El futuro es de todos

DNP
Departamento
Nacional de Planeación

El futuro es de todos

DNP
Departamento
Nacional de Planeación

Análisis de los PND a través del tiempo utilizando minería de texto

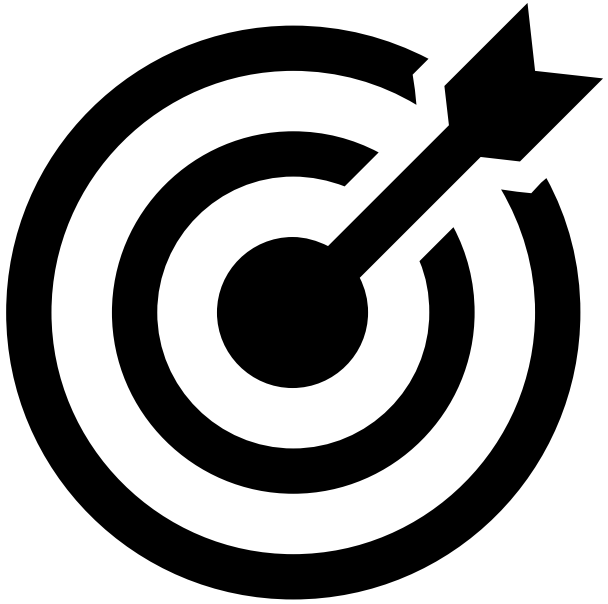
Unidad de Científicos de Datos
Dirección de Desarrollo Digital

Octubre, 2019

1. Descripción del proyecto
2. Determinación de palabras claves por sector
3. Estudio de PND por sector: 1990 – 2014
4. Estudio de términos más relevantes por PND: 1970 - 2018

1. Descripción del proyecto

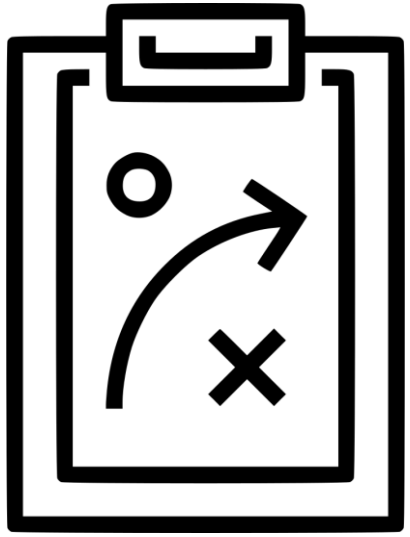
Objetivo



Realizar un análisis de la evolución del **contenido** y las **temáticas** más relevantes tratadas en los PND a través del tiempo.

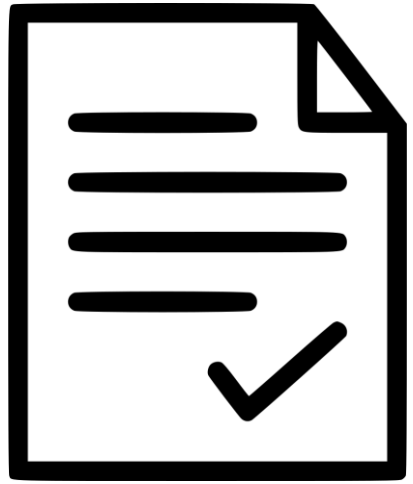
¿Qué información se utilizó?

1



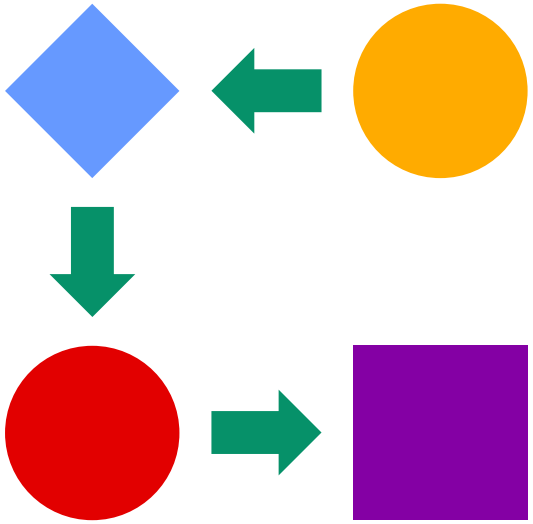
Planes Nacionales
de Desarrollo
(1970 a 2018)

2



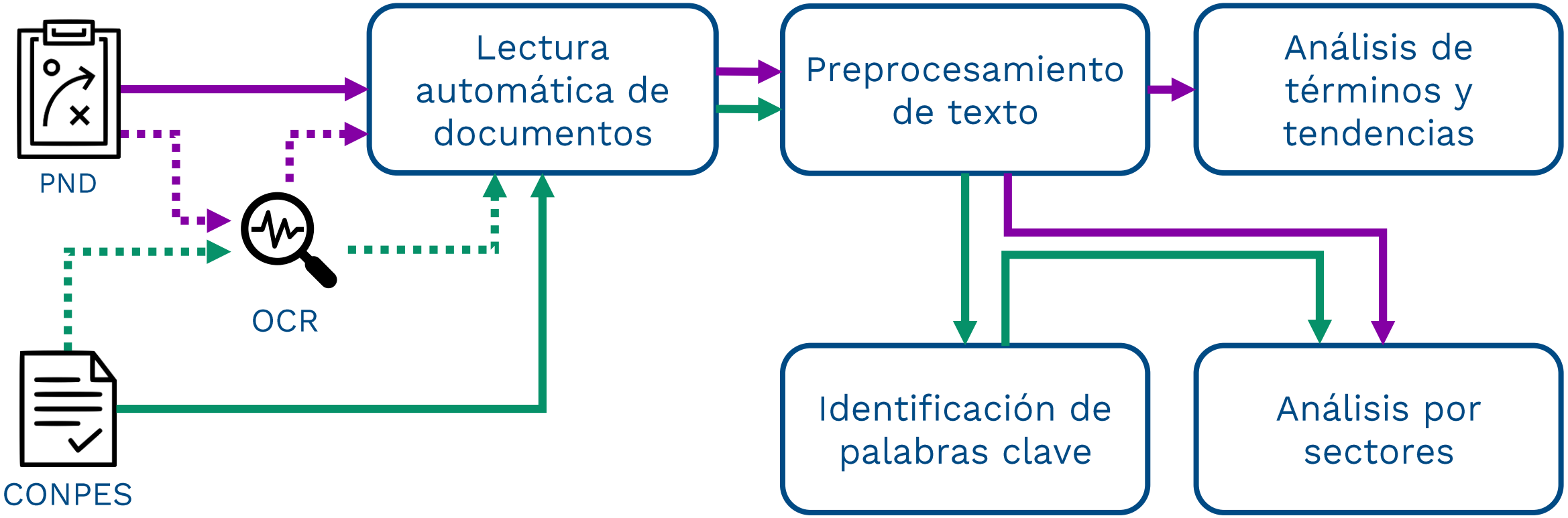
600 Documentos
CONPES más recientes
(corte: febrero 2018)

Metodología

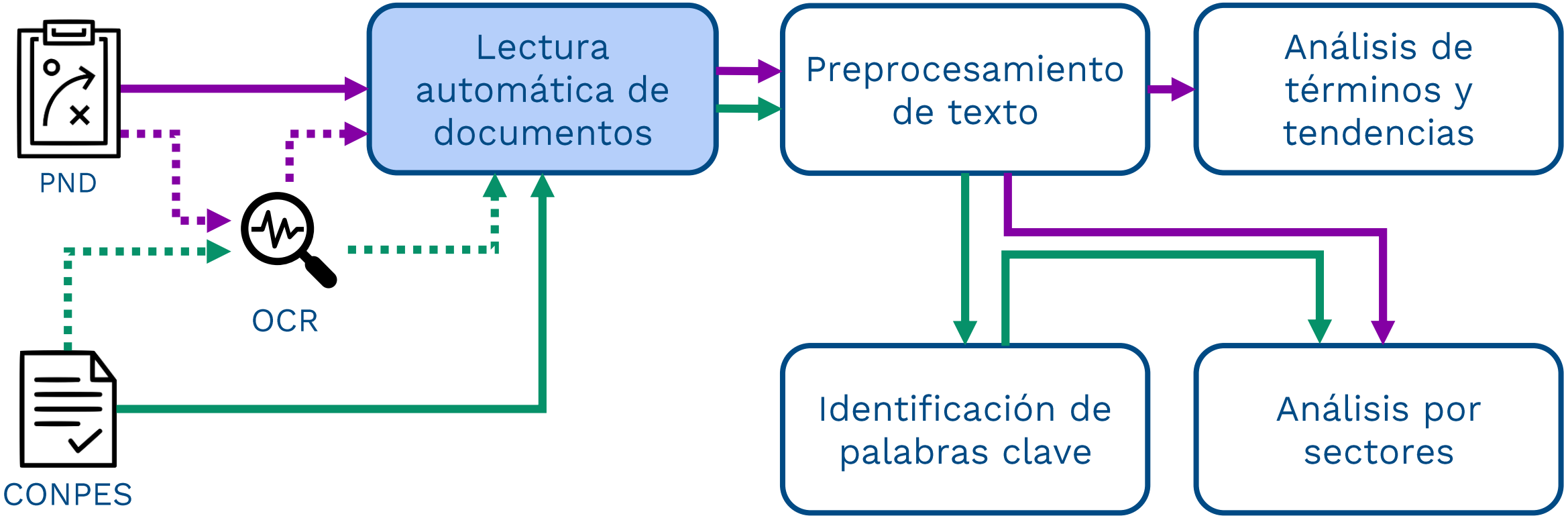


1. Técnicas de procesamiento y vectorización de textos
2. Algoritmo de identificación automática de palabras clave
3. Análisis a nivel sectorial y temporal

Vista general del proyecto



Vista general del proyecto



Lectura automatizada de documentos

PND

7 Documentos

3450 páginas

1,35 millones de palabras

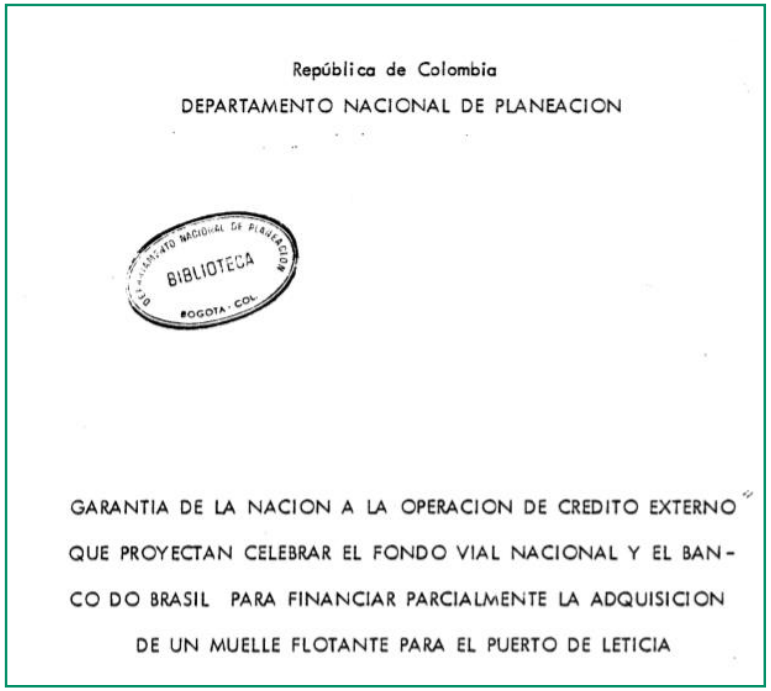
CONPES

132 Documentos

5025 páginas

1,45 millones de palabras

¿Qué es un OCR?



República de Colombia

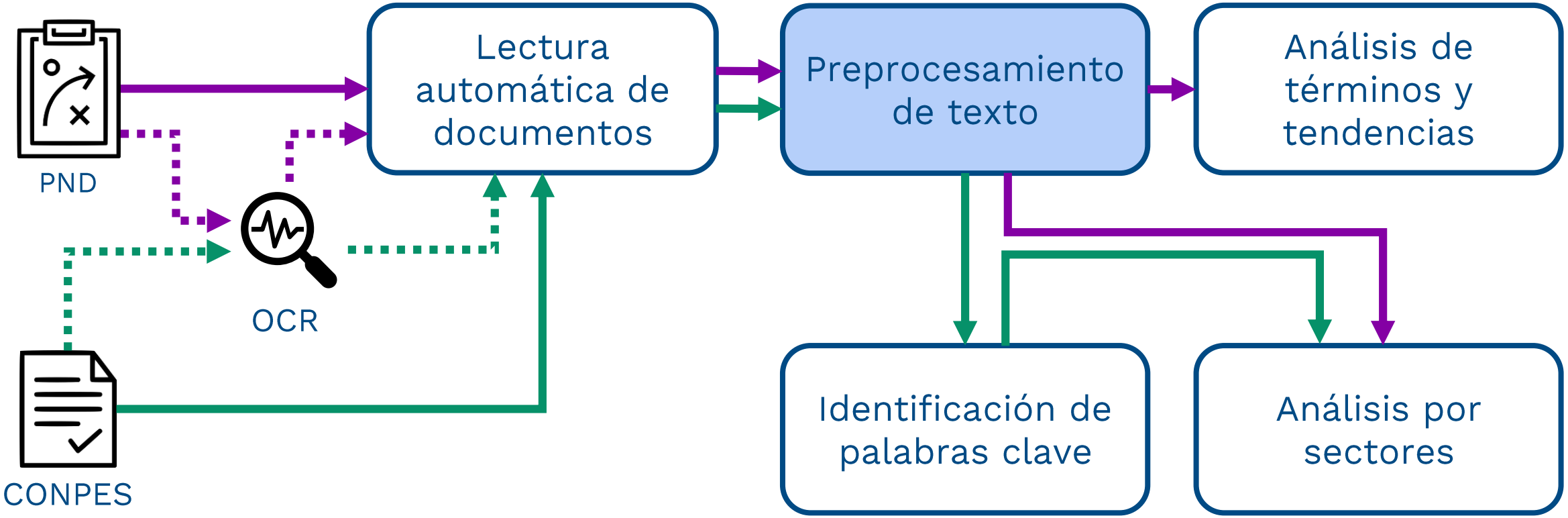
DEPARTAMENTO NACIONAL DE PLANEACION

GARANTÍA DE LA NACIÓN A LA OPERACIÓN DE CRÉDITO EXTERNO QUE PROYECTAN CELEBRAR EL FONDO VIAL NACIONAL Y EL BAN CO DO BRASIL PARA FINANCIAR PARCIALMENTE LA ADQUISICIÓN DE UN MUELLE FLOTANTE PARA EL PUERTO DE LETICIA

Texto escaneado (imagen)

Texto plano

Vista general del proyecto



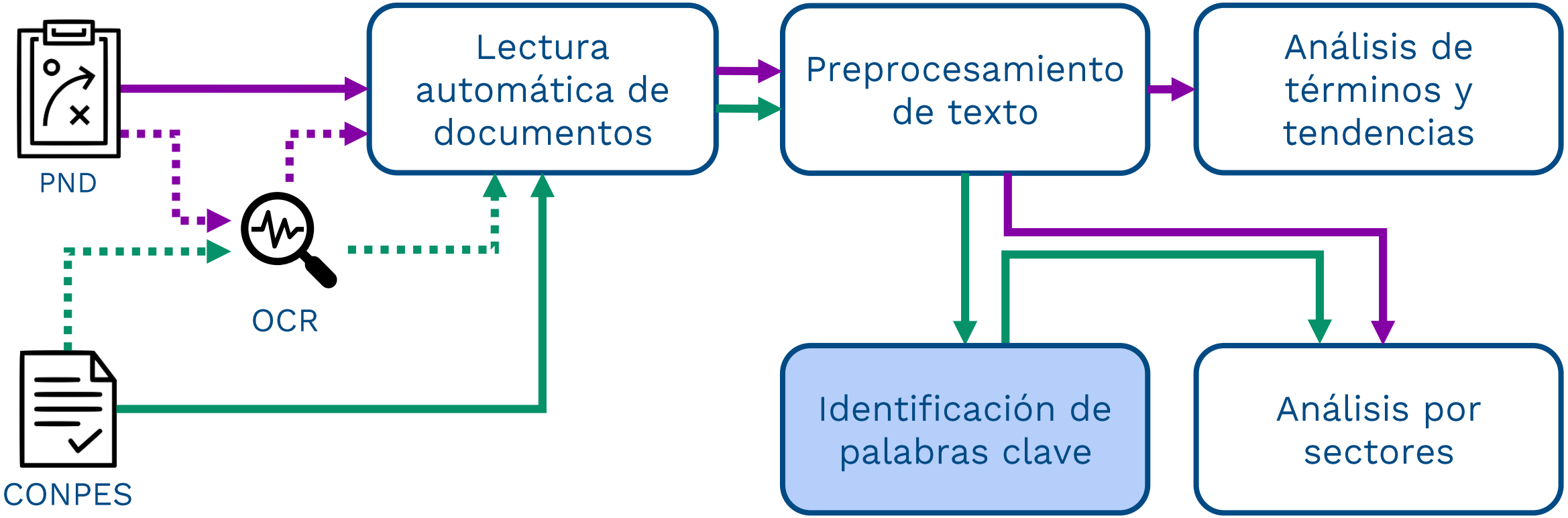
Preprocesamiento de texto

1. Números, puntuación y caracteres especiales
2. *Stopwords*
   - Zonas geográficas de Colombia
   - Palabras que no agregan significado (conectores, etc.)
   - Términos irrelevantes en el análisis (plan, desarrollo, etc.)

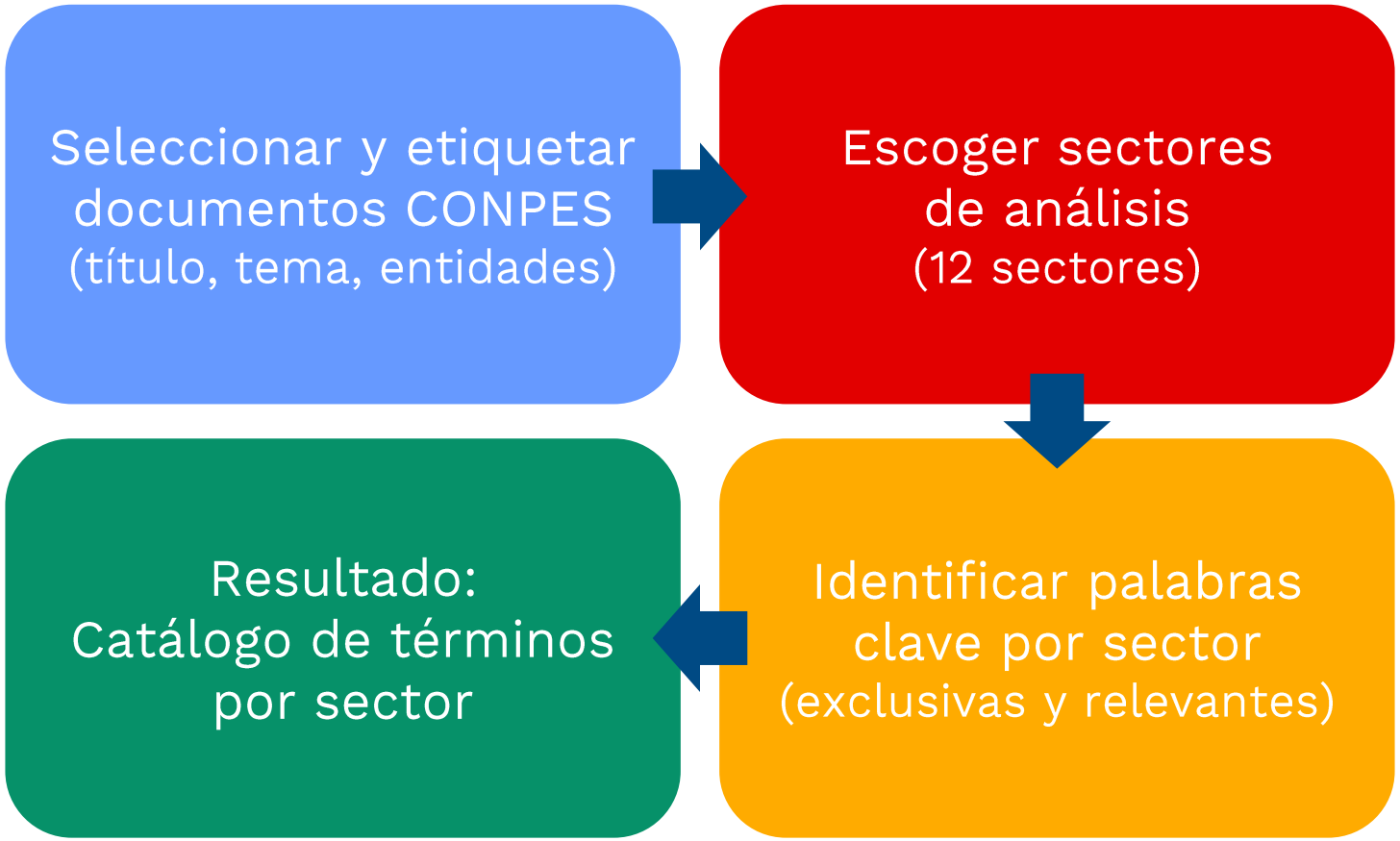
Reducción (PND) a **660 mil** palabras (49%)

2. Identificación de palabras clave por sector

Vista general del proyecto



Identificación de palabras clave



¿Qué sectores se escogieron?

132 Documentos CONPES

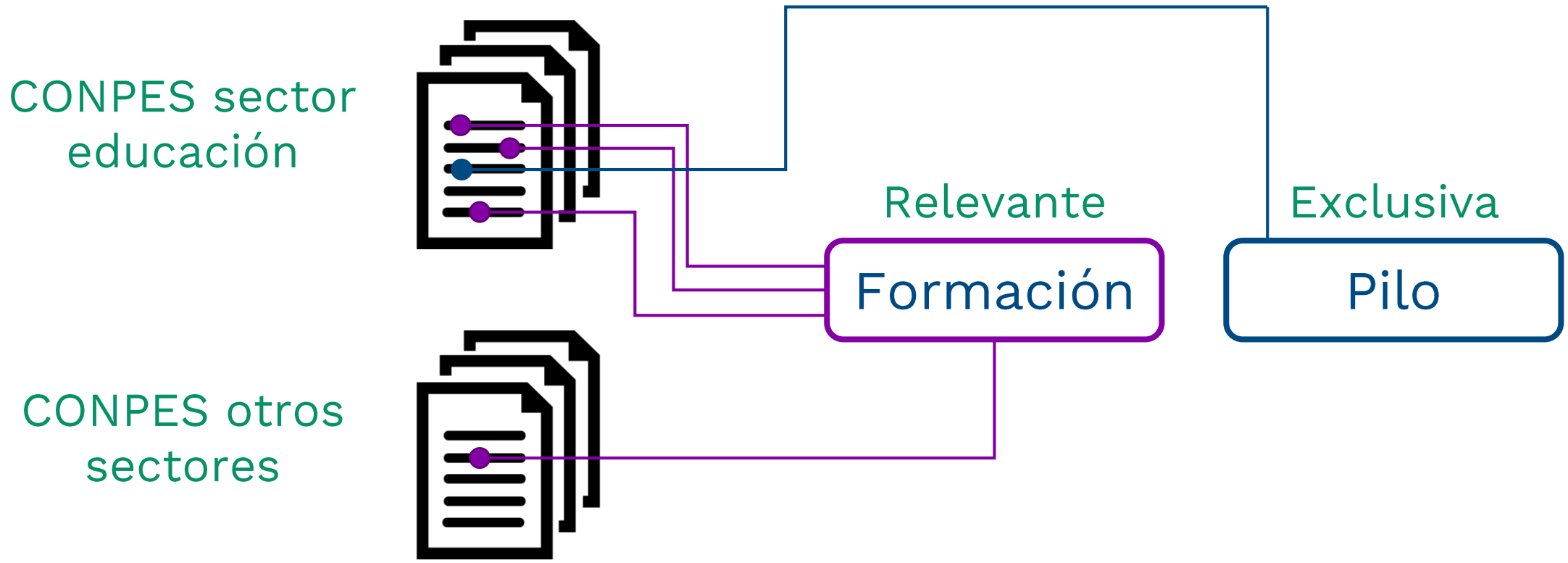
(12 sectores x 11 documentos)

- Inclusión social y reconciliación
- Comercio, industria y turismo
- Ambiente y desarrollo sostenible
- Educación
- Salud y protección social
- Otros siete sectores*

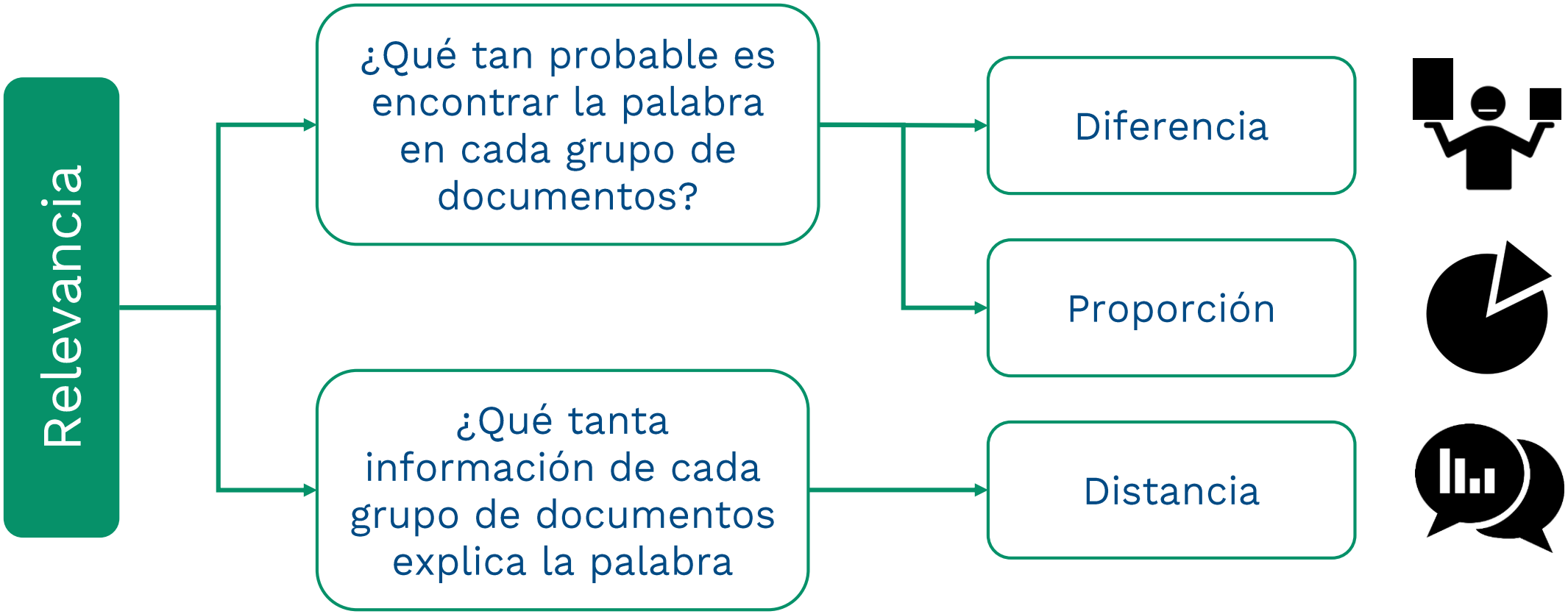
Detalle de documentos CONPES

Sector	Documentos CONPES										
TRANSPORTE	3916	3900	3899	3889	3857	3846	3844	3840	3836	3823	3820
CULTURA, DEPORTE Y RECREACION	3812	3803	3783	3733	3659	3658	3623	3471	3462	3409	3208
EDUCACION	3914	3883_AnexoB	3880	3872	3862	3831	3809	3790	3768	3708	3674
VIVIENDA	3897	3869	3859	3848	3746	3740	3725	3583	3488	3486	3476
AGUA POTABLE Y SANEAMIENTO BASICO	3883_AnexoD	3874	3858	3810	3798	3780	3715	3614	3574	3570	3551
AGRICULTURA	3811	3763	3675	3577	3556	3514	3477	3468	3401	3376	3375
INCLUSIÓN SOCIAL Y RECONCILIACIÓN	3867	3850	3784	3731	3726	3660	3616	3607	3597	3554	3411
AMBIENTE Y DESARROLLO SOSTENIBLE	3716	3700	3697	3680	3624	3594	3550	3459	3451	3344	3343
SALUD Y PROTECCION SOCIAL	3887	3883_AnexoA	3861	3843	3755	3622	3605	3494	3415	3338	3337
MINAS Y ENERGIA	3873	3855	3839	3587	3560	3517	3510	3453	3363	3356	3347
TELECOMUNICACIONES	3898	3854	3769	3701	3670	3650	3579	3518	3506	3457	3371
COMERCIO, INDUSTRIA Y TURISMO	3866	3771	3709	3668	3640	3639	3628	3621	3620	3546	3527

Comparación de documentos



Indicadores de relevancia para cada palabra

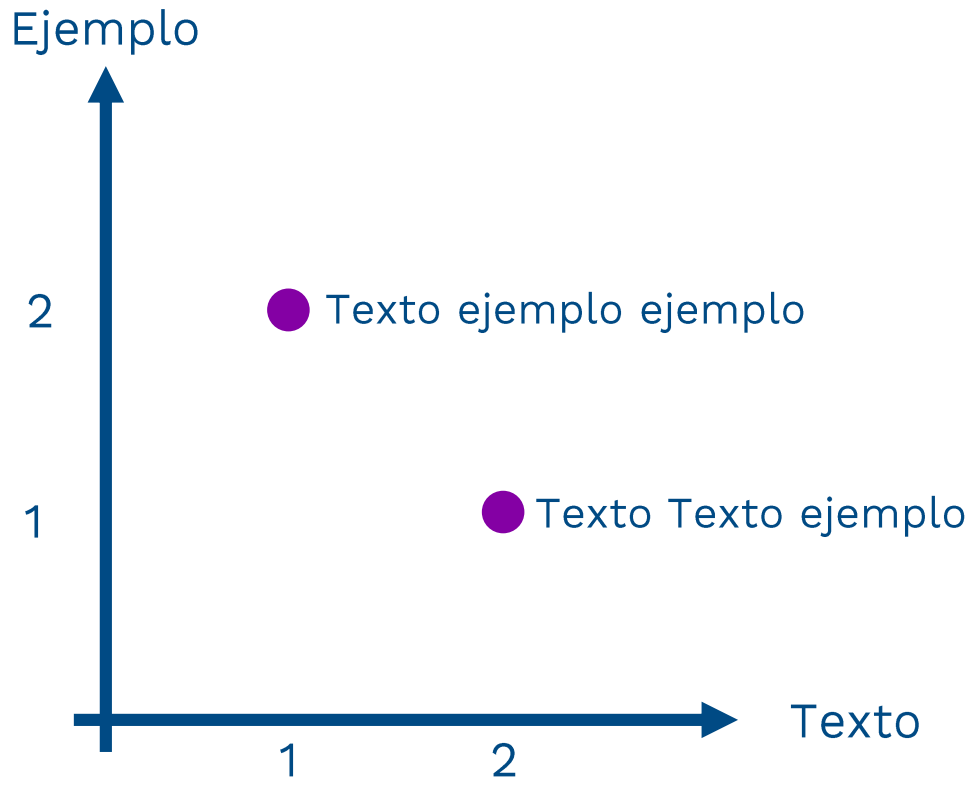


Vectorización del texto

Representación numérica de un texto

“Texto Texto ejemplo”

(2,1)

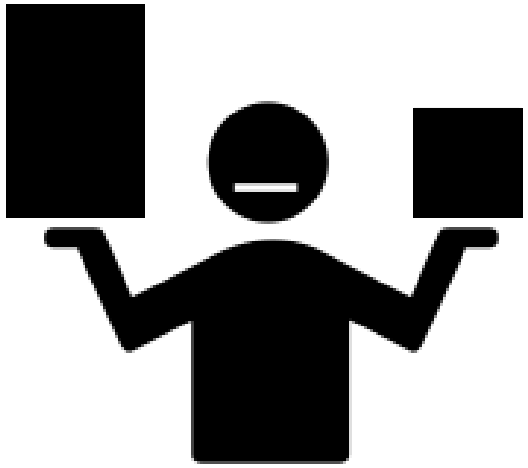


Modelo de *Bag of Words* (BOW)

	un	texto	ejemplo	otro	de	más
un texto ejemplo	1/3	1/3	1/3	0	0	0
otro texto ejemplo	0	1/3	/3	1/3	0	0
ejemplo de texto	0	1/3	1/3	0	1/3	0
un texto más texto	1/4	2/4	0	0	0	1/4

La frecuencia sobre el número total de palabras se puede interpretar como la probabilidad de que una palabra esté en un texto

Indicador de probabilidad: diferencia



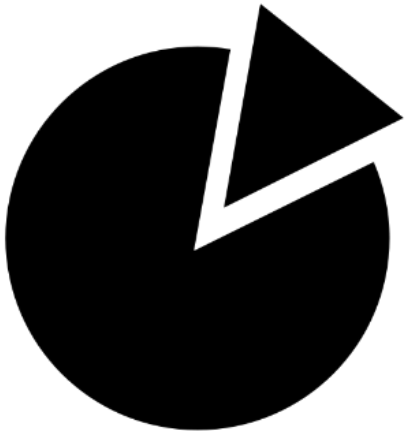
$$\frac{Frec.\ palabra\ en\ corpus\ Edu}{Total\ palabras\ en\ corpus\ Edu} - \frac{Frec.\ palabra\ en\ corpus\ Otros}{Total\ palabras\ en\ corpus\ Otros}$$

Probabilidad de que la palabra aparezca en un documento de educación

Probabilidad de que la palabra aparezca en un documento de otros sectores

Diferencia

Indicador de probabilidad: proporción



$$\frac{Frec.\ palabra\ en\ corpus\ Edu}{Total\ palabras\ en\ corpus\ Edu} \div \frac{Frec.\ palabra\ en\ corpus\ Otros}{Total\ palabras\ en\ corpus\ Otros}$$

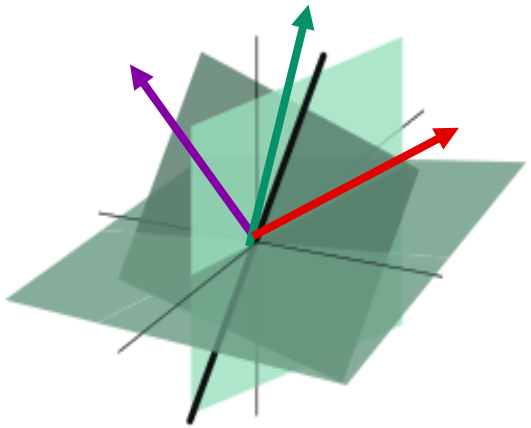
Probabilidad de que la palabra aparezca en un documento de educación

Probabilidad de que la palabra aparezca en un documento de otros sectores

Razón

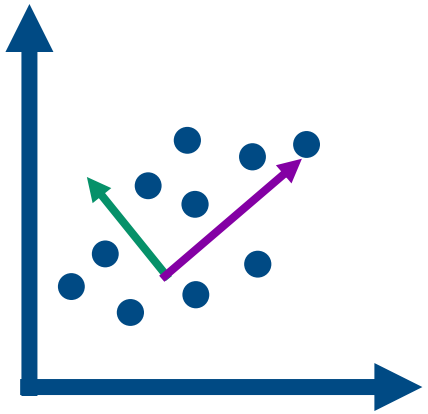
Indicador de información

1



Vectorizar los textos:
$\boldsymbol{Edu}_{BoW}$ y $\boldsymbol{Otros}_{BoW}$

2



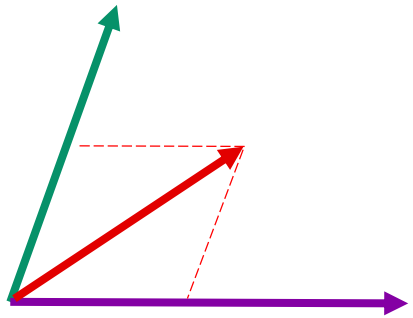
Calcular la matriz de proyección ($\boldsymbol{V}$) al espacio base de $\boldsymbol{Otros}_{BoW}$ utilizando ACP

Se puede interpretar como el espacio de "documentos de política pública de distintos temas"

ACP: Análisis de Componentes Principales

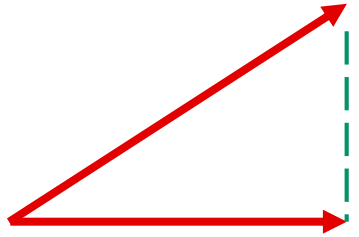
Indicador de información

3



Se encuentra la componente de $\boldsymbol{Edu}_{BoW}$ que se proyecta sobre la base de $\boldsymbol{Otros}_{BoW}$:
$(\boldsymbol{Edu}_{BoW}\boldsymbol{V})\boldsymbol{V}^{\boldsymbol{T}}$

4



Se calcula
$\boldsymbol{Edu}_{BoW} - (\boldsymbol{Edu}_{BoW}\boldsymbol{V})\boldsymbol{V}^{\boldsymbol{T}}$

Se puede interpretar como la parte de cada página que es exclusiva de “documentos del sector S”.

Integración de indicadores

$$f(p) = \frac{\sum_{i \in listas} \text{rank}(p, i)}{\left(\sum_{i \in listas} \text{I}\{p \in i\} \right)^2}$$

Index rank

Nótese que

$p \in i \quad \forall \; p, i$

Catálogos de términos por sector

Transporte

transmilenio
plmb
sitp

transporte
invías
proyecto
troncal
pasajeros
corredor
calzada
conpes
documento
metro

Educación

pnie
pilo
val

educación
formación
superior
jornada
condonación
calidad
colfuturo
icetex
aulas
paga

Minas y Energía

gmdc
microcentral
kwh

gecelca
producción
proyecto
hidroeléctrica
biocombustibles
mw
programa
ecuario
etanol
refinerías

Inclusión y Reconciliación

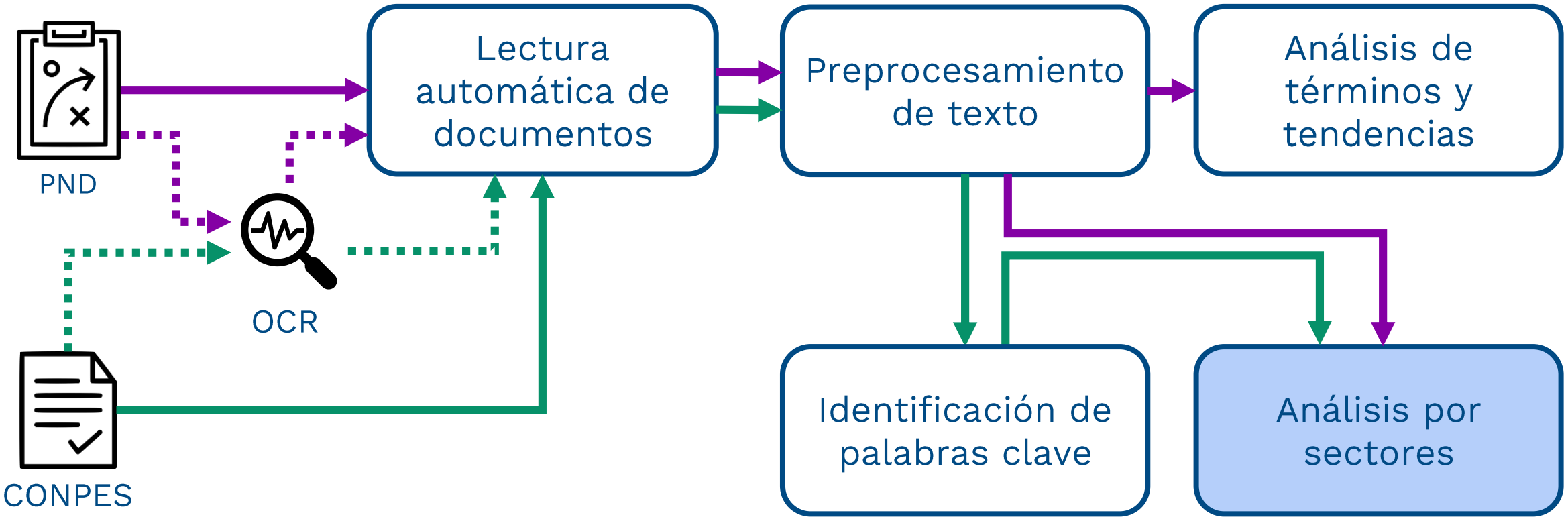
acr
desmovilizados
ddhh

afrocolombiana
reintegración
pped
víctimas
social
posconflicto
juntos
reparación
atención
violaciones

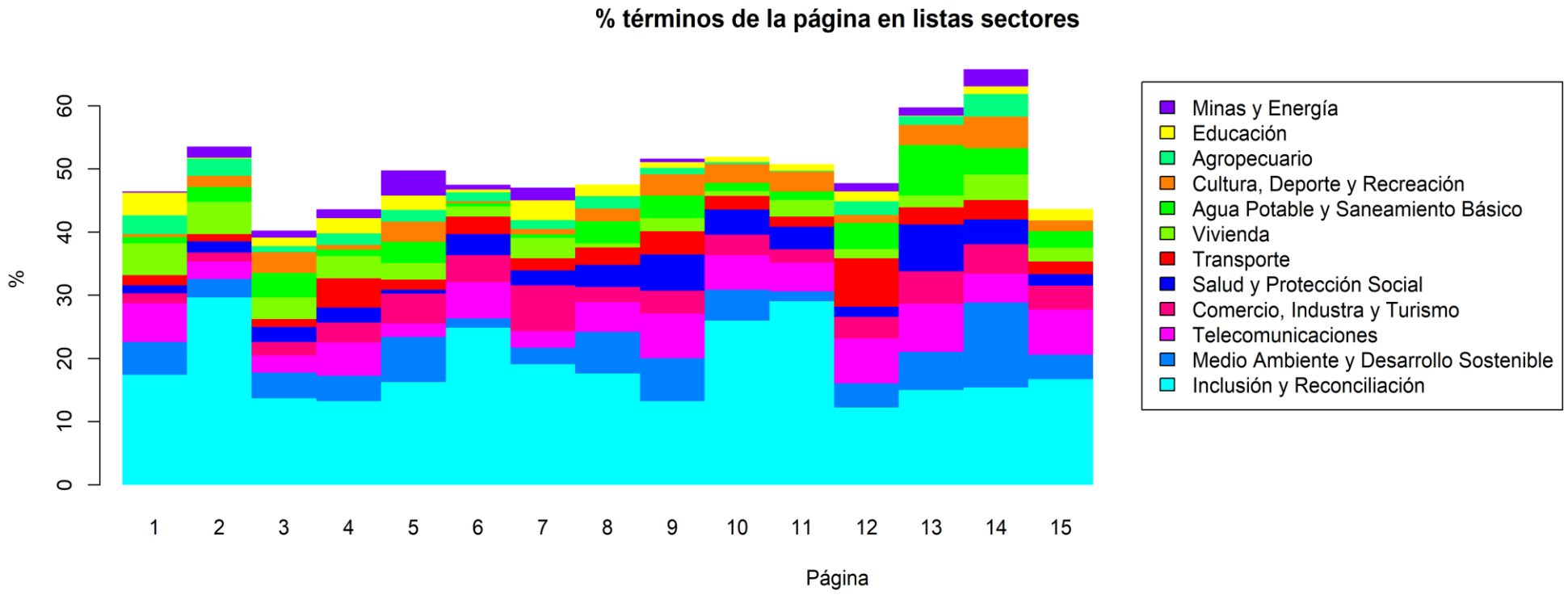
50 exclusivas y 200 relevantes en cada sector

3. Estudio de PND por sector: 1990 – 2014

Vista general del proyecto



Ejemplo categorización



PND 2014-2018, Capítulo 2: “COLOMBIA EN PAZ”

El método *identifica correctamente* que la mayoría de términos en el capítulo corresponden a “Inclusión y Reconciliación”.

Resultados categorización PND (1994-2014)

1990	1994	1998	2002	2006	2010	2014
Comercio	Comercio	Inclusión	Inclusión	Inclusión	Inclusión	Inclusión
Salud	Inclusión	Comercio	Telecom.	Comercio	Comercio	Ambiente
Educación	Ambiente	Ambiente	Comercio	Ambiente	Ambiente	Comercio
Inclusión	Salud	Educación	Salud	Telecom.	Educación	Telecom.
Vivienda	Educación	Salud	Educación	Educación	Telecom.	Transporte
Tansporte	Telecom.	Telecom.	Ambiente	Salud	Salud	Educación
Ambiente	Vivienda	Cultura	Vivienda	Vivienda	Transporte	Salud
Agua	Cultura	Vivienda	Cultura	Agua	Vivienda	Vivienda
Cultura	Transporte	Transporte	Agua	Transporte	Cultura	Agua
Agro.	Agua	Agua	Energía	Cultura	Agro.	Cultura
Telecom.	Agro.	Energía	Transporte	Agro.	Energía	Agro.
Energía	Energía	Agro.	Agro	Energía	Agua	Energía

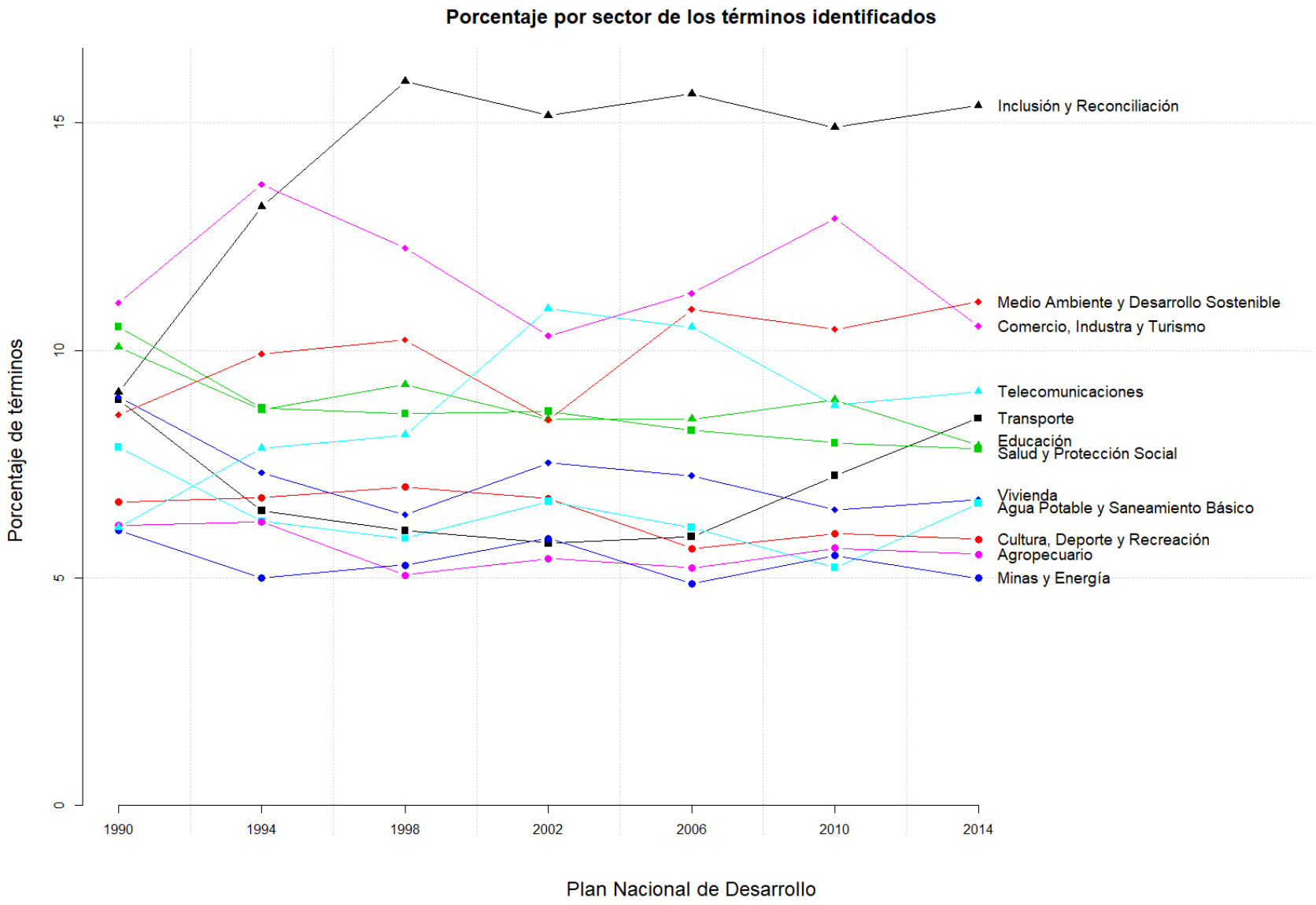
En promedio, cerca del 20% de los términos se reconocen como de alguno de los 12 sectores

Sectores con más términos reconocidos

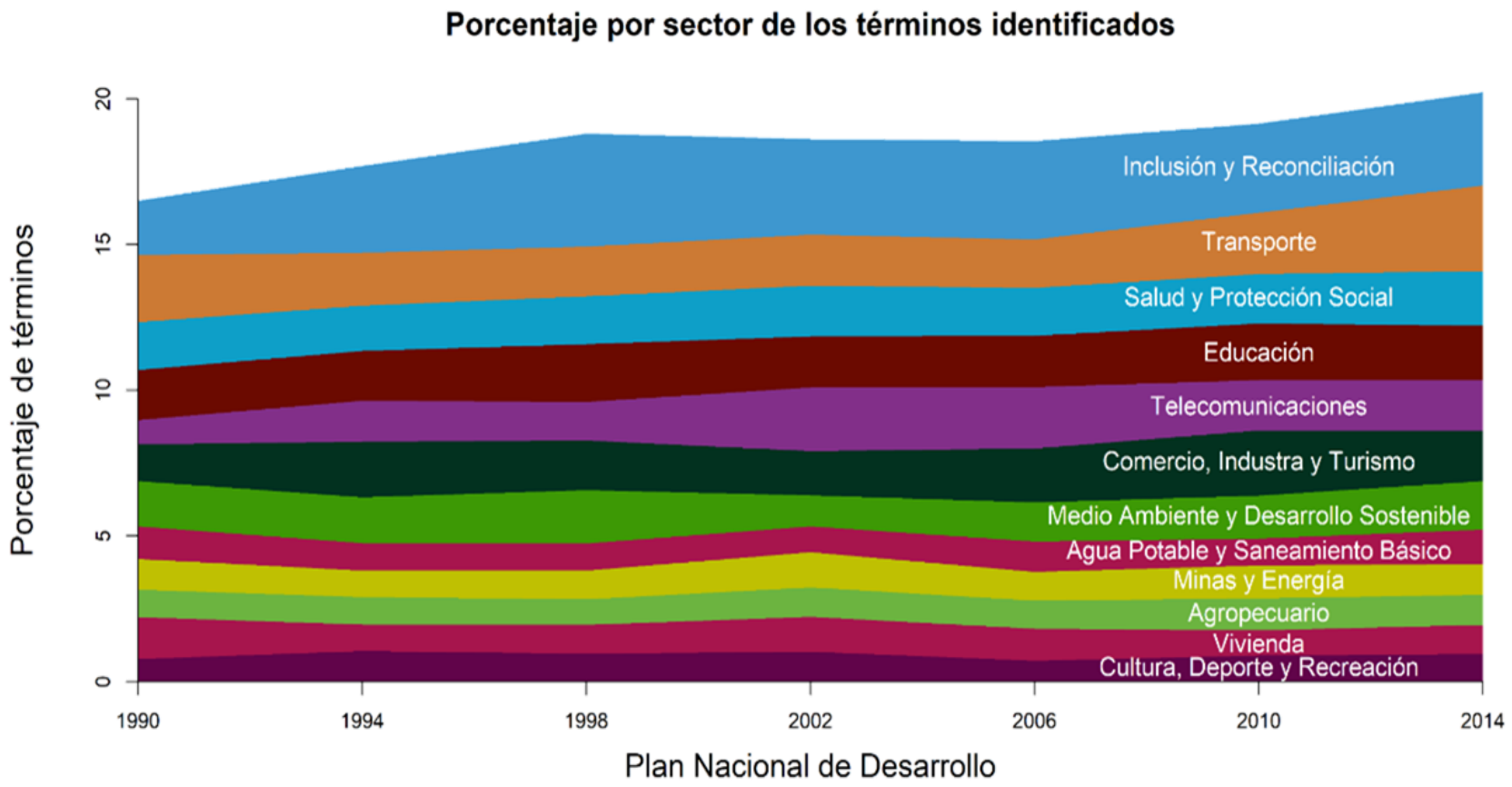
Inclusión social y reconciliación

Comercio, industria y turismo

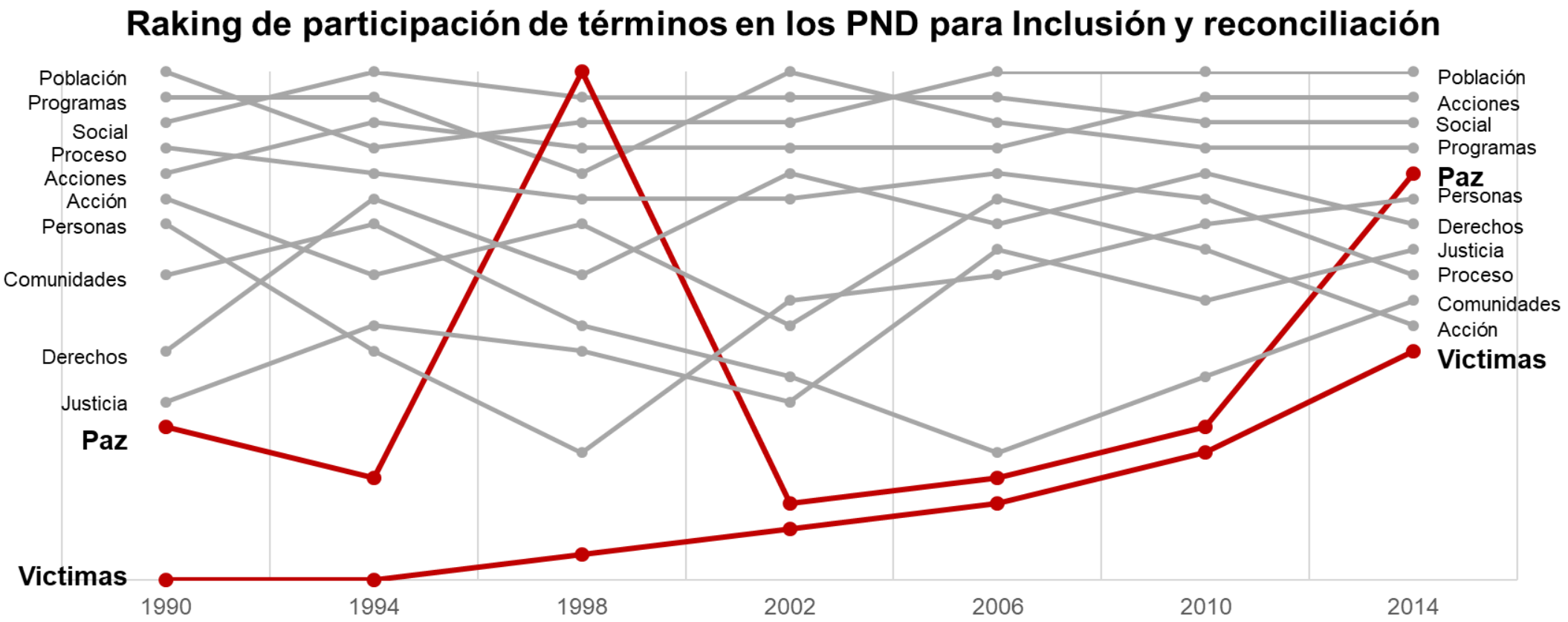
Sectores en el tiempo



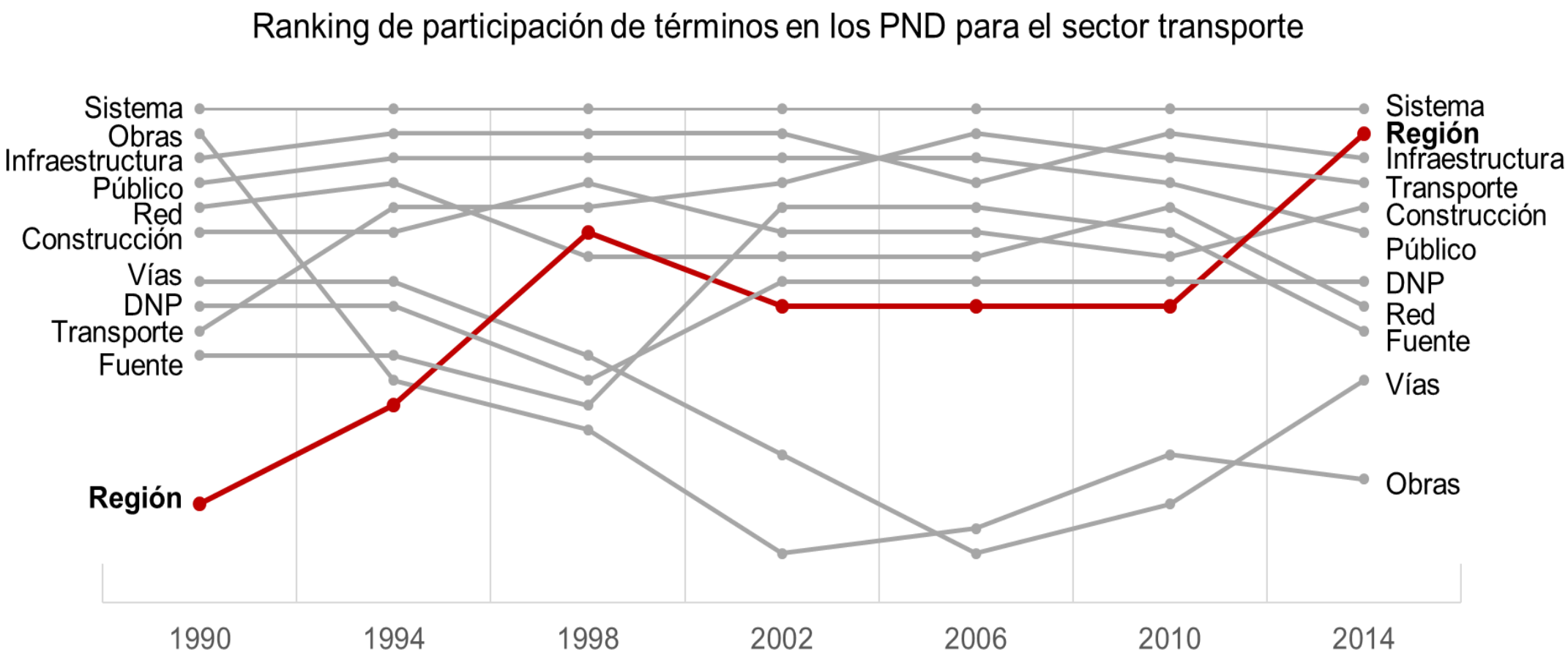
Sectores en el tiempo



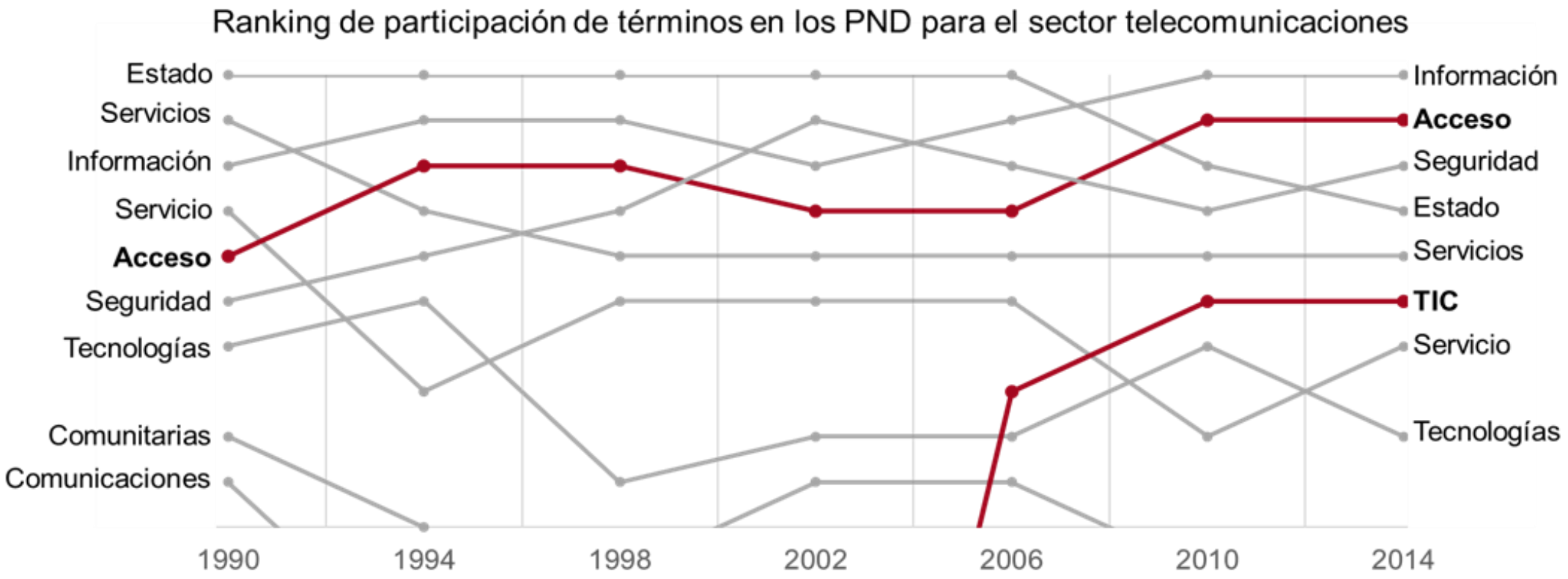
Ejemplo evolución de términos



Ejemplo evolución de términos

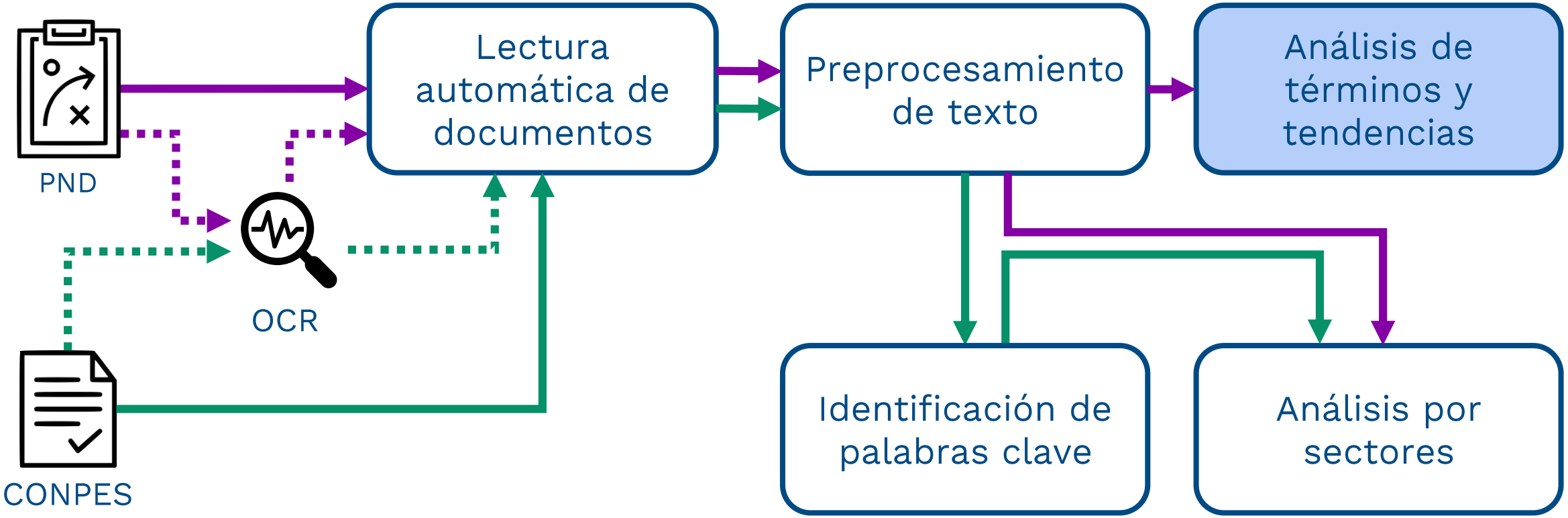


Ejemplo evolución de términos



4. Estudio de términos más relevantes por PND: 1970 - 2018

Vista general del proyecto



Descripción del análisis

Se quiso extender el alcance del análisis de los planes de desarrollo, identificando qué términos han tenido relevancia, y cómo esta ha cambiado durante los años.

1. Términos más frecuentes en los planes de desarrollo

2. Análisis de tendencias para identificar qué términos han ganado o perdido relevancia con los años

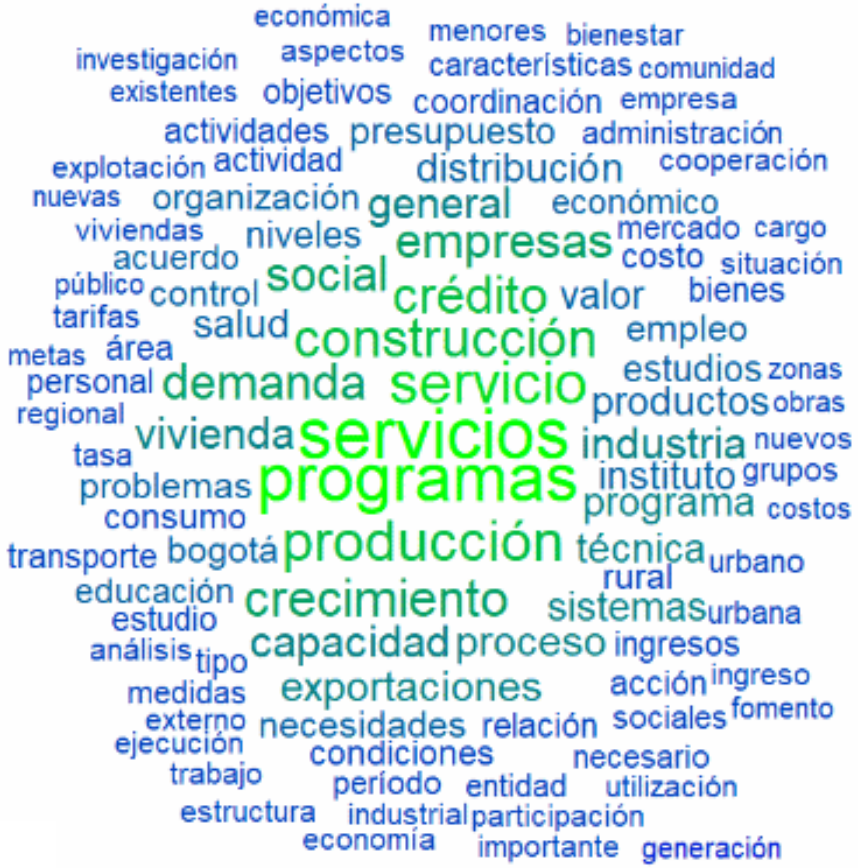
Términos más frecuentes en los PND

El primer paso fue identificar, para cada PND entre 1970 y 2018:

- 100 palabras más frecuentes
- 100 n-gramas de orden 2 más frecuentes
- 50 n-gramas de orden 3 más frecuentes
- 50 n-gramas de orden 4 más frecuentes

Términos más frecuentes: Palabras

PND 1970



Términos más frecuentes: Bi-gramas

PND 1970

Tendencias en la relevancia de los términos

El objetivo en este análisis es identificar cómo va cambiando la relevancia de algunos términos (palabras y bi-gramas) con el paso de los años.

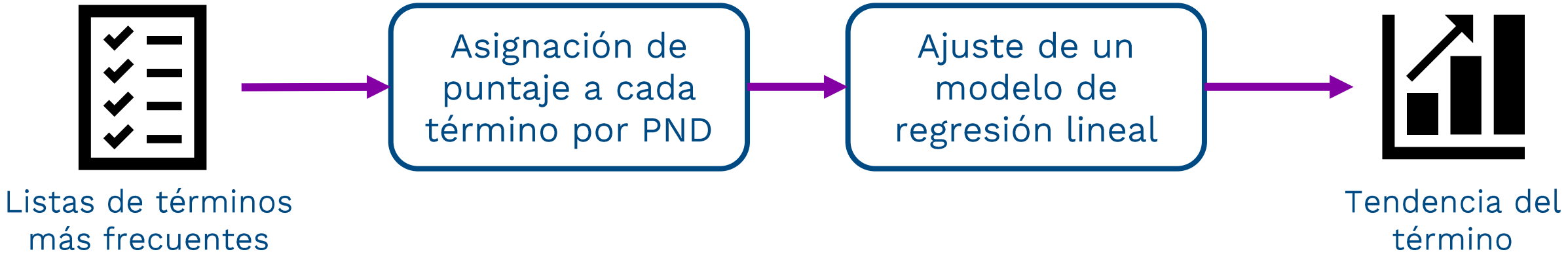
Términos que han perdido relevancia con el tiempo

Términos que se han mantenido relevantes

Términos que han ganado relevancia con el tiempo

Tendencias en la relevancia de los términos

Para cada término que haya estado dentro de los más frecuentes para al menos un PND, se estimó un modelo lineal.



Tendencias en la relevancia: Palabras

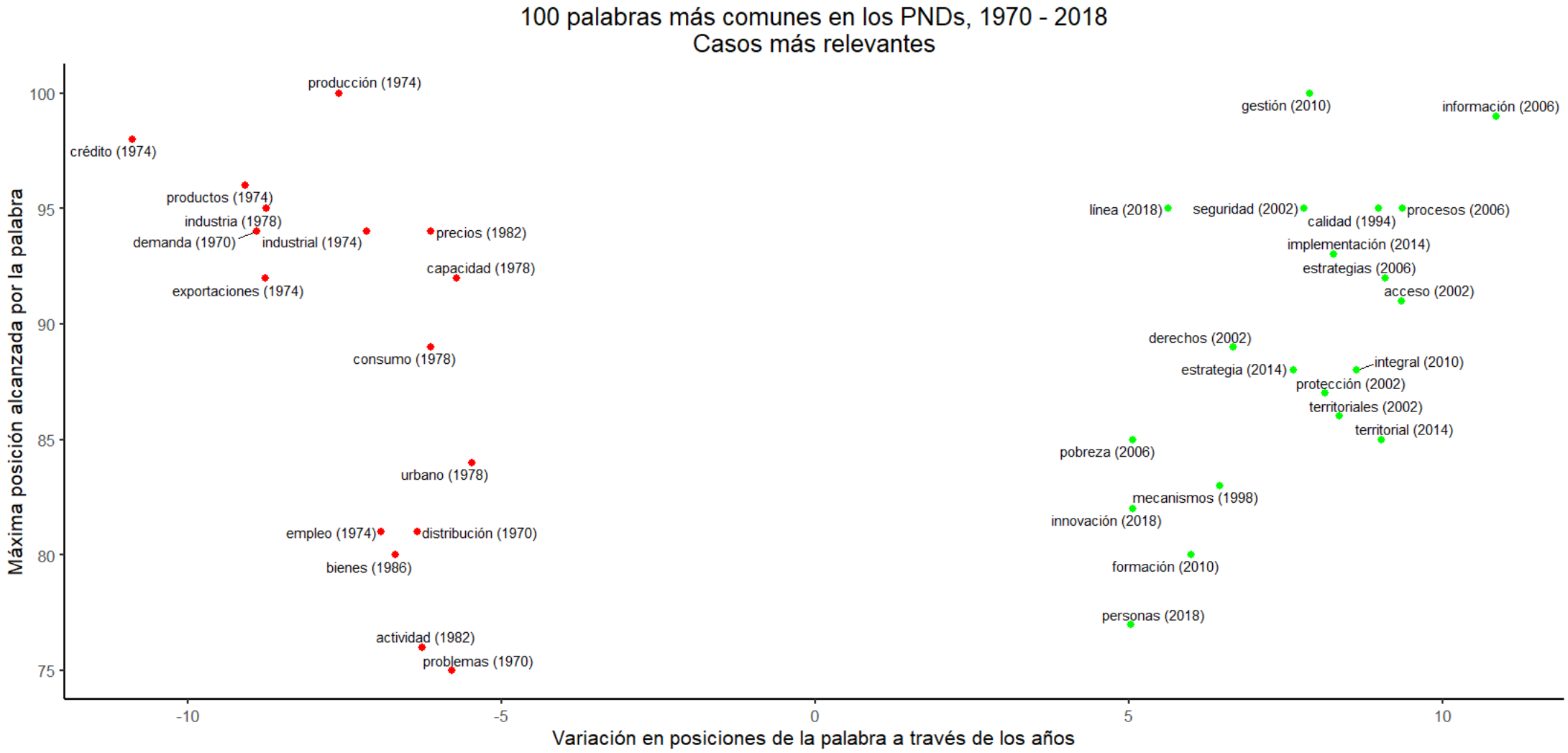


Figura: Máxima posición alcanzada por cada término versus su variación en posiciones con el tiempo

Tendencias en la relevancia: Palabras

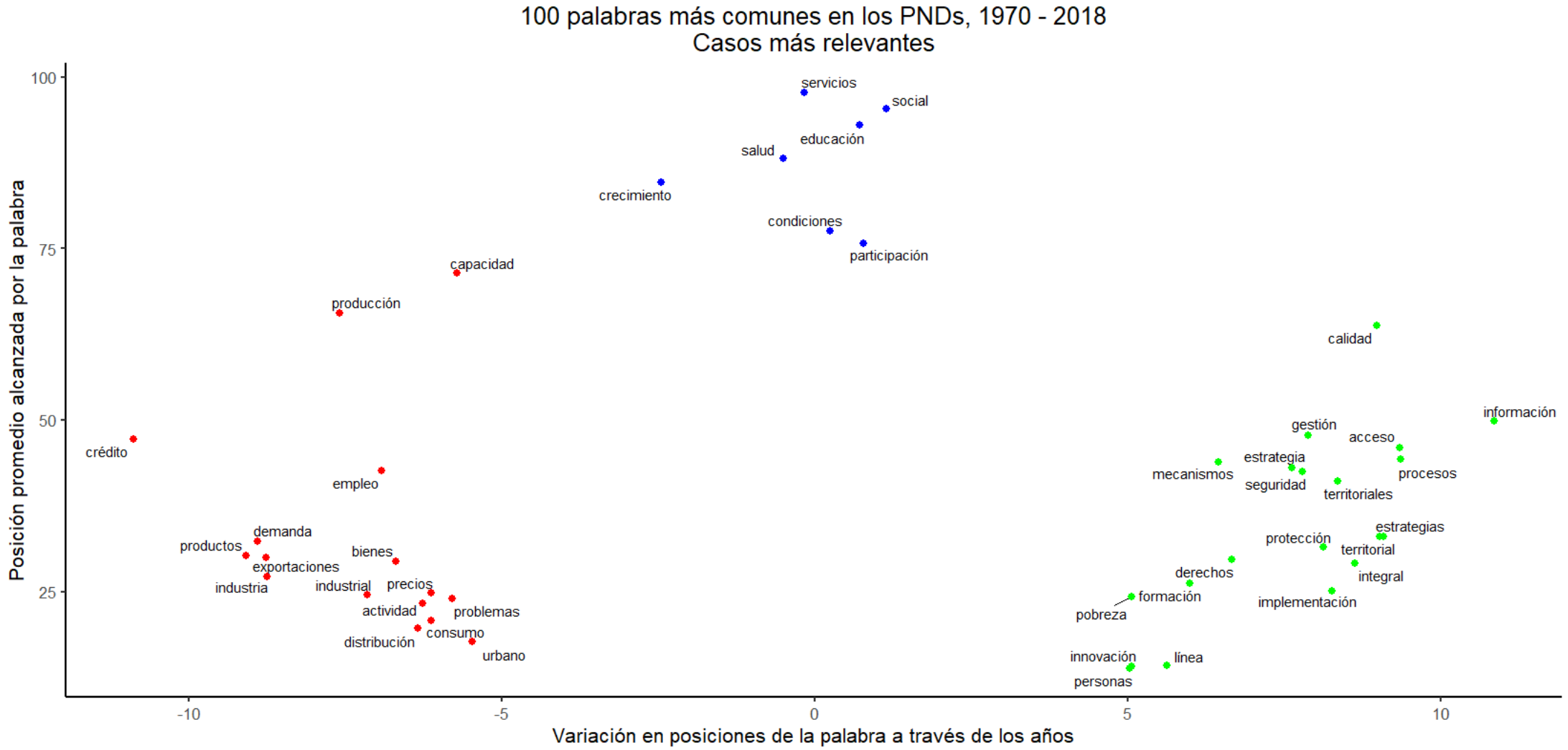


Figura: Posición promedio alcanzada por cada término versus su variación en posiciones con el tiempo

Tendencias en la relevancia: Bigramas

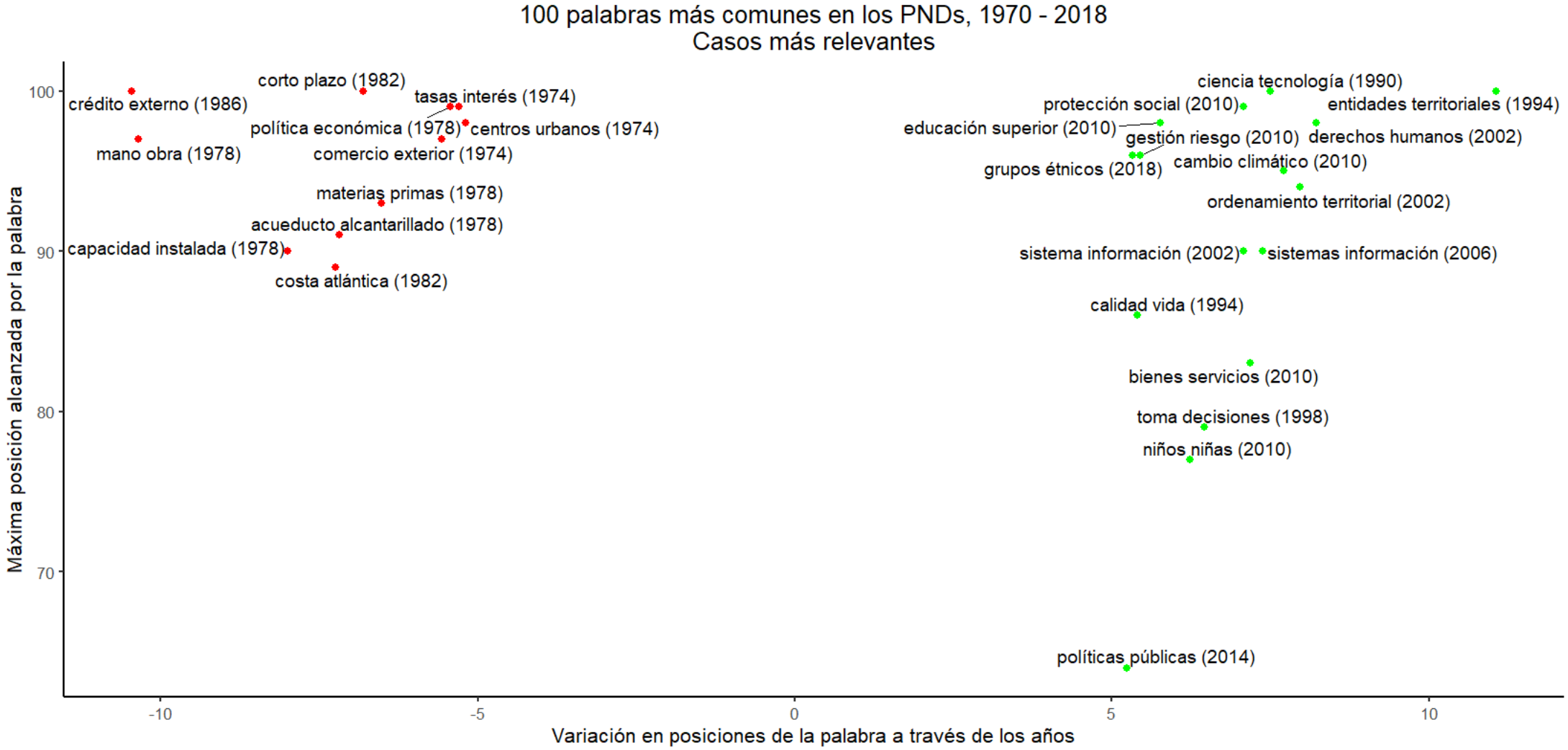


Figura: Máxima posición alcanzada por cada término versus su variación en posiciones con el tiempo

Tendencias en la relevancia: Bigramas

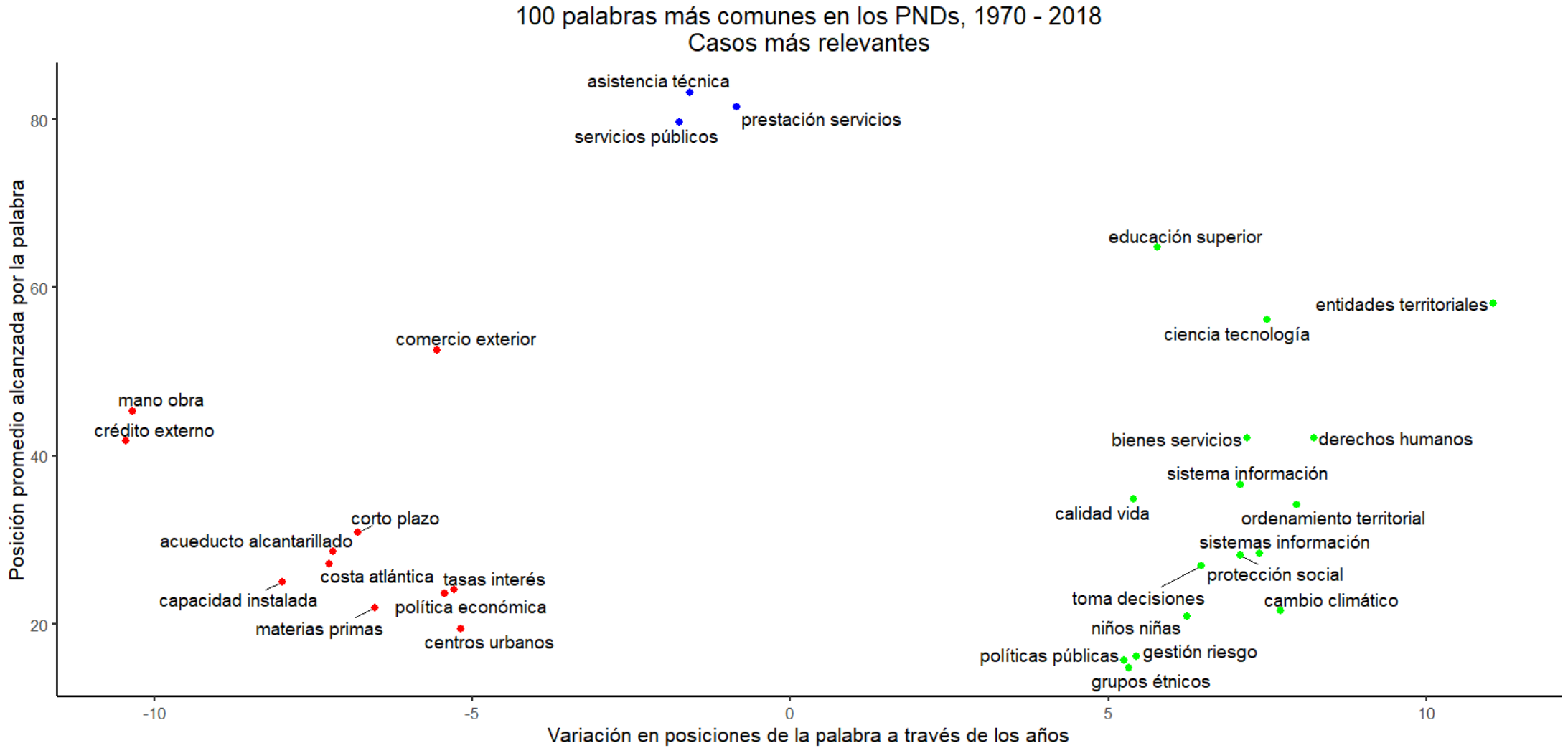
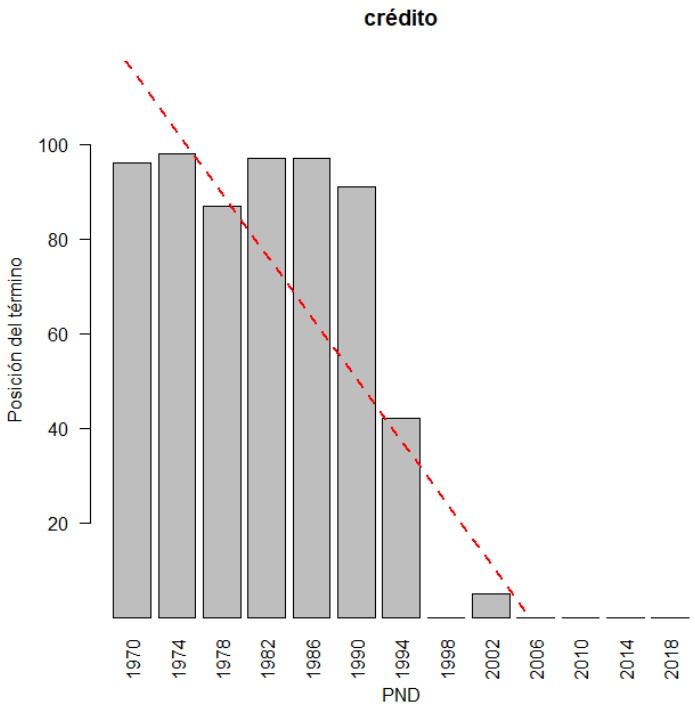
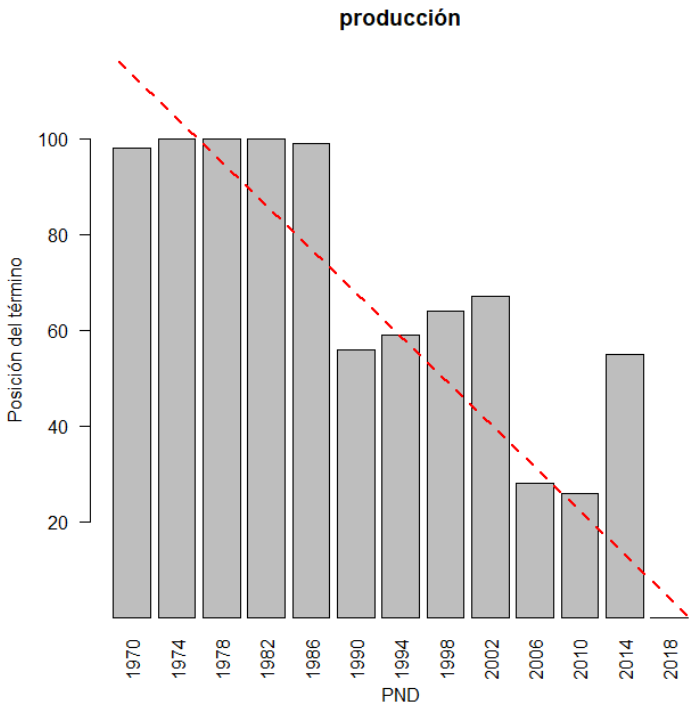


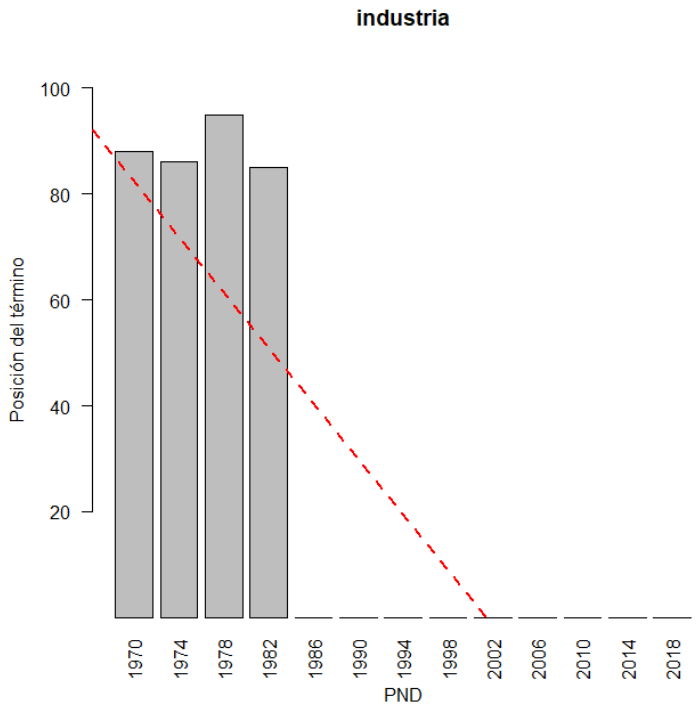
Figura: Posición promedio alcanzada por cada término versus su variación en posiciones con el tiempo

Términos que han perdido relevancia

Palabras

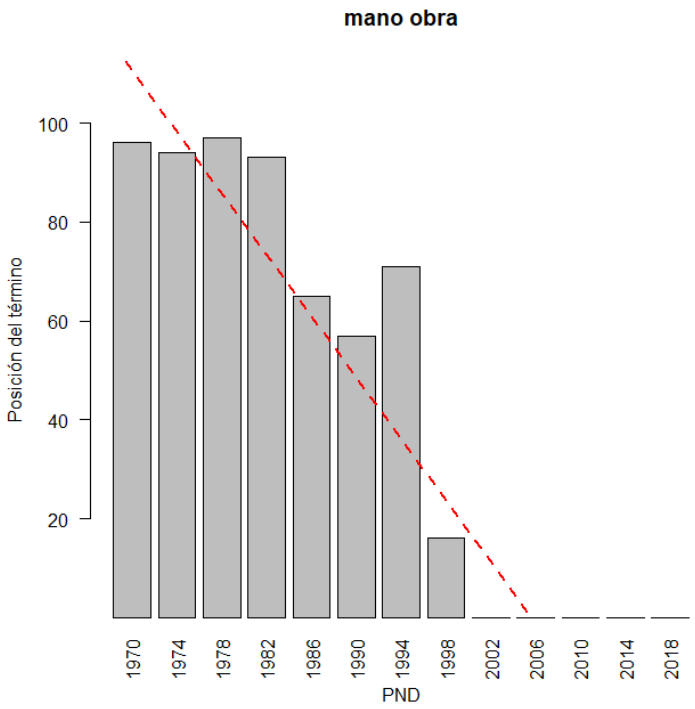


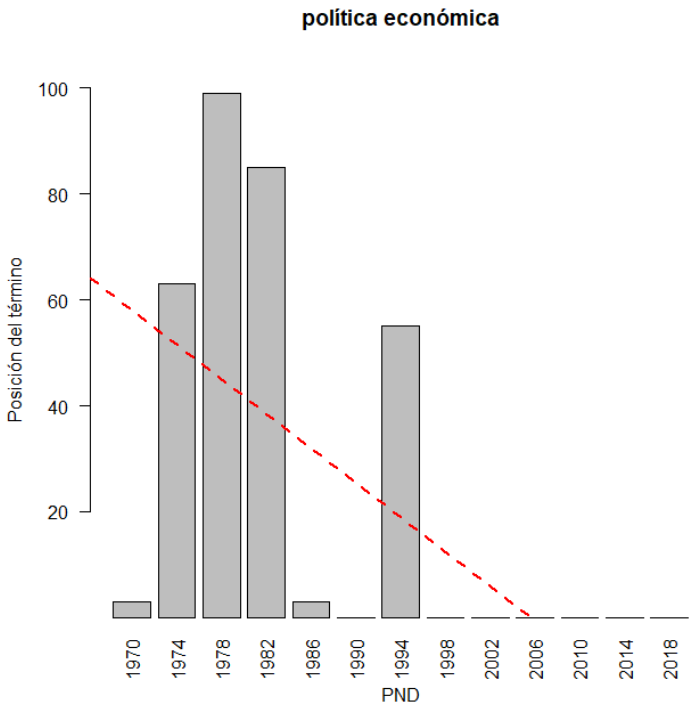


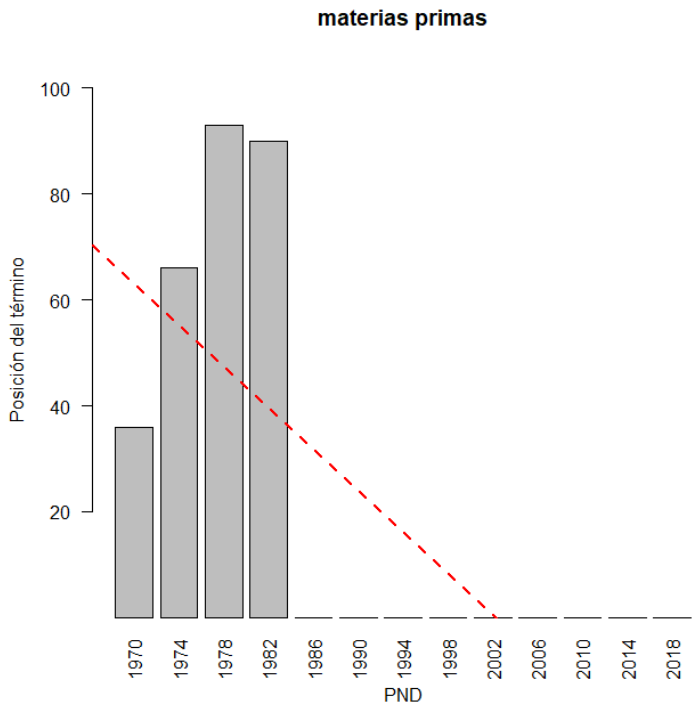


Términos que han perdido relevancia

Bigramas

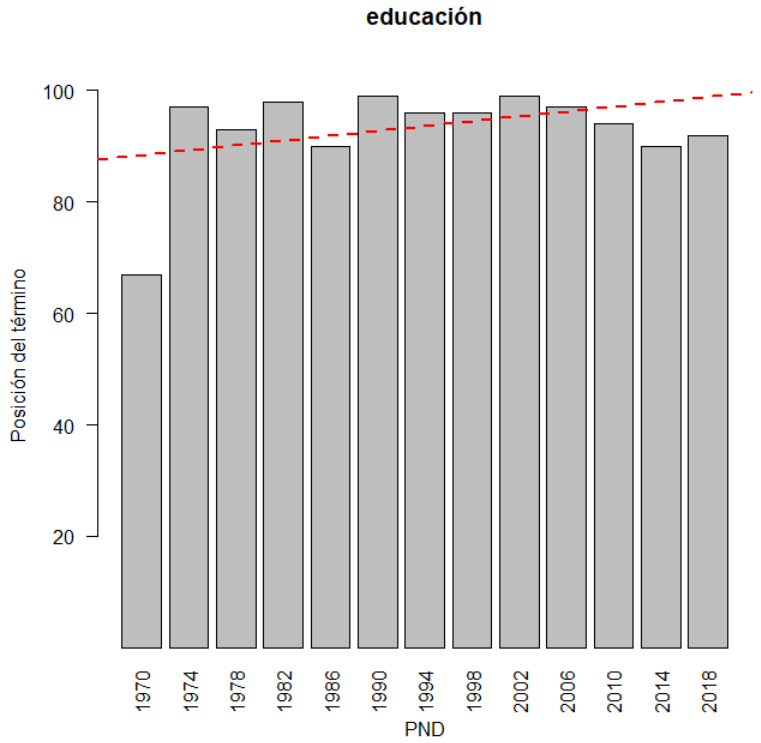


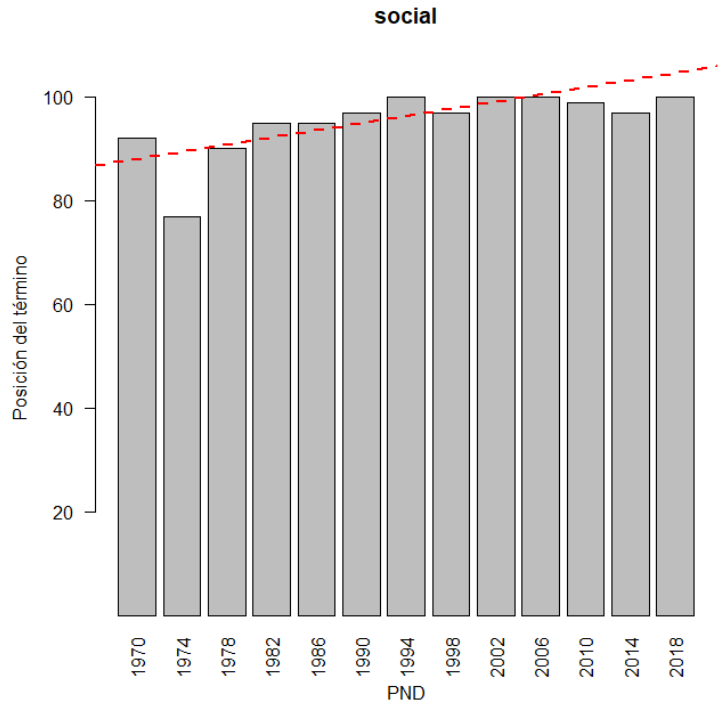




Términos que han mantenido relevancia

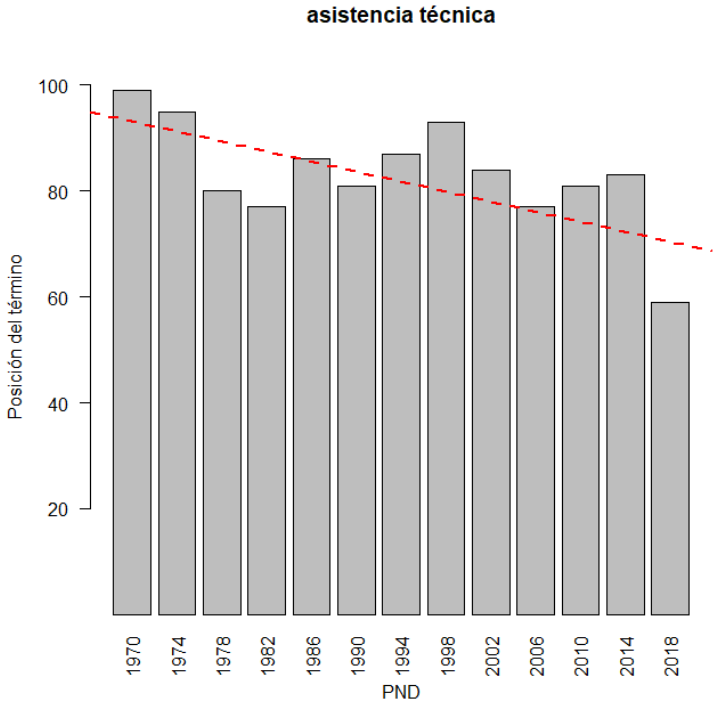
Palabras

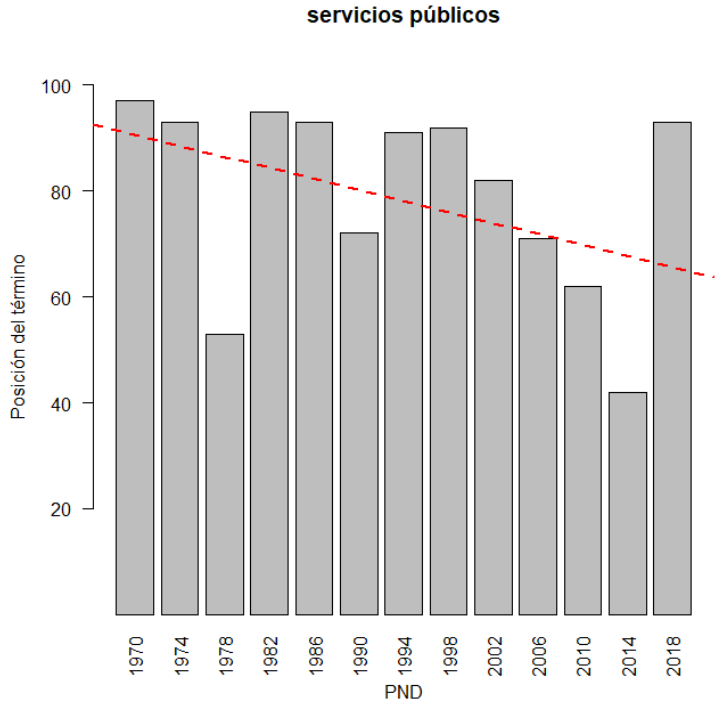


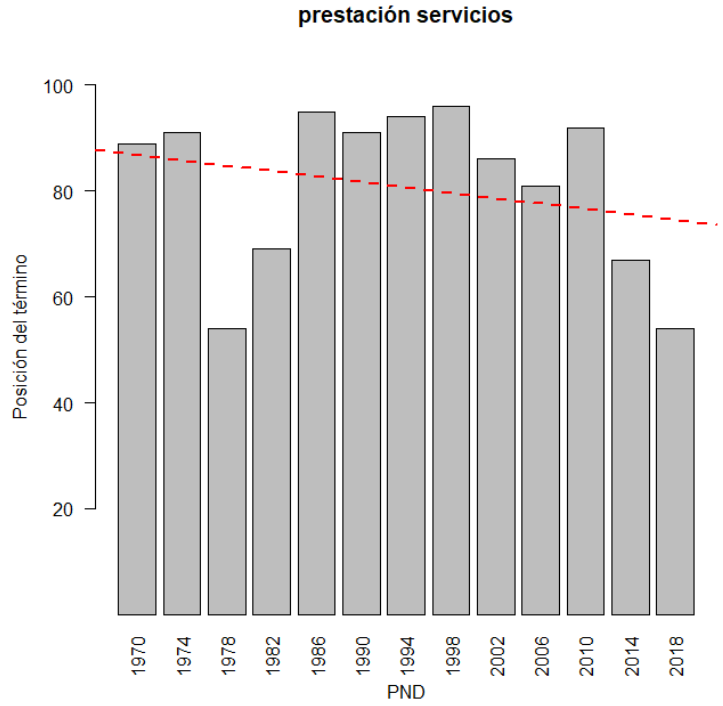


Términos que han mantenido relevancia

Bi-gramas

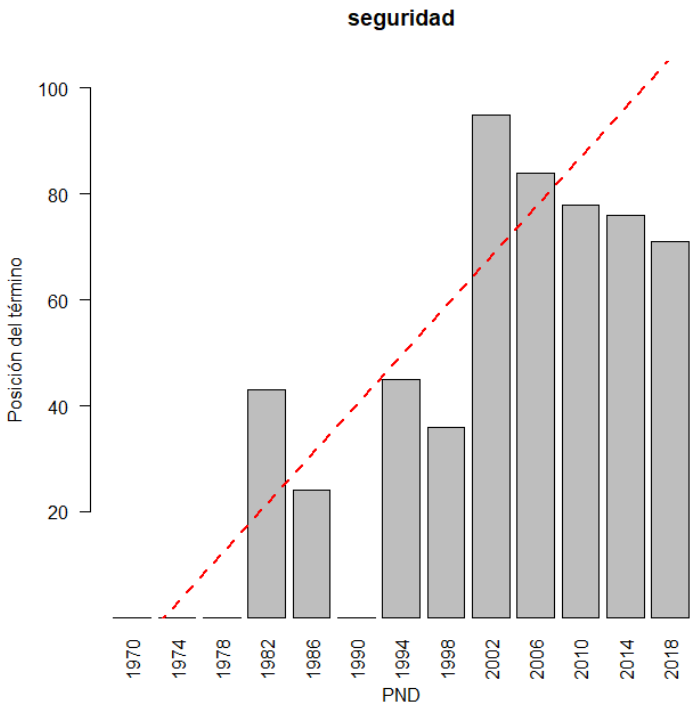


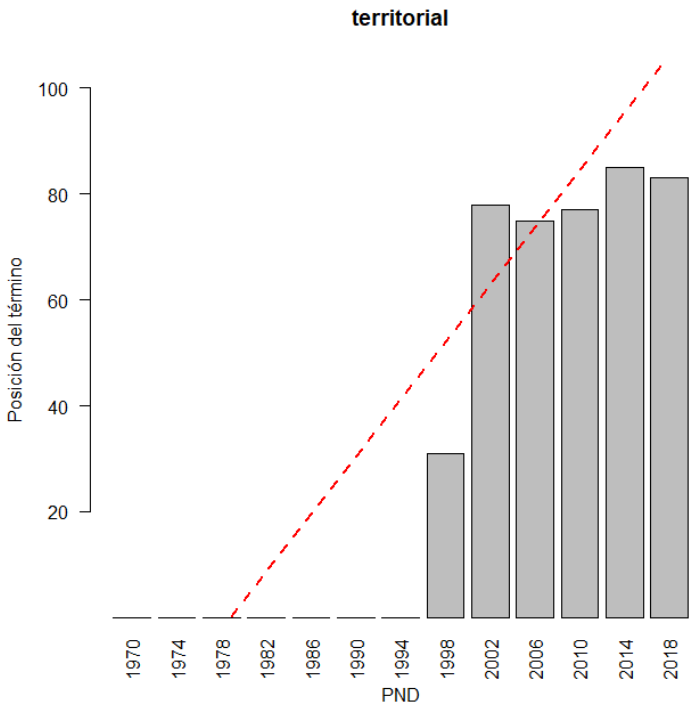


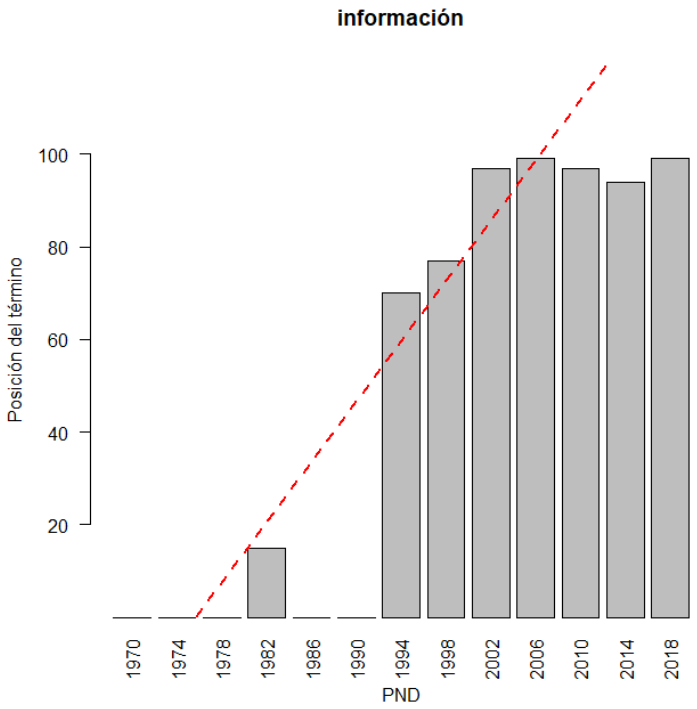


Términos que han ganado relevancia

Palabras

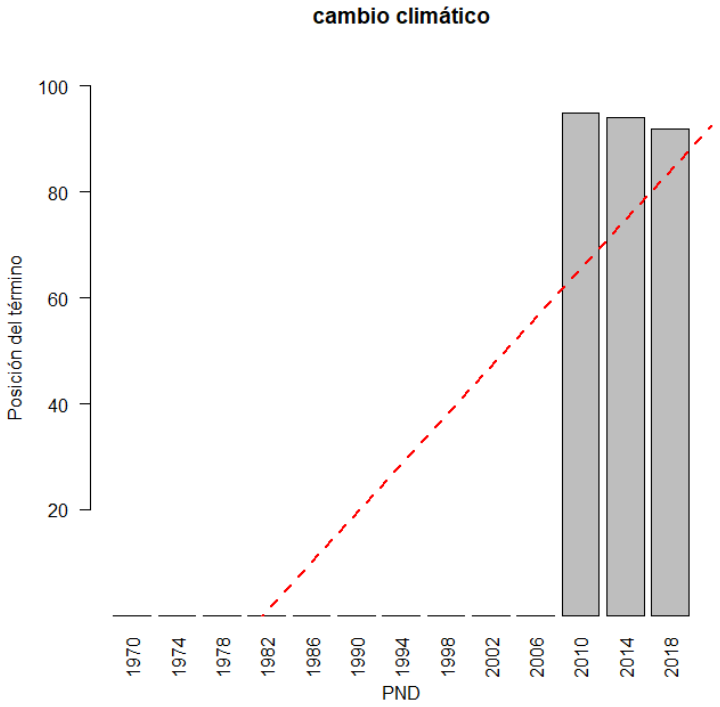


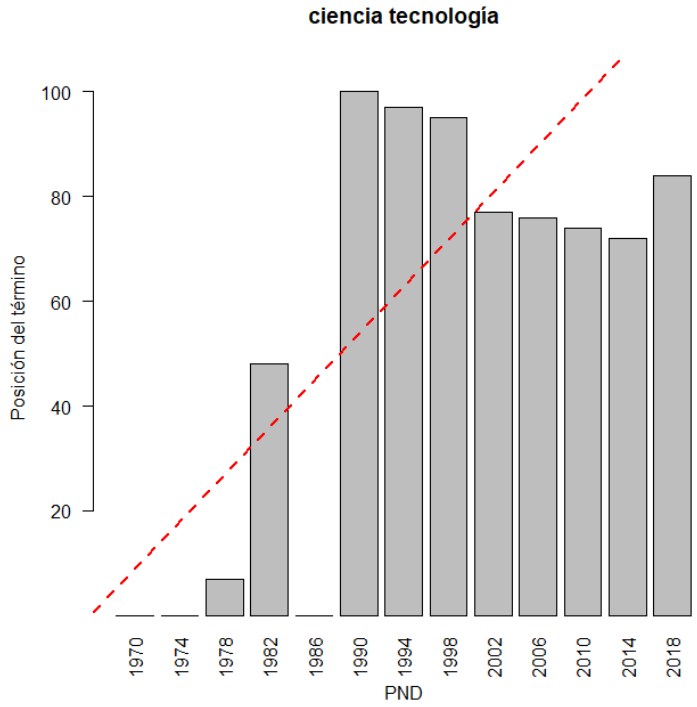


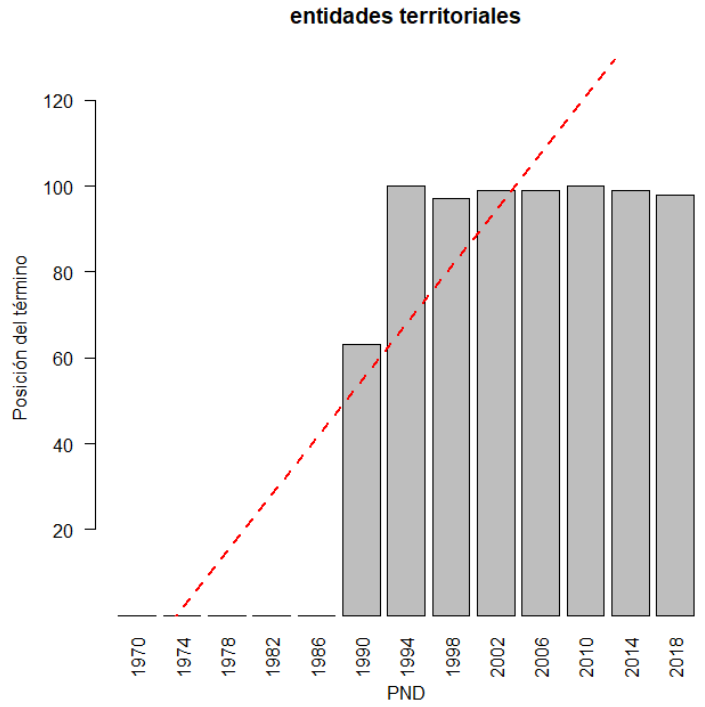


Términos que han ganado relevancia

Bi-gramas







Conclusiones y recomendaciones

1. A través de técnicas de minería de texto y análisis exploratorios es posible obtener un panorama general de cómo ha variado en el tiempo el énfasis que se hace en ciertos temas y sectores en los PND.

Conclusiones y recomendaciones

2. La metodología de identificación de palabras clave para una temática en específico, implementada en este proyecto, permite obtener indicadores objetivos sobre la relevancia de ciertos términos.

Conclusiones y recomendaciones

3. La metodología de identificación de palabras clave no necesariamente debe sustituir el criterio experto. Ambos enfoques pueden ser utilizados de manera complementaria.

El futuro es de todos

DNP
Departamento
Nacional de Planeación