

# Clasificación anticipada de textos: cuándo clasificar

---

LIC. JUAN MARTÍN LOYOLA

CONICET



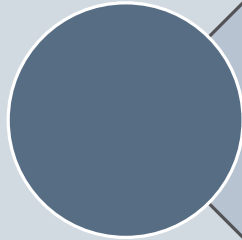
Universidad  
Nacional de  
San Luis

---

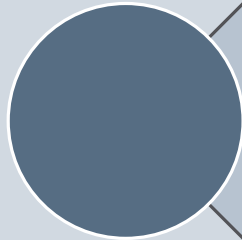
I M A S L

# Clasificación anticipada de textos

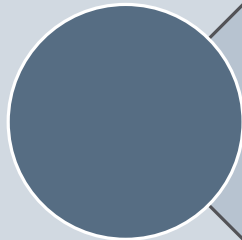
---



Modelo predictivo capaz de determinar la clase a la que pertenece un documento tan pronto como sea posible.



Documentos procesados secuencialmente.



El objetivo es intentar hacer predicciones confiables de la categoría de los textos con la mínima información.

# Motivación y posibles aplicaciones

---

## Motivación:

- Necesidad de respuesta rápida en entornos de *streaming*.
- Beneficio (científico, económico o social) de tener respuesta temprana de la categorización.

## Posibles aplicaciones:

- Detección de predadores sexuales en conversaciones de chat.
- Prevención de hostigamiento cibernético.
- Descubrimiento de tópicos que se convertirán en tendencia en redes sociales.
- Detección de discurso suicida.

# Objetivos

---

1

Primera aproximación al problema de la clasificación anticipada de textos.

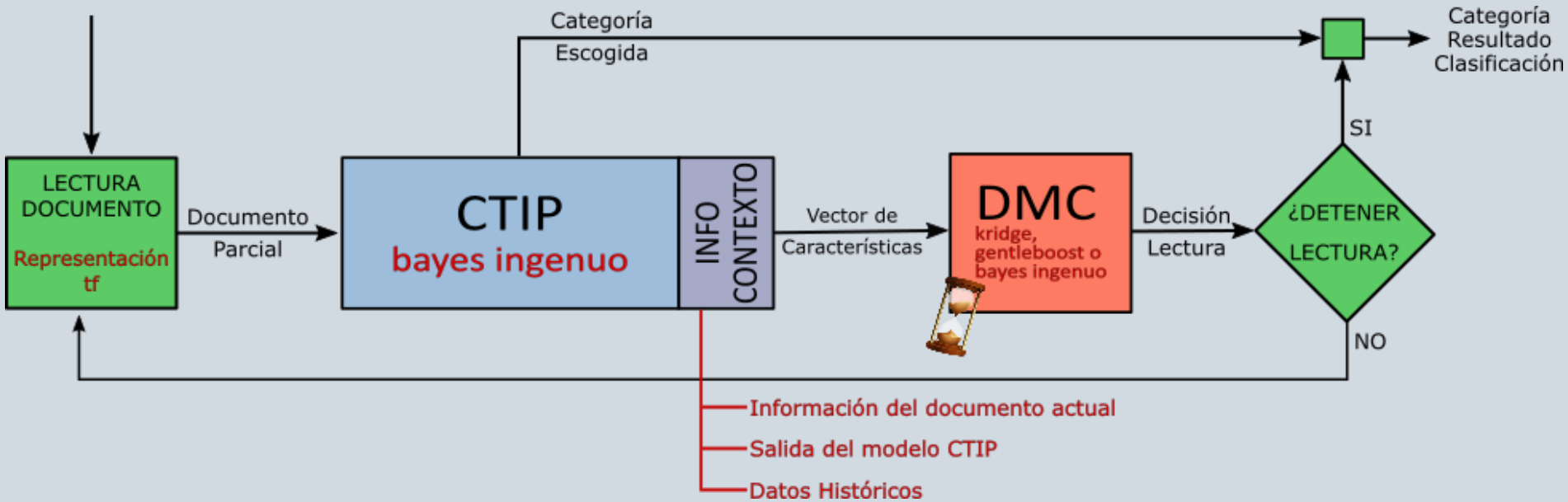
2

Definir un enfoque simple para los problemas que forman parte de la clasificación anticipada de textos.

3

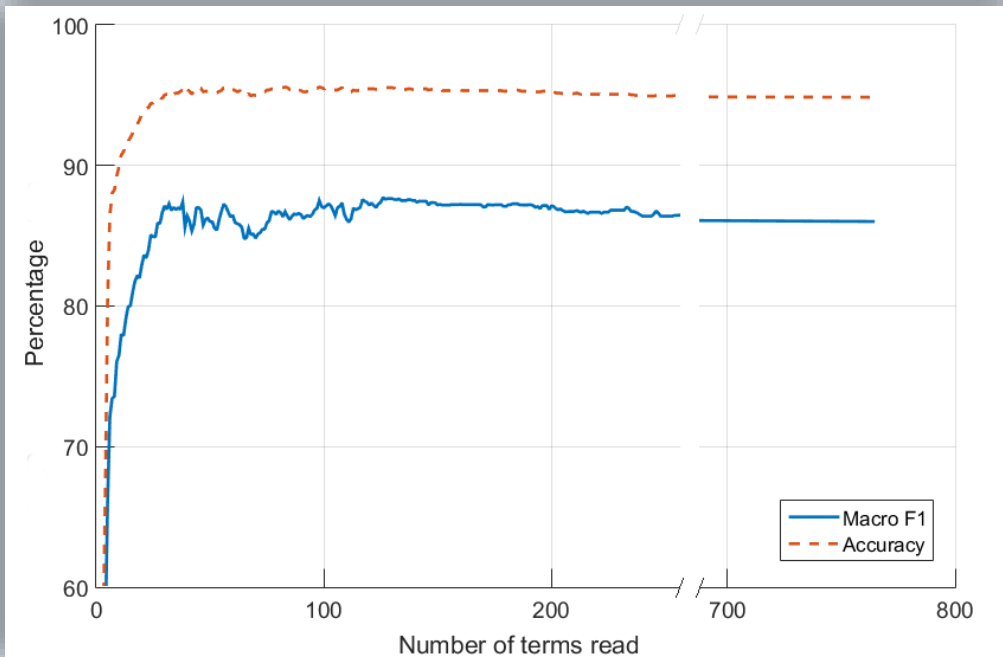
Comparar el desempeño con respecto a un clasificador estándar.

# Método



CTIP = Clasificador de Texto con Información Parcial  
DMC = Decisión del Momento de la Clasificación

# Resultados CTIP y DMC



- No es necesario procesar todo el documento para llegar al punto de “saturación”.
- A veces, la lectura completa de los documentos trae una disminución en el desempeño.
- El ensamble de aprendices débiles (redes neuronales) muestra un mejor desempeño.

Modelo	Precisión	Exhaustividad	Medida $F_1$
Kridge	54,46 %	57,52 %	55,94
Bayes Ingenuo	56,19 %	56,79 %	56,48
<b>Gentleboost</b>	<b>60,12 %</b>	<b>78,87 %</b>	<b>68,23</b>

# Resultados modelo lineal versus modelo temporal

---

Tipo de Modelo	Medida $F_1$	Promedio de Términos no Leídos	$EDT_{10}^*$
Estándar	85,97	0	0,73
Temporal	78,99	41,21	0,57

- Reducción en el desempeño (Medida  $F_1$ ) frente al modelo lineal relacionada a no procesar la totalidad del documento.
- Disminución considerable de la cantidad de términos leídos.
- Valor de medida temporal muy favorable.

# Conclusiones

---

01

Se definió un marco de referencia para la clasificación anticipada de textos de etiqueta única.

02

Se construyó un modelo base sobre el corpus R8 para comparar nuevos enfoques de clasificación de textos anticipada.


03

Se generalizó la medida de desempeño temporal para problemas multi-clase.



# Trabajos a futuro

---



Análisis impacto de diferentes representaciones de documentos sobre las distintas partes de la arquitectura.

Evaluar el desempeño de la arquitectura usando modelos clásicos de aprendizaje automático como las máquinas de soporte vectorial.

Utilizar otros corpus de datos para evaluar el modelo, por ejemplo el creado para la competencia ERISK 2017 (<http://early.irlab.org/>).

Analizar la medida de desempeño temporal para problemas de distinta características (cantidad de clases, cantidad de documentos, etc.).

# ¡Gracias!

---

## Directores:

- Dr. Marcelo Luis Errecalde – Universidad Nacional de San Luis
- Dr. Hugo Jair Escalante - Instituto Nacional de Astrofísica, Óptica y Electrónica (Puebla, México)
- Dr. Manuel Montes y Gomez - Instituto Nacional de Astrofísica, Óptica y Electrónica (Puebla, México)

## Información de contacto:

- Lic. Juan Martín Loyola
- Universidad Nacional de San Luis
- [jmloyola@outlook.com](mailto:jmloyola@outlook.com)

# Trabajos relacionados

---

- Dulac-Arnold, G., Denoyer, L., & Gallinari, P. (2011, Abril). Text Classification: A Sequential Reading Approach. En *ECIR* (pp. 411-423).
- Escalante, H. J., Montes-y-Gómez, M., Villaseñor-Pineda, L., & Errecalde, M. L. (2015). Early text classification: a Naïve solution. *arXiv preprint arXiv:1509.06053*.
- Dachraoui, A., Bondu, A., & Cornuéjols, A. (2015, Septiembre). Early classification of time series as a non myopic sequential decision making problem. En *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 433-447). Springer, Cham.
- Losada, D. E., & Crestani, F. (2016, Septiembre). A Test Collection for Research on Depression and Language Use. En *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 28-39). Springer International Publishing.

# Referencias posibles aplicaciones

---

- Escalante, H. J., Villatoro-Tello, E., Juárez, A., Montes-y-Gómez, M., & Pineda, L. V. (2013, June). Sexual predator detection in chats with chained classifiers. En *WASSA@ NAACL-HLT* (pp. 46-54).
- Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of Textual Cyberbullying. *The Social Mobile Web*, 11(02).
- Benhardus, J., & Kalita, J. (2013). Streaming trend detection in twitter. *International Journal of Web Based Communities*, 9(1), 122-139.
- Pestian, J. P., Matykiewicz, P., & Grupp-Phelan, J. (2008, June). Using natural language processing to classify suicide notes. En *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing* (pp. 96-97). Association for Computational Linguistics.

# Referencias modelos

---

- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1), 1-47.
- Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2), 337-407.
- Robert, C. (2014). Machine Learning, a Probabilistic Perspective, 493-495.
- Escalante, H. J., Montes-y-Gómez, M., Villaseñor-Pineda, L., & Errecalde, M. L. (2015). Early text classification: a Naïve solution. *arXiv preprint arXiv:1509.06053*.

# Modelo ganador competencia ERISK (ERDE 50)

---

- M. L. Errecalde, M. P. Villegas, D. G. Funez, M. J. Garciarena Ucelay, y L. C. Cagnina (2017). Temporal Variation of Terms as concept space for early risk prediction.
- M. P. Villegas, D. G. Funez, M. J. Garciarena Ucelay, L. C. Cagnina, y M. L. Errecalde (2017). LIDIC - UNSL's participation at eRisk 2017: Pilot task on Early Detection of Depression Notebook for eRisk at CLEF 2017.

# Corpus de datos R8

---

- Documentos que aparecieron en el servicio de noticias de Reuters y que fueron categorizados manualmente por el personal de Reuter Ltd.
- Ocho clases distintas.
- Pre-procesado.
- Distribución de documentos entre clases bastante sesgada.
- Longitud de documentos muy diversas.

# Pre-procesado corpus de datos

---

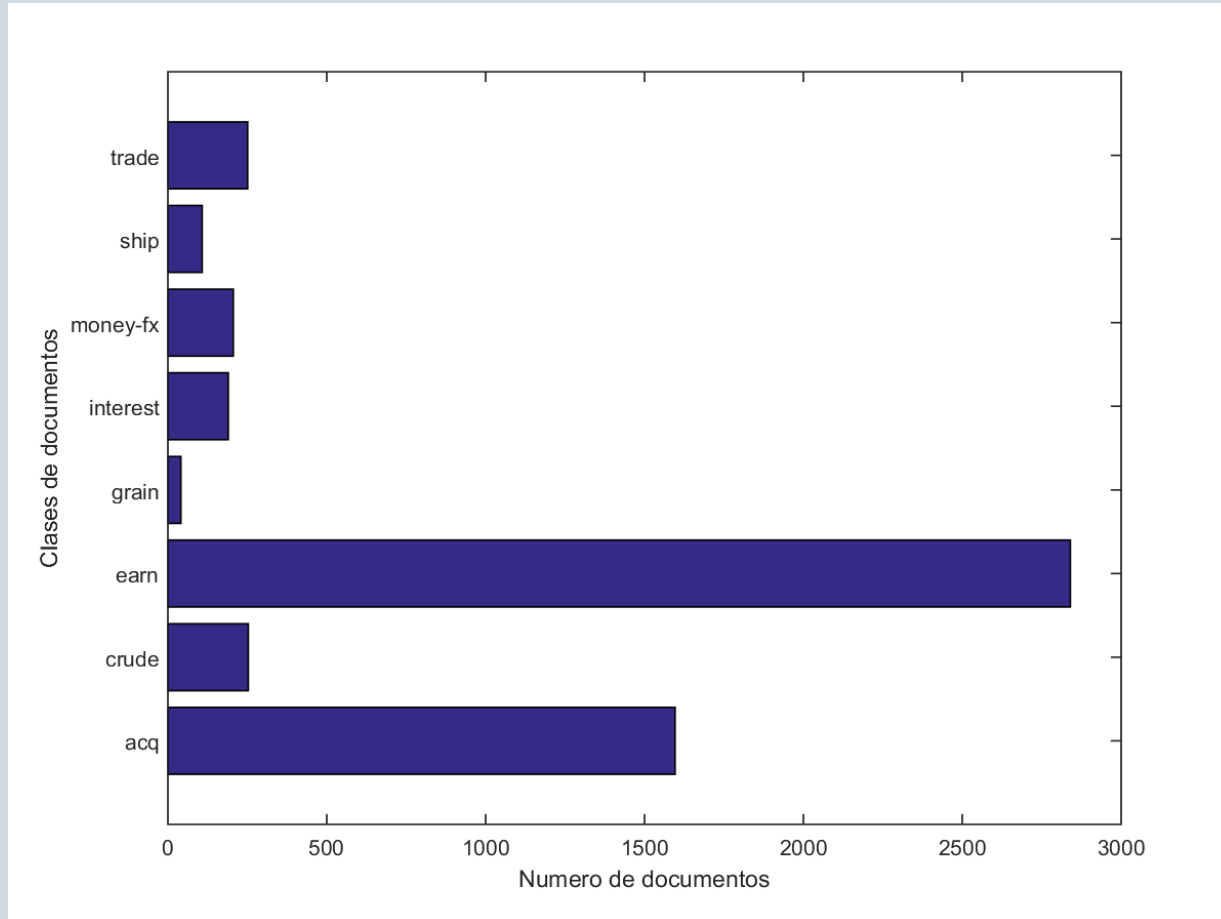
1. Sustituir los caracteres de TAB, NUEVA-LINEA y RETORNO por un carácter de ESPACIO;
2. Mantener las letras únicamente, es decir, convertir los caracteres de puntuación, números, etcétera en un carácter de ESPACIO;
3. Pasar todas las letras a minúscula;
4. Transformar múltiples caracteres de ESPACIO consecutivos en un único carácter de ESPACIO;
5. El título y subtítulo de cada documento se agrega al comienzo del documento.



# Corpus de datos R8

## Número de documentos de cada clase para entrenamiento

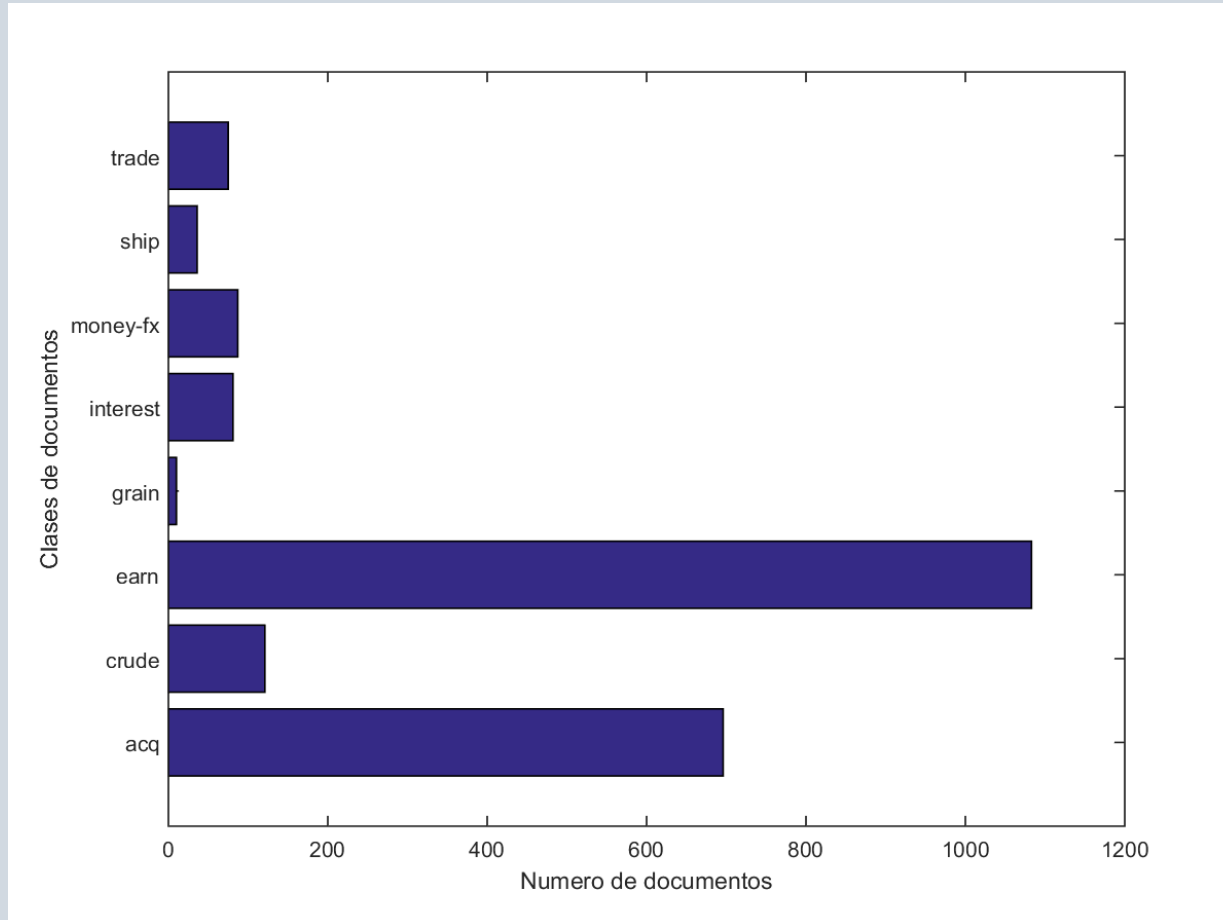
---



# Corpus de datos R8

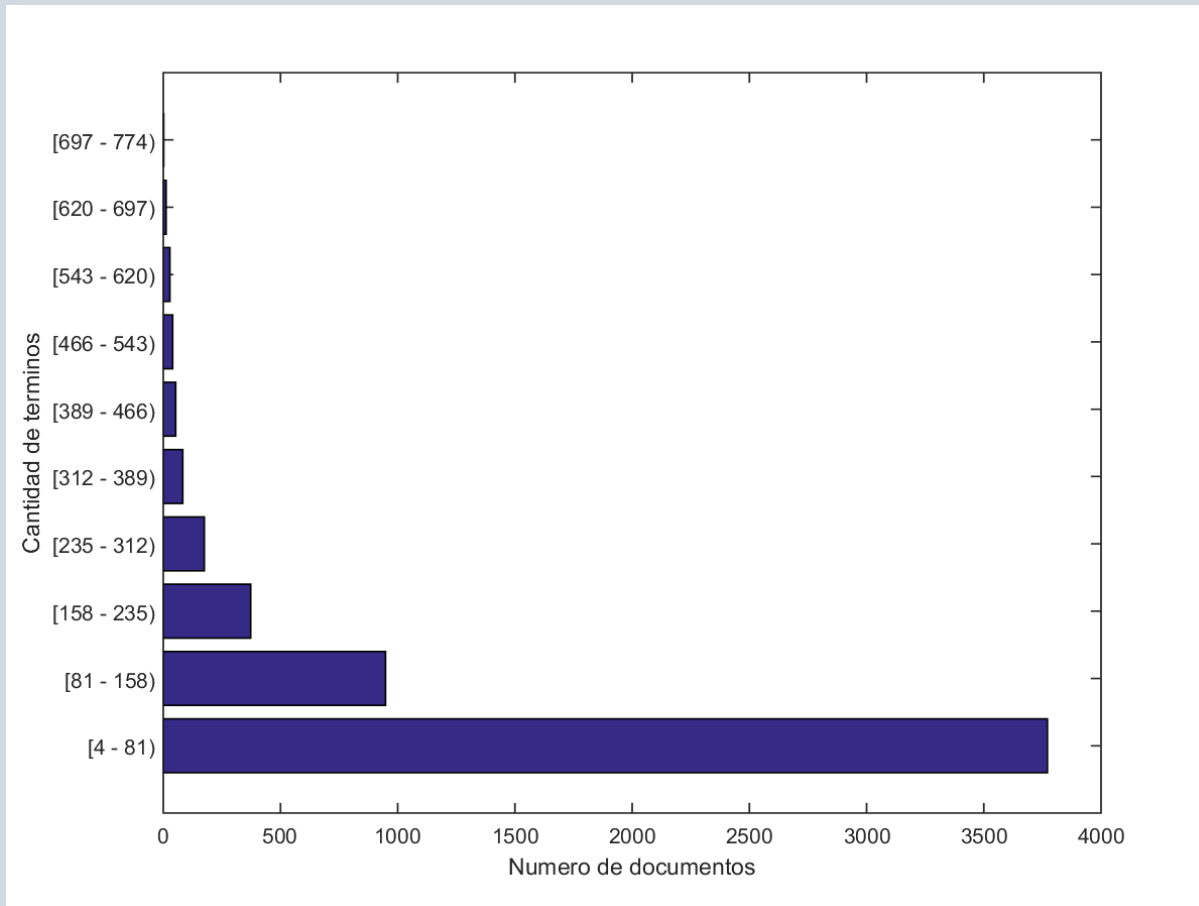
## Número de documentos de cada clase para testeo

---



# Corpus de datos R8

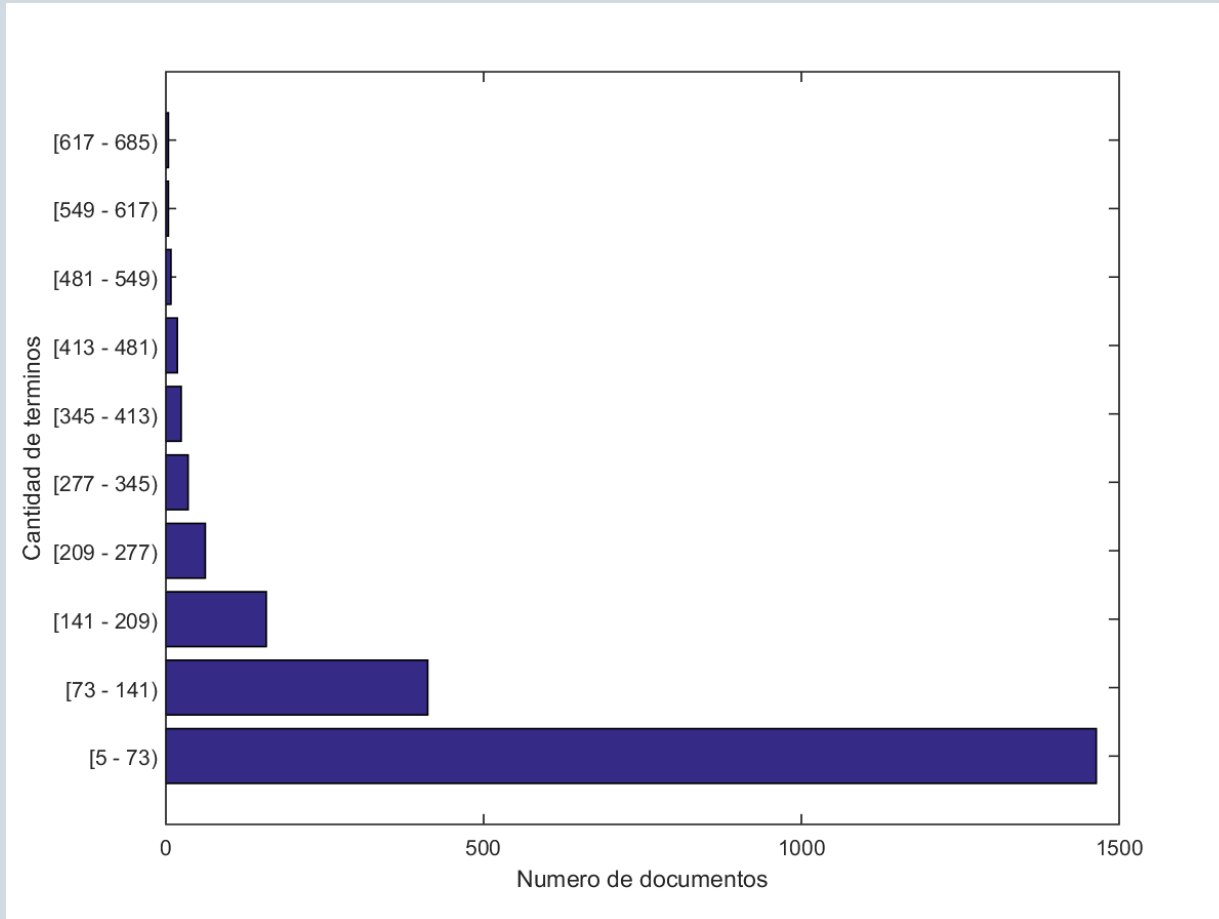
## Cantidad de términos para cada documento en entrenamiento



# Corpus de datos R8

## Cantidad de términos para cada documento en testeo

---



# Medidas de desempeño que consideran el tiempo

---

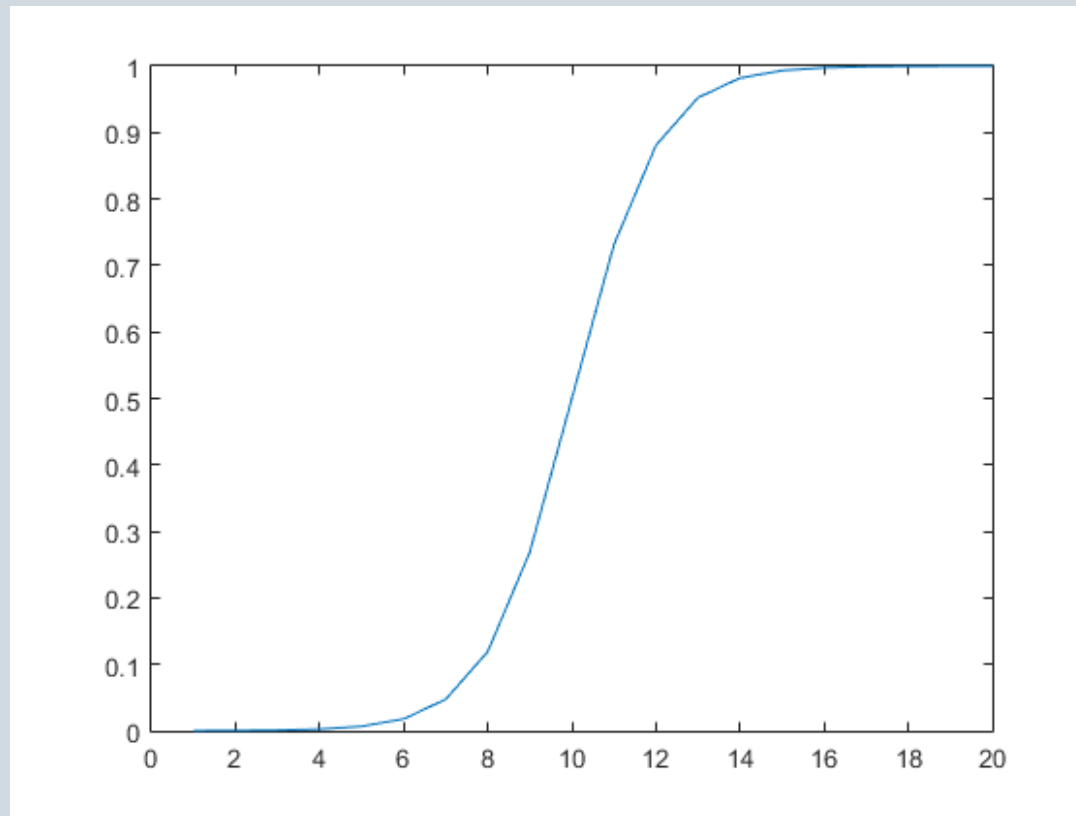
$$ERDE_o(d, k) = \begin{cases} c_{fp} & \text{cuando la decisión } d \text{ es positiva pero es incorrecta} \\ c_{fn} & \text{cuando la decisión } d \text{ es negativa pero es incorrecta} \\ lc_o(k) \cdot c_{vp} & \text{cuando la decisión } d \text{ es positiva y es correcta} \\ 0 & \text{cuando la decisión } d \text{ es negativa y es correcta} \end{cases}$$

Donde  $d$  es el documento actual,  $k$  es el tiempo (medido en cantidad de términos),  $c_{fp}$ ,  $c_{fn}$  y  $c_{vp}$  son los costos asociados a una decisión FP, FN, VP respectivamente. El factor  $lc_o(k) \in [0,1]$  representa el costo asociado al retraso en detectar los verdaderos positivos.

$$lc_o(k) = 1 - \frac{1}{1 + e^{k-o}}$$

# Ejemplo $lc_o(k)$ con $o = 10$

---



# Adaptación medida de desempeño temporal para más de dos clases

---

$$\text{ERDE}_o(d, k, i) = \begin{cases} c_{fp}^i & \text{if it is a } FP_i \\ c_{fn}^i & \text{if it is a } FN_i \\ lc_o(k) \cdot c_{tp}^i & \text{if it is a } TP_i \\ 0 & \text{if it is a } TN_i \end{cases}$$

$$\text{EDE}_o(d, k) = \sum_{i=1}^{|\mathcal{C}|} \text{ERDE}_o(d, k, i)$$

$$\text{EDE}_o(d, k) = \begin{cases} lc_o(k) \cdot c_{tp}^i & \text{when the category } i \text{ chosen by CPI is correct} \\ c_{fn}^j + c_{fp}^l & \text{when the category } j \text{ chosen by CPI is incorrect} \\ & \text{and the category } l \text{ was correct.} \end{cases}$$

# Gentleboost (Gentle AdaBoost)

---

- La salida de los otros algoritmos de aprendizaje ("aprendices débiles") se combina en una suma ponderada que representa la salida final del clasificador potenciado.
- AdaBoost es adaptativo en el sentido de que los aprendices débiles posteriores son ajustados en favor de aquellas instancias mal clasificadas por los clasificadores anteriores.
- AdaBoost es sensible a datos ruidosos y valores atípicos.
- Gentleboost no minimiza el error de testeo de forma codiciosa sino que lo hace en pasos limitados.



# Kridge

---

- Kernel ridge regression combina la Ridge Regression (mínimos cuadrados lineales con la regularización de la norma  $l_2$ ) con el truco del kernel.