRIPCIÓN DEL PROBLEMA

- EL DESAFÍO PROPUESTO ES HACER UN CLASIFICADOR QUE ENTRENE CON LOS PROYECTOS QUE YA FUERON GIRADOS A COMISIONES EN EL PASADO Y PUEDA SUGERIR A QUÉ COMISIONES DEBEN ASIGNARSE LOS NUEVOS PROYECTOS QUE INGRESAN.
- CONSIDERAMOS LAS 45 COMISIONES A LAS QUE PUEDE SER ASIGNADO CADA PROYECTO, COMO UNA ETIQUETA. Y DADO QUE CADA PROYECTO PUEDE ASIGNARSE A MÁS DE UNA COMISIÓN, ESTAMOS ANTE UN PROBLEMA DE CLASIFICACIÓN MULTI-LABEL.
- SCIKIT-LEARN PROVEE UNA LIBRERÍA ESPECÍFICA PARA PROBLEMAS DE CLASIFICACIÓN MULTI-LABEL: **SCIKIT-MULTILEARN**.
- EN ESTE DESAFÍO VAMOS A UTILIZAR CLASIFICADORES MULTILABEL Y CLASIFICADORES SINGLE-LABELS ADAPTANDO EL ESPACIO DE ETIQUETAS A ESTOS ÚLTIMOS
- SE VA A EVALUAR EL RESULTADO UTILIZANDO MÉTRICAS PROPIAS DEL PROBLEMA DE CLASIFICACIÓN MULTILABEL
- FINALMENTE VAMOS A TESTEAR EL COMPORTAMIENTO DE LOS MEJORES CLASIFICADORES OBTENIDOS, MEDIANTE UNA API WEB Y UNA APLICACIÓN CLIENTE

ANALISIS DE DATOS

DATASET ORIGINAL

TITULO	GIRO_INICIADORA
DECLARASE LA EMERGENCIA LABORAL EN EL "INSTITU...	LEGISLACION DEL TRABAJO;PRESUPUESTO Y HACIENDA
DENOMINASE A LA RUTA NACIONAL N° 5 COMO "RUTA ...	TRANSPORTES
INSTITUYESE EL 11 DE MAYO DE CADA AÑO COMO "DI...	LEGISLACION GENERAL;CULTURA
DECLARASE EL 30 DE NOVIEMBRE DE CADA AÑO COMO ...	LEGISLACION GENERAL;AGRICULTURA Y GANADERIA
ESTABLECESE CON CARACTER DE "FIESTA NACIONAL D...	CULTURA;TURISMO;LEGISLACION GENERAL
DECLARAR MONUMENTO NATURAL A LA ESPECIE "CHINC...	RECURSOS NATURALES Y CONSERVACION DEL AMBIENTE...
REGIMEN DE PROMOCION DE LA PRODUCCION Y/O ELAB...	AGRICULTURA Y GANADERIA;PRESUPUESTO Y HACIENDA
PROHIBICION DE USO Y DISTRIBUCION DE ARTICULOS...	SEGURIDAD INTERIOR;INDUSTRIA
COMERCIALIZACION DE SUPLEMENTOS DIETARIOS. REG...	ACCION SOCIAL Y SALUD PUBLICA;COMERCIO
INSTITUYESE EL 26 DE MAYO DE CADA AÑO COMO "DI...	LEGISLACION GENERAL;ACCION SOCIAL Y SALUD PUBLICA
COMISION BICAMERAL DE AMBIENTE - CBA -. CREACI...	PETICIONES, PODERES Y REGLAMENTO;PRESUPUESTO Y...
DECLARASE "FIESTA NACIONAL DE LA TRADICION GAU...	LEGISLACION GENERAL;TURISMO;CULTURA
CONTRATO DE TRABAJO - LEY 20744 -. MODIFICACIO...	LEGISLACION DEL TRABAJO;DISCAPACIDAD
DECLARASE DE INTERES NACIONAL LA PROMOCION DE ...	FAMILIA, MUJER, NIÑEZ Y ADOLESCENCIA;ACCION SO...
RED NACIONAL DE BANCOS DE LECHE HUMANA Y REGIS...	ACCION SOCIAL Y SALUD PUBLICA;FAMILIA, MUJER, ...

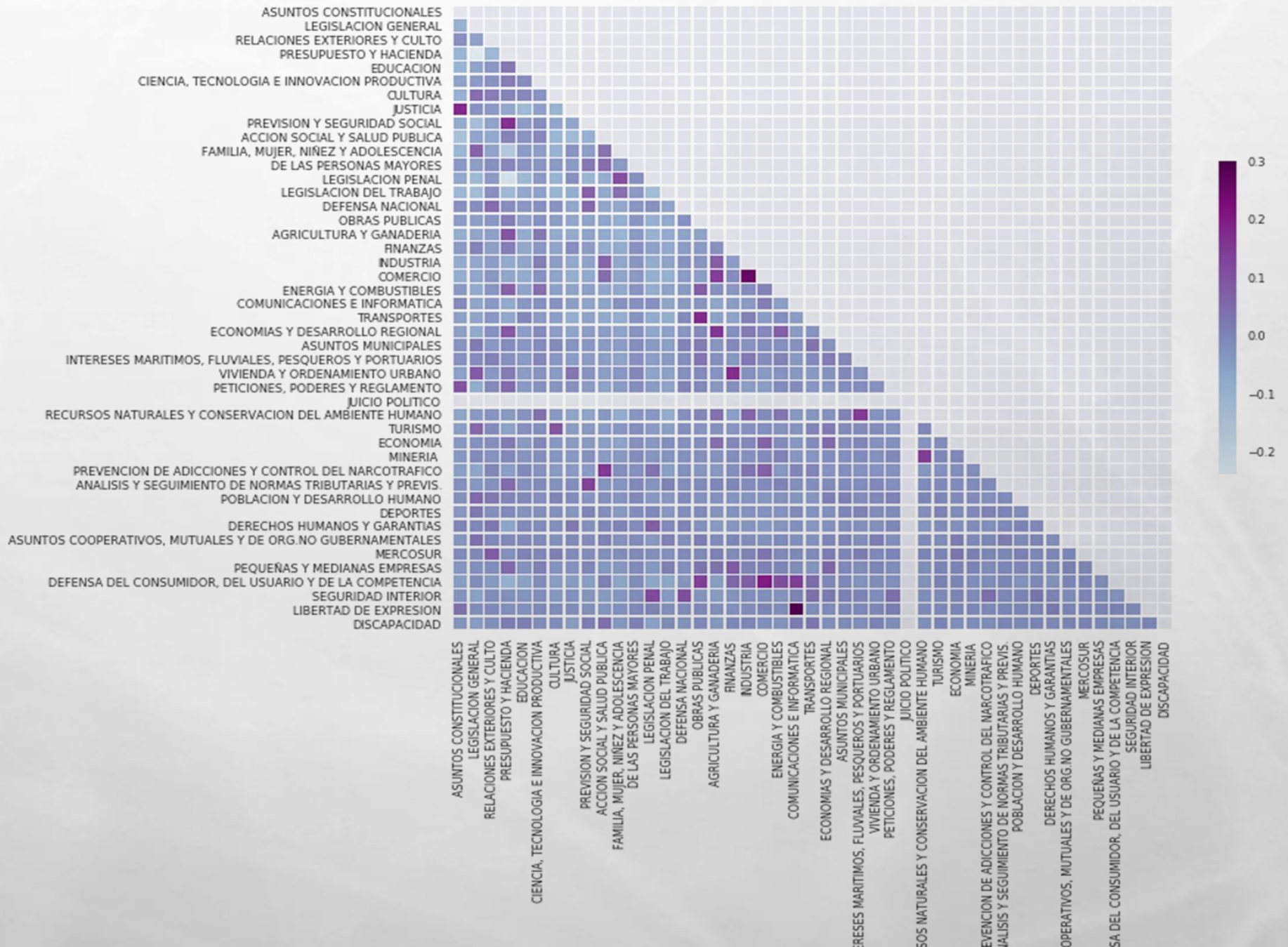
textos

comisiones (labels)

FUENTE: www.hcdn.gob.ar/datos.hcdn.gob.ar

Correlación entre las etiquetas

La correlación entre etiquetas indica que puede ser mejor elegir un clasificador que además de basarse en la relevancia de los features tenga en cuenta la correlación entre los targets



Desbalance entre las combinaciones de etiquetas



PREPARACIÓN DE LOS DATOS

PREPARACION DE LOS TEXTOS

1. NORMALIZACION DE MAYUSCULAS Y ACENTOS
2. ELIMINACION DE STOP WORDS
3. TOKENIZACION
4. STEMMING
5. VECTORIZACIÓN

PREPARACION DE LOS TEXTOS

TOKENIZER/STEMMING

PREVENCION DE LOS DEFECTOS DEL TUBO NEURAL A PARTIR DE GARANTIZAR LA PROVISION GRATUITA DE ACIDO FOLICO.

1 REGIMEN.

EXPRESAR BENEPLACITO POR LA PARTICIPACION Y EL SUBCAMPEONATO QUE OBTUVO LA SELECCION NACIONAL DE FUTBOL PARA CIEGOS "LOS MURCIELAGOS" EN EL MUNDIAL 2018 REALIZADO EN EL MUNDIAL 2018 REALIZADO EN LA CIUDAD DE MADRID, ESPAÑA.

EXPRESAR BENEPLACITO POR EL CENTESIMO QUINTO ANIVERSARIO DE LOS OLIVOS TENIS CLUB UBICADO EN LA LOCALIDAD DE VICENTE LOPEZ, PROVINCIA DE BUENOS AIRES, A CELEBRARSE EL 25 DE OCTUBRE DE 2018.

RECURSOS PARA ATENDER EL PROGRAMA DE REPARACION HISTORICA PARA JUBILADOS Y PENSIONADOS Y EL SISTEMA INTEGRADO PREVISIONAL ARGENTINO - SIPA -. MODIFICACIONES DE LAS LEYES 27260 Y 26425.

MINISTERIOS - LEY 22520 -. INCORPORACION DEL ARTICULO 8 BIS, ESTABLECIENDO QUE LOS MINISTROS O EL JEFE DE GABINETE SALIENTE, DEBEN CONCURRIR EN UN PLAZO DE 10 DIAS ANTE LAS CAMARAS DEL H. CONGRESO DE LA NACION.

INICIATIVA POPULAR. REGLAMENTACION

0	regim	folic	acid	gratuit	provision	garantiz	part	neural	tub
1	espa	madr	ciud	juni	realiz	2018	mundial	murcielag	cieg
2	octubr	celebr	air	buen	provinci	lopez	vicent	local	ubic
3	26425	27260	ley	modif	sip	argentin	previsional	integr	sistem
4	nacion	congres	h	cam	dias	plaz	concurr	deb	salient
5	24747	derogacion	constitucion	reglamentac	popul	inici	articul	ley	naciona
6	conex	cuestion	26093	cre	biocombust	sustent	uso	produccion	mezcl
7	ucar	rural	cambi	unid	eliminacion	decret	efect	dej	dispong

COUNT VECTORIZER

DOC	WORD	WEIGHT
0	19348	0.21087651429112053
0	9091	0.38067157993316886
0	23401	0.3728456674995482
0	17206	0.39648422474229345
0	18212	0.1719145872064076
0	12217	0.2513436609886885
0	19644	0.266300864546739
0	12663	0.26092061444721093
0	3532	0.349530753447365
0	11837	0.3663307618964071
0	20349	0.1622117698456229

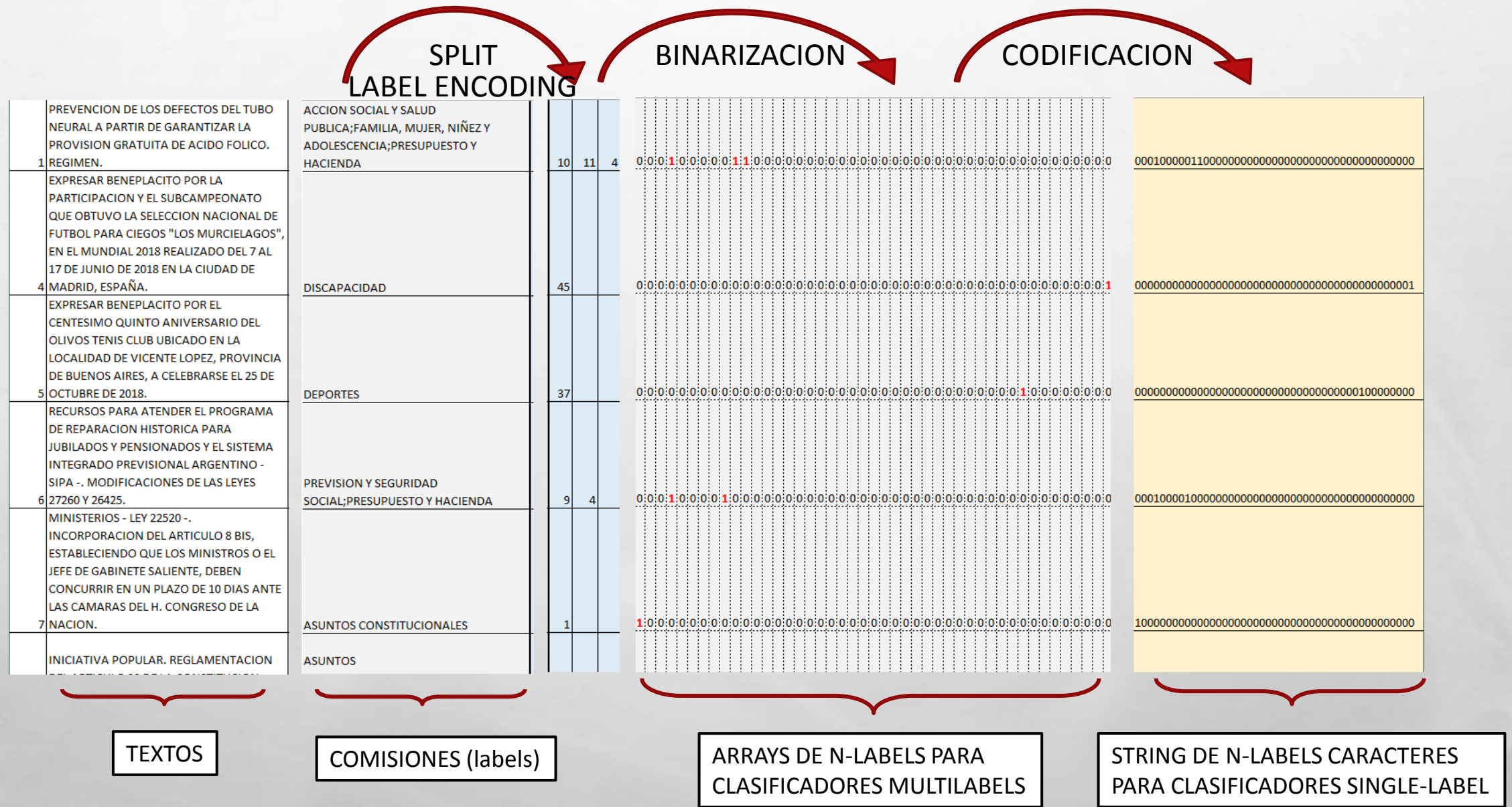
TFIDF TRANSFORMER

19348	prevencion
9091	defect
23401	tub
17206	neural
18212	part
12217	garantiz
19644	provision
12663	gratuit
3532	acid
11837	folic
20349	regim

PREPARACIÓN DE LAS ETIQUETAS

1. SPLIT
2. LABEL ENCODING
3. CONVERSIÓN A MATRIZ DE 0/1 (N-FEATURES, N-LABELS) (PARA UTILIZACIÓN DE CLASIFICADORES CON CAPACIDAD MULTILABEL)
4. CONVERSIÓN A CÓDIGO DE 0/1 (STRING) (PARA UTILIZACIÓN DE CLASIFICADORES SINGLE-LABEL)

PREPARACIÓN DE LAS ETIQUETAS



CLASIFICADORES

CLASIFICADORES A PROBAR

- Modelos de clasificación que toman el espacio de etiquetas y aplican un clasificador de base (SVM, Logistic Regression, Naive Bayes, etc), a cada etiqueta:
 - CHAIN CLASSIFIER
 - BINARY RELEVANCE
 - LABEL POWERSSET
- Modelos de clasificación adaptados para abordar el problema de la clasificación multilabel:
 - MLKNN
- Modelos de clasificación single-label, utilizando cada combinación de etiquetas como una etiqueta distinta:
 - SGD CLASSIFIER
 - LOGISTIC REGRESSION
 - MULTINOMIAL NAIVE BAYES

CLASIFICADORES MULTILABEL

BINOMIAL RELEVANCE

- ESTE CLASIFICADOR SE UBICA DENTRO DE LA CATEGORÍA DE CLASIFICADORES QUE **TRANSFORMAN EL PROBLEMA DE MULTI-LABEL A SINGLE-LABEL**. BÁSICAMENTE ASIGNA CADA ETIQUETA A CADA REGISTRO COMO SI SE TRATARA DE UN PROBLEMA DE UNA SOLA ETIQUETA, SEPARANDO EL PROBLEMA EN TANTOS PROBLEMAS COMO ETIQUETAS HAYA.
- DESPUÉS DE SEPARAR EL PROBLEMA, EL CLASIFICADOR BR PUEDE UTILIZAR CUALQUIER OTRO CLASIFICADOR "SINGLE-LABEL" QUE SE LE CONFIGURE PARA ASIGNAR CADA ETIQUETA.
- LA DESVENTAJA DE ESTE CLASIFICADOR ES QUE NO TOMA EN CUENTA LA CORRELACIÓN ENTRE ETIQUETAS.

\mathbf{x}	Y_1	\mathbf{x}	Y_2	\mathbf{x}	Y_3	\mathbf{x}	Y_4
$\mathbf{x}^{(1)}$	0	$\mathbf{x}^{(1)}$	1	$\mathbf{x}^{(1)}$	1	$\mathbf{x}^{(1)}$	0
$\mathbf{x}^{(2)}$	1	$\mathbf{x}^{(2)}$	0	$\mathbf{x}^{(2)}$	0	$\mathbf{x}^{(2)}$	0
$\mathbf{x}^{(3)}$	0	$\mathbf{x}^{(3)}$	1	$\mathbf{x}^{(3)}$	0	$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	1	$\mathbf{x}^{(4)}$	0	$\mathbf{x}^{(4)}$	0	$\mathbf{x}^{(4)}$	1
$\mathbf{x}^{(5)}$	0	$\mathbf{x}^{(5)}$	0	$\mathbf{x}^{(5)}$	0	$\mathbf{x}^{(5)}$	1

CLASSIFIER CHAIN

- ESTE CLASIFICADOR TAMBIÉN SE UBICA DENTRO DE LA CATEGORÍA DE CLASIFICADORES QUE **TRANSFORMAN EL PROBLEMA DE MULTI-LABEL A SINGLE-LABEL**.
- PERO EN LUGAR DE ASIGNAR CADA ETIQUETA INDIVIDUALMENTE, ASIGNA CADA ETIQUETA TENIENDO EN CUENTA NO SOLO LAS VARIABLES INDEPENDIENTES SINO LAS ETIQUETAS QUE YA ASIGNÓ, POR LO QUE PUEDE FUNCIONAR MEJOR QUE EL ANTERIOR CUANDO HAY UNA FUERTE CORRELACIÓN ENTRE LAS ETIQUETAS.
- PARA CADA ESLABÓN DE LA CADENA UTILIZA EL CLASIFICADOR SINGLE-LABEL QUE SE LE CONFIGURE

X	y1
x1	0
x2	1
x3	0

X	y1	y2
x1	0	1
x2	1	0
x3	0	1

X	y1	y2	y3
x1	0	1	1
x2	1	0	0
x3	0	1	0

X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0

LABEL POWERSET

- ESTE TERCER CLASIFICADOR TRANSFORMA EL PROBLEMA DE "**MULTI-LABEL**" A "**MULTI-CLASS SINGLE-LABEL**", YA QUE CONVIERTE CADA COMBINACIÓN EXISTENTE DE ETIQUETAS, EN UN ÚNICO VALOR DE UNA NUEVA Y ÚNICA ETIQUETA, QUE VA A SER LA QUE VA A UTILIZAR PARA RESOLVER EL PROBLEMA.
- PARA CADA ESLABÓN DE LA CADENA UTILIZA EL CLASIFICADOR SINGLE LABEL QUE SE LE CONFIGURE

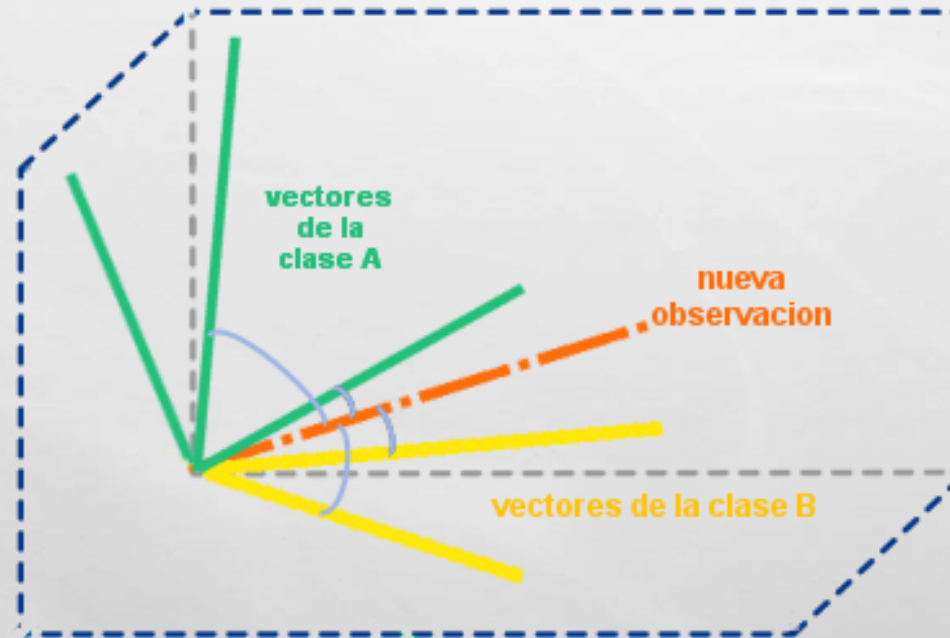
X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0
x4	0	1	1	0
x5	1	1	1	1
x6	0	1	0	0



X	y1
x1	1
x2	2
x3	3
x4	1
x5	4
x6	3

MLkNN

- A DIFERENCIA DE LOS ANTERIORES ESTE CLASIFICADOR NO REQUIERE QUE SE LE CONFIGURE UN CLASIFICADOR BASE, YA QUE ES UNA ADAPTACIÓN DEL CLASIFICADOR KNN AL PROBLEMA DE LA CLASIFICACIÓN MULTILABEL.



CLASIFICADORES SINGLE LABEL

CLASIFICADORES QUE SOLO PUEDEN ASIGNAR UNA ETIQUETA POR OBSERVACIÓN. SE UTILIZAN EN EL CASO EN ESTUDIO EMPLEANDO CADA COMBINACIÓN DE ETIQUETAS COMO UNA ÚNICA ETIQUETA. (SE UTILIZAN SIN EMBARGO LOS ARRAYS DE ETIQUETAS PARA MEDIR LOS RESULTADOS):

- LOGISTIC REGRESSION
- MULTINOMIAL NAIVE BAYES
- SGD CLASSIFIER

METRICAS

MÉTRICAS

LAS MÉTRICAS DE EVALUACIÓN DE DESEMPEÑO DE LOS CLASIFICADORES MULTI-LABEL SON DIFERENTES DE LAS UTILIZADAS PARA LOS CLASIFICADORES SINGLE-LABEL, YA QUE ADEMÁS DEL CONCEPTO DE ACIERTO Y ERROR DEBE TENERSE EN CUENTA EL CONCEPTO DE ACIERTO PARCIAL.

LAS MÉTRICAS A UTILIZAR SERÁN:

- HAMMING LOSS:
- PRECISION, RECALL Y F-SCORE
- JACCARD SIMILARITY
- EXACT MATCH (TAMBIÉN LLAMADO SUBSET ACCURACY)

MÉTRICAS

JACCARD SIMILARITY: MIDE LA SIMILITUD ENTRE EL SET DE ETIQUETAS PREDICHAS Y EL SET DE ETIQUETAS REALES, DIVIDIENDO EL TAMAÑO DE LA INTERSECCIÓN ENTRE LAS ETIQUETAS PREDICHAS Y LAS ETIQUETAS VERDADERAS POR EL TAMAÑO DE LA UNIÓN DE AMBAS

$$\text{Jaccard Similarity} = \frac{1}{N} \sum_{k=1}^N \left(\frac{\|Y_k \cap Z_k\|}{\|Y_k \cup Z_k\|} \right)$$

ACCURACY SCORE: EN PROBLEMAS DE CLASIFICACIÓN MULTILABEL, ESTA MÉTRICA SE DENOMINA TAMBIÉN “EXACT MATCH” O “SUBSET ACCURACY”, Y ES LA MÉTRICA MÁS ESTRICTA, YA QUE DEVUELVE EL PORCENTAJE DE OBSERVACIONES QUE OBTUVIERON TODAS SUS ETIQUETAS CORRECTAMENTE CLASIFICADAS

$$\text{Subset-Accuracy} = \frac{1}{N} \sum_{k=1}^N \mathbf{1}(Y_k = Z_k)$$

HAMMING LOSS: FUNCIÓN DE PÉRDIDA QUE INDICA LA PROPORCIÓN DE ERRORES EN LA CLASIFICACIÓN DE LOS DOCUMENTOS Y SE COMPUTA COMO LA DIFERENCIA SIMÉTRICA ENTRE LAS CATEGORÍAS PREDICHAS Y LAS VERDADERAS, SOBRE EL TOTAL DE CATEGORÍAS EXISTENTES Y DE OBSERVACIONES.

$$\text{Hamming-Loss} = \frac{1}{N} \sum_{k=1}^N \left(\frac{\|Y_k \cup Z_k\| - \|Y_k \cap Z_k\|}{M} \right)$$

Y = etiquetas predichas

Z = etiquetas verdaderas

N = cantidad de observaciones

M = cantidad de etiquetas

COMPARACION DE RESULTADOS

CLASSIFICADORES MULTI-LABEL (CLASIFICADOR BASE: LOGISTIC REGRESSION) *			
	CLASSIFIER CHAIN	BINARY RELEVANCE	LABEL POWERSET
ACCURACY SCORE	0.6045	0.5635	0.6638
JACCARD SIMILARITY	0.7102	0.6967	0.7504
HAMMING LOSS	0.0182	0.0172	0.0164

CLASSIFICADOR MLkNN (kNN MULTI-LABEL)**	
ACCURACY SCORE	0.6081
JACCARD SIMILARITY	0.7157
HAMMING LOSS	0.0175

CLASSIFICADORES SINGLE-LABEL *			
	LOGISTIC REGRESSION	SGD CLASSIFIER	MULTINOMIAL NAIVE BAYES
ACCURACY SCORE	0.6782	0.6139	0.6456
JACCARD SIMILARITY	0.7595	0.6956	0.7292
HAMMING LOSS	0.0122	0.0151	0.0139

* Con 13 mil registros aleatorios

* Con 11 mil registros aleatorios

APLICACIÓN DE PRUEBA

WEB SERVICE



2) SELECCIONAR
UN
CLASIFICADOR

1) INPUT
Sumario del
proyecto

Clasificador de proyecto x 127.0.0.1:8080/lr/classify?te x +

127.0.0.1:8080/lr/classify?texto=MODIFICACION DEL CODIGO PENAL INCREMENTANDO LA PENA DEL DELITO DE DEFR 170%

JSON Raw Data Headers

Save Copy Filter JSON

```
giro_propuesto:
  0: "LEGISLACION PENAL"
otras comisiones: []
similares:
  0:
    0: 1.0007711382152524
    1:
      0: "codigo"
      1: "penal"
      2: "modificacion"
      3: "defraudacion"
      4: "delito"
    2: "MODIFICACION DEL Art. 173 DEL CODIGO PENAL, SOBRE DELITO DE DEFRAUDACION"
    3:
      0: "LEGISLACION PENAL"
      4: "1177-D-2003"
    1: [...]
    2: [...]
```

3) OUTPUT
Comisiones
propuestas

4) OUTPUT
Proyectos
similares

APLICACIÓN CLIENTE



1) INPUT
Sumario del proyecto

2) SELECCIONAR UN CLASIFICADOR

SUMARIO DEL PROYECTO

CONVOCATORIA A CONSULTA POPULAR VINCULANTE POR INTERRUPCION VOLUNTARIA DEL EMBARAZO.

GIRO PROPUESTO:

ASUNTOS CONSTITUCIONALES
PRESUPUESTO Y HACIENDA

Clasificador SGD

Regresión logística

Borrar

Distancia	Texto	comisiones	expediente
0.88	CONVOCATORIA A UNA CONSULTA POPULAR.	ASUNTOS CONSTITUCIONALES,PRESUPUESTO Y HACIENDA	0881-D-1999
0.91	CONSULTA POPULAR - LEY 25432 -. MODIFICACION DE LOS ARTICULOS 1° Y 6°, SOBRE CONSULTA POPULAR VINCULANTE Y NO VINCULANTE.	ASUNTOS CONSTITUCIONALES	6167-D-2015
0.91	CONSULTA POPULAR - LEY 25432 -. MODIFICACION DE LOS ARTICULOS 1° Y 6°, SOBRE CONSULTA POPULAR VINCULANTE Y NO VINCULANTE.	ASUNTOS CONSTITUCIONALES	9056-D-2014
0.91	CONSULTA POPULAR - LEY 25432 -. MODIFICACION DE LOS ARTICULOS 1° Y 6°, SOBRE CONSULTA POPULAR VINCULANTE Y NO VINCULANTE.	ASUNTOS CONSTITUCIONALES	1053-D-2018

3) OUTPUT
Comisiones propuestas

4) OUTPUT
Proyectos similares

CONVOCATORIA A CONSULTA POPULAR VINCULANTE POR INTERRUPCION VOLUNTARIA DEL EMBARAZO.

GIRO PROPUESTO:

ASUNTOS CONSTITUCIONALES
PRESUPUESTO Y HACIENDA

Clasificador SGD

Regresión logística

Borrar

Distancia	Texto	comisiones	expediente
0.88	CONVOCATORIA A UNA CONSULTA POPULAR.	ASUNTOS CONSTITUCIONALES,PRESUPUESTO Y HACIENDA	0881-D-1999
0.91	CONSULTA POPULAR - LEY 25432 -. MODIFICACION DE LOS ARTICULOS 1° Y 6°, SOBRE CONSULTA POPULAR VINCULANTE Y NO VINCULANTE.	ASUNTOS CONSTITUCIONALES	6167-D-2015
0.91	CONSULTA POPULAR - LEY 25432 -. MODIFICACION DE LOS ARTICULOS 1° Y 6°, SOBRE CONSULTA POPULAR VINCULANTE Y NO VINCULANTE.	ASUNTOS CONSTITUCIONALES	9056-D-2014
0.91	CONSULTA POPULAR - LEY 25432 -. MODIFICACION DE LOS ARTICULOS 1° Y 6°, SOBRE CONSULTA POPULAR VINCULANTE Y NO VINCULANTE.	ASUNTOS CONSTITUCIONALES	1053-D-2018

MUCHAS GRACIAS!