El futuro es de todos

DNP
Departamento
Nacional de Planeación

El futuro es de todos

DNP
Departamento Nacional de Planeación

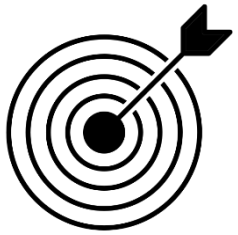
Clasificación de documentos del Diario Oficial

Unidad de Científicos de Datos
Dirección de Desarrollo Digital

Junio, 2019

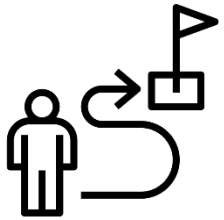
Descripción del proyecto

Es necesario cuantificar las regulaciones impuestas para cada uno de los sectores productivos del país



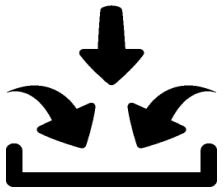
Objetivo

Implementar técnicas de procesamiento de texto para clasificar normativas dentro de 9 sectores productivos.



Metodología

1. Procesamiento y limpieza de los documentos
2. Distinguir el tipo de palabras asociadas entre sectores
3. Proponer modelos de clasificación

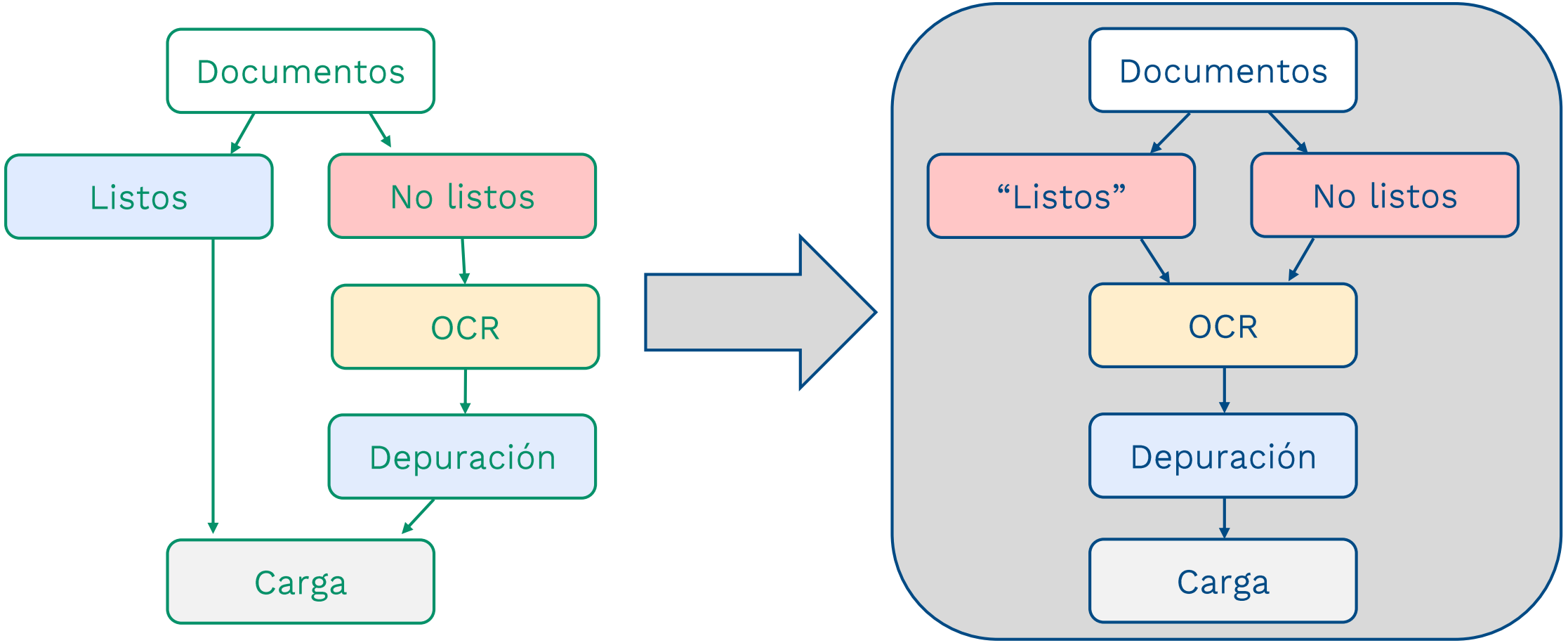


Insumos

386 documentos compuestos por: Leyes, reformas y decretos, con su correspondiente clasificación (etiqueta).

Carga de los documentos

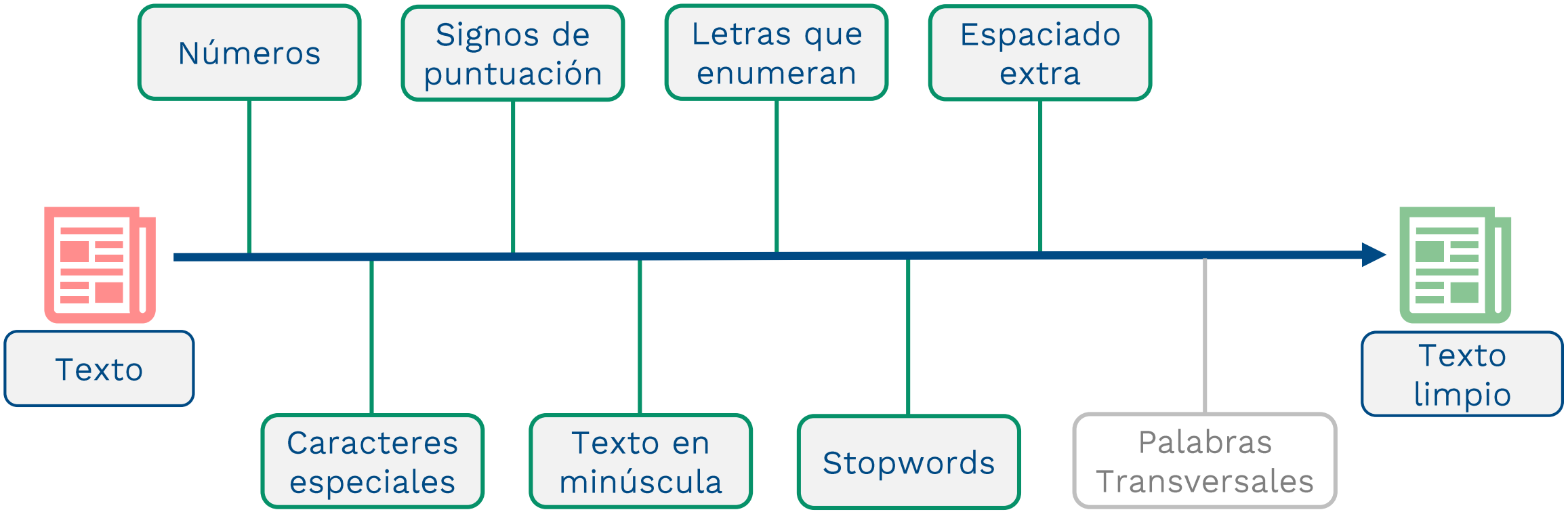
Uso de un algoritmo de OCR (*tesseract*)



OCR: Reconocimiento óptico de caracteres

Limpieza del texto

Se prepararon los documentos para aumentar la calidad de los insumos



Identificación de palabras

Se identificaron las palabras más frecuentes para cada sector

Sectores

1. Agricultura
2. Minería
3. Manufacturas
4. Electricidad
5. Construcción
6. Comercio
7. Transporte
8. Finanzas
9. Servicios

Agricultura

Identificación de palabras

Para el sector de Manufacturas, se identificaron las palabras más frecuentes por subsector

Manufacturas

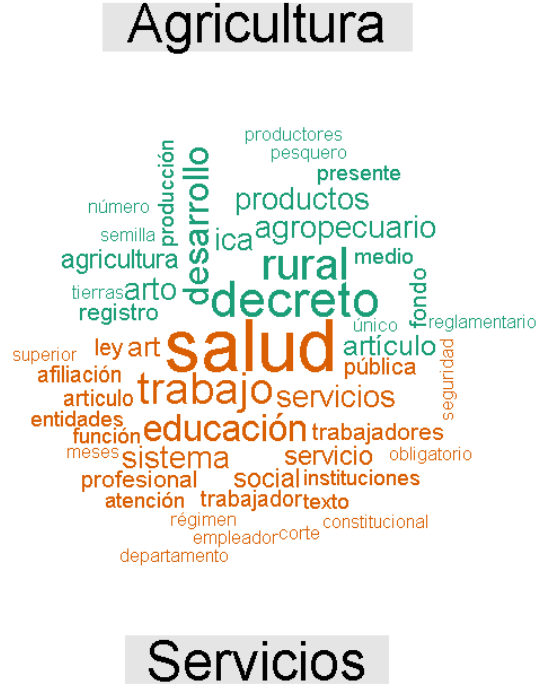
1. Derivados petróleo
2. Madera y corcho
3. Papel e imprenta
4. Productos de metal
5. Minerales no metálicos
6. Productos ABC
7. Textiles y cuero
8. Generales

Derivados de petróleo

Identificación de palabras

Comparación de palabras frecuentes entre normativas (sector-sector ó subsector-subsector)

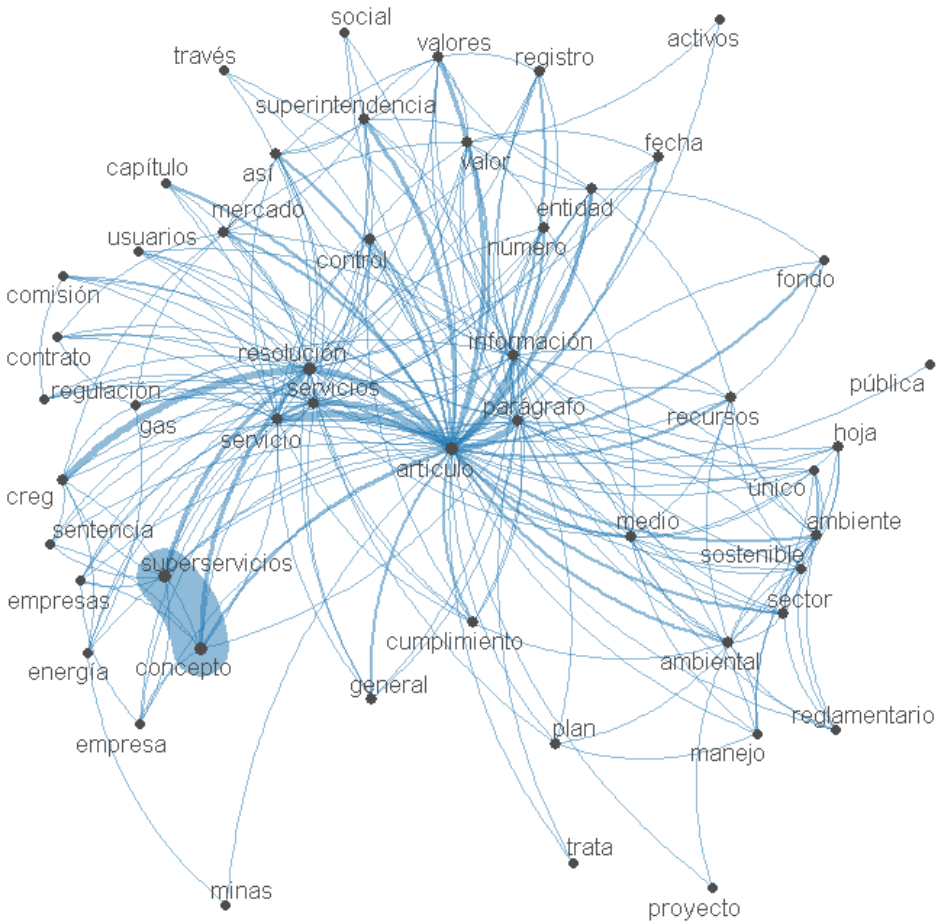
Palabras en común



Palabras discriminantes

Identificación de *stopwords*

Conformación de red de conexiones entre palabras para identificar aquellas que son transversales



Crear una lista de palabras no necesarias:

ej: decreto, artículo, parágrafo, resolución

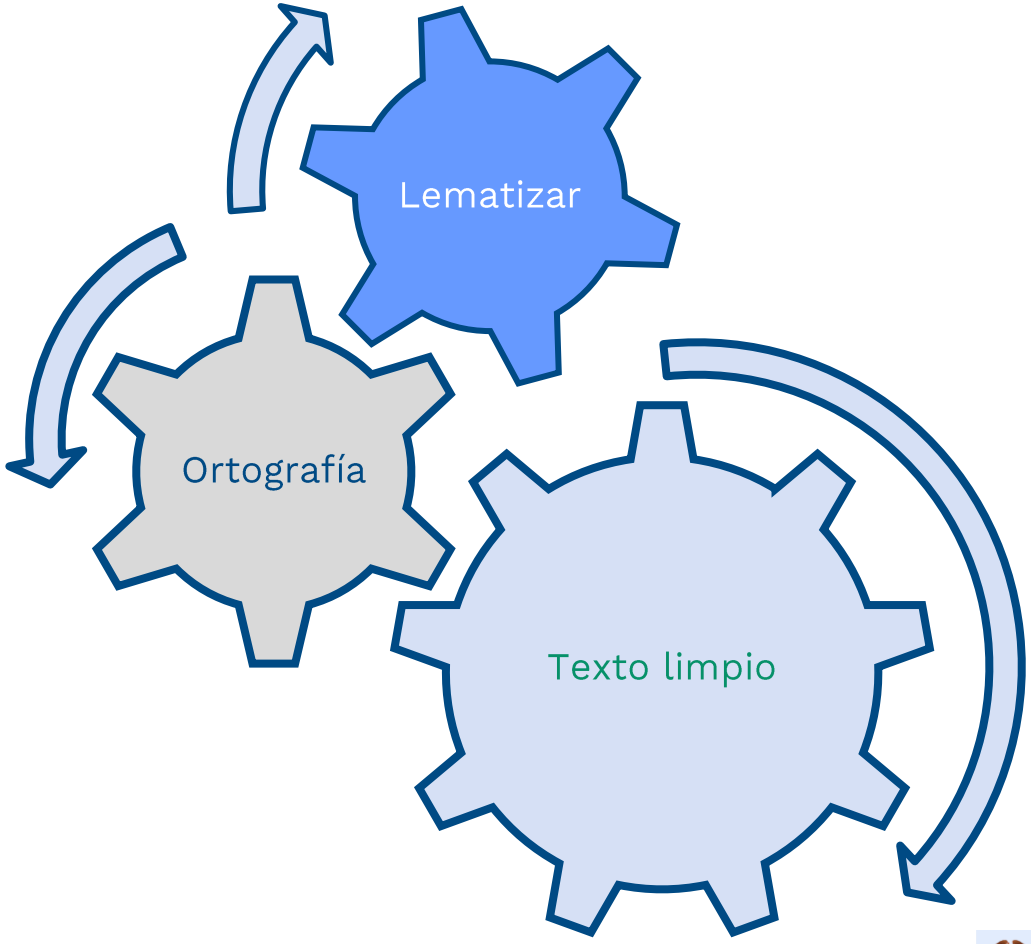
Mejora de los insumos

Se está trabajando en mejorar la calidad de los insumos

Mejor conteo de las palabras frecuentes en los sectores

Ej: “hablamos y habló el hablador”
> hablar y hablar el hablar

Corregir errores de ortografía o del OCR
Ej: “anbiental”, “alimenfo”
> ambiental, alimento



Vectorización

Representación numérica del texto contenido en los documentos

Doc 1

Doc 2

Doc 386

	X0.	...	X99	sector_economico
1	3.6112919	...	-0.1030051	1
2	1.1316650	...	0.6562225	1
386	-0.4359001	...	0.1919476	9

Primeros modelos

Se empezó la clasificación con dos tipos de modelos supervisados

Nombre	Precisión
Gradient Boosting Model	55,88%
Super Vector Machine	61,76%

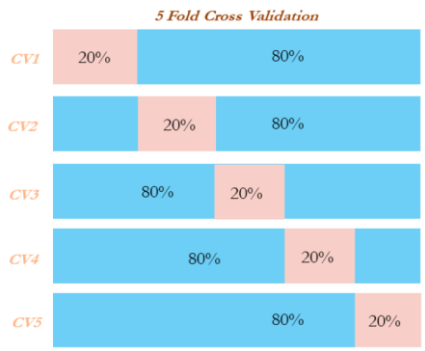
Búsqueda del modelo

Diferentes algoritmos para modelos supervisados

Modelo	Precisión
Gradient Boosting Model	55,88%
Naive Bayes	36,24%
Neural Network	45,76%
Random Trees	54,12%
Extreme Gradient Boosting	63,20%
Super Vector Machine	61,74%
★ Random Forest	65,88%
Lineal Discriminant Analysis	52,37%

Hiperparámetros

Conjunto de parámetros que exploran las combinaciones posibles en la configuración inicial del modelo



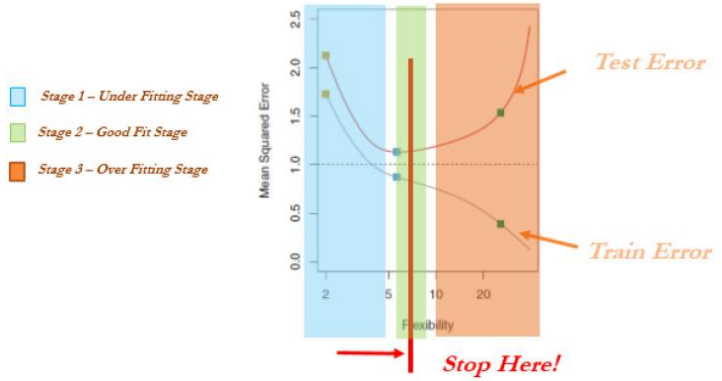
Validación cruzada

17 repeticiones

sector_economico	n
1	56
2	23
3	108
4	78
5	22
6	14
7	36
8	23
9	26

Balanceo de clases

Upsampling



Aprender-Memorizar

70% entrenamiento – 30% prueba

Objetivo: Obtener un mejor rendimiento en la clasificación

Resultados de clasificación

Conjunto de prueba con el algoritmo Random Forest

Funciones objetivo:

- **Precision (P):** TP/(TP+FP)
- **Recall (R)/sensitivity** TP/(TP+FN)

88,23% de precisión
91,83% de sensibilidad

Prediction \ Truth	1	2	3	4	5	6	7	8	9
1	12	0	4	0	0	0	0	0	1
2	0	2	0	0	1	0	0	0	0
3	4	4	28	0	0	1	1	0	1
4	0	0	0	23	1	1	0	1	0
5	0	0	0	0	4	0	0	0	0
6	0	0	0	0	0	2	0	0	0
7	0	0	0	0	0	0	9	0	0
8	0	0	0	0	0	0	0	5	0
9	0	0	0	0	0	0	0	0	5

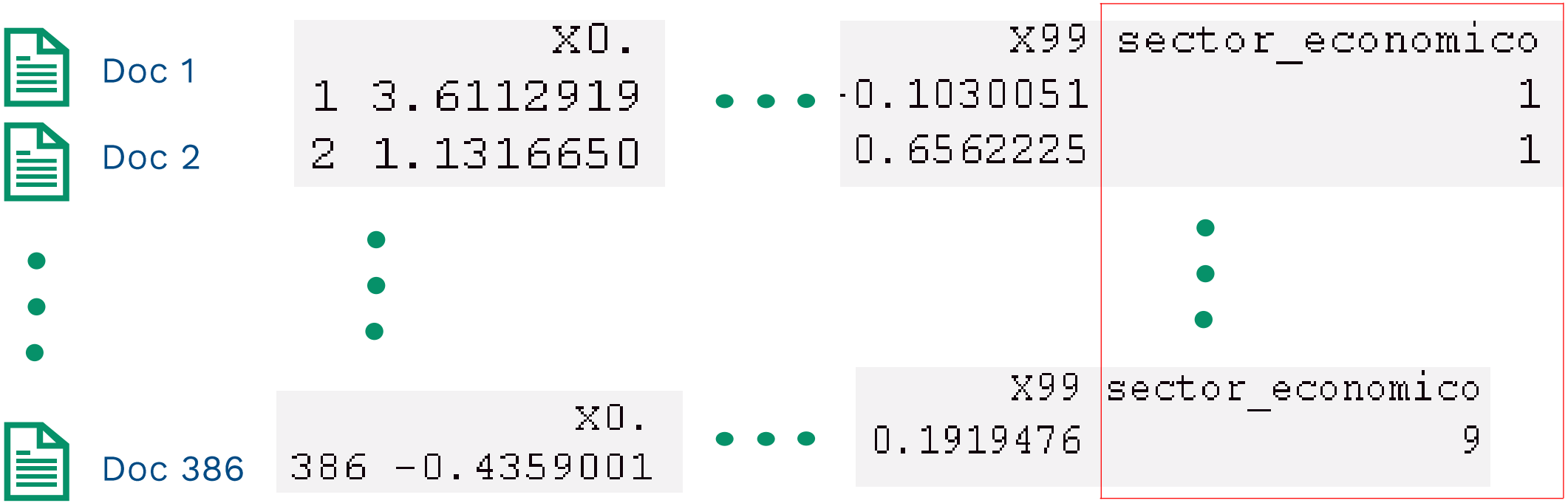
Sectores

1. Agricultura
2. Minería
3. Manufacturas
4. Electricidad
5. Construcción
6. Comercio
7. Transporte
8. Finanzas
9. Servicios

Falso Positivo: Es clasificado en el sector cuando en realidad no pertenece a él. (12)
Falso Negativo: No se clasifica en el sector cuando en realidad sí pertenece a él. (8)

Vectorización

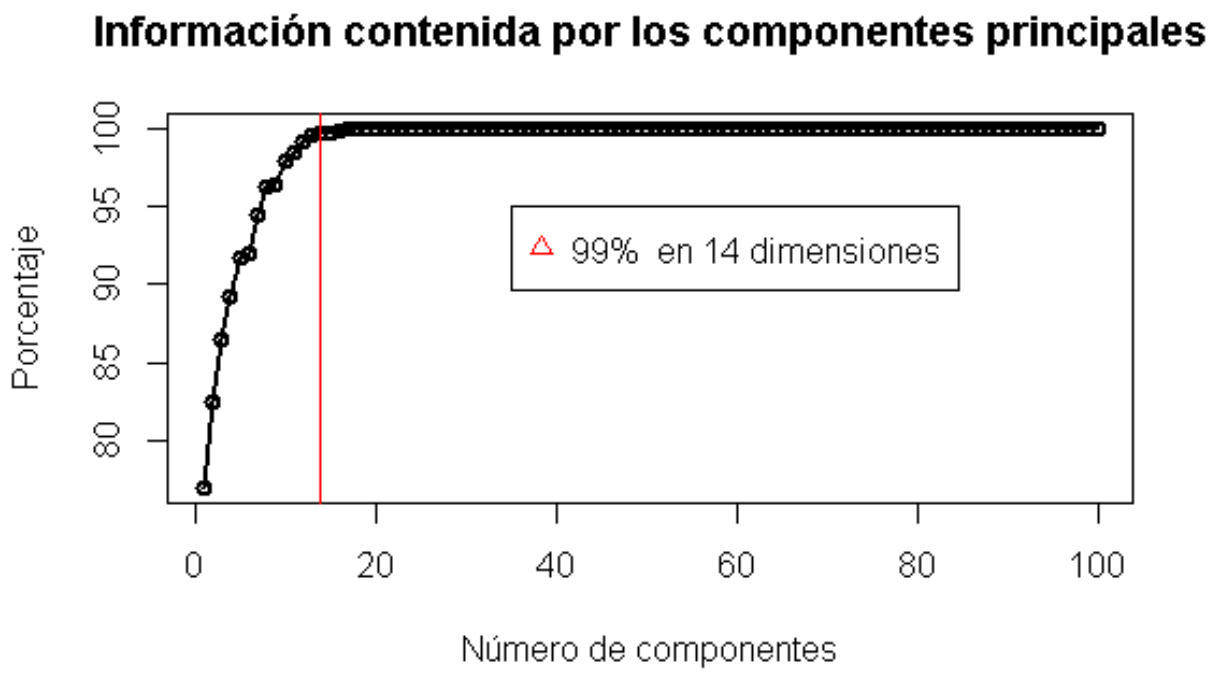
Representación numérica del texto contenido en los documentos



Algoritmo: Doc2Vec
Programa: Python

Componentes principales

Explora el número de dimensiones necesarias para reunir la información a partir del proceso de vectorización con el algoritmo Doc2Vec



Nuevos escenarios:

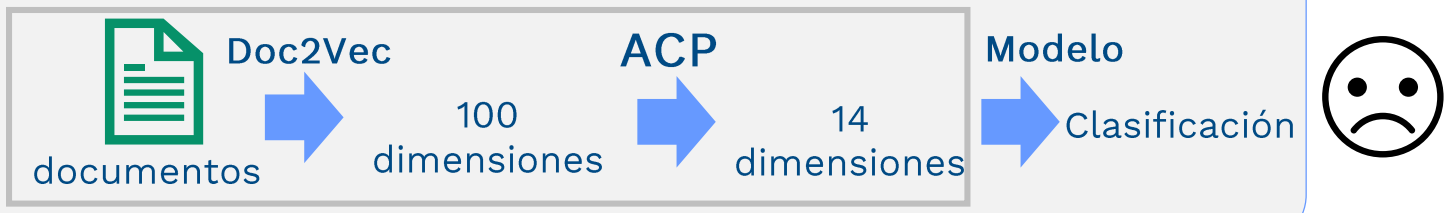
1. Resumir el 99% de las proyecciones hechas por la vectorización en 14 componentes y luego clasificar

2. Hacer una proyección en el proceso de vectorización con menos dimensiones

Escenarios

Escenarios:

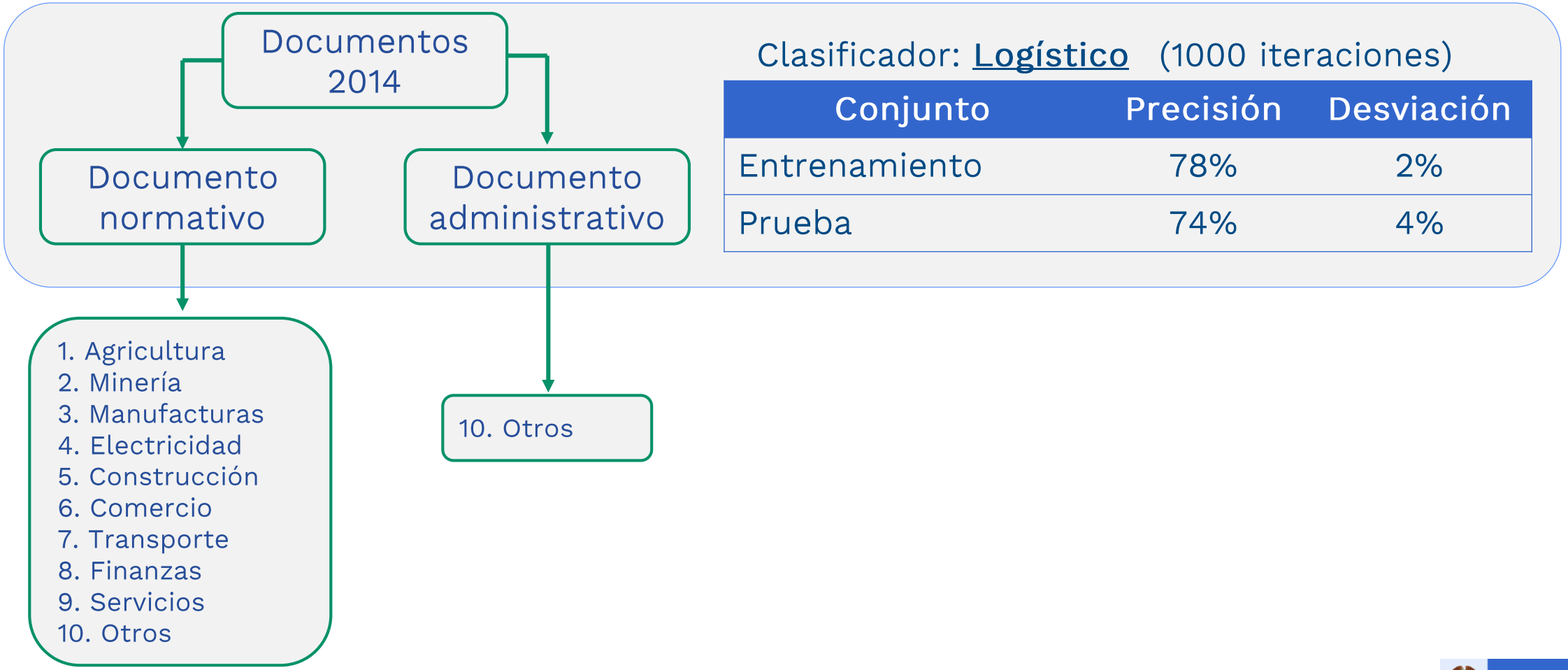
1. Resumir el 99% de las proyecciones hechas por la vectorización en 14 componentes y luego clasificar



2. Hacer una proyección en el proceso de vectorización con menos dimensiones

Diagrama del proceso

Resultados obtenidos para el clasificador sustancial



Clasificador: **Logístico** (1000 iteraciones)

Conjunto	Precisión	Desviación
Entrenamiento	78%	2%
Prueba	74%	4%

El futuro es de todos

DNP
Departamento
Nacional de Planeación