



27/10/2025

RECOMENDADOR EDITORIAL PARA HERNAN CASCIARI

Que más personas sigan leyendo y vuelvan

PRESENTED BY:

Gabriel Moreno, Adrian Suarez, Luis Rubiano, Juan Sebastian Sierra y Gianluca Cicco

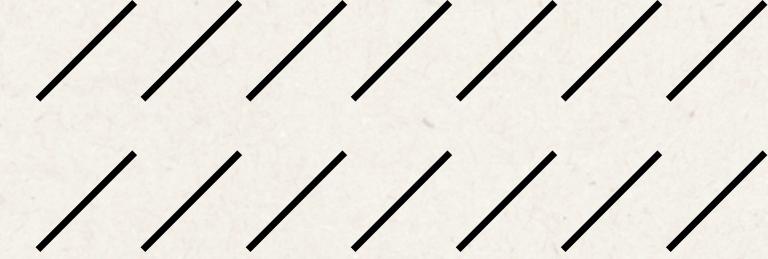


Ciencia de datos y
econometría
aplicada

Índice



| | |
|-----------|-----------------------|
| 01 | Introducción |
| 02 | Limpieza de los datos |
| 03 | TF-IDF |
| 04 | LDA |
| 05 | Recomendación |
| 06 | Evaluación |



Introducción



Hernán Casciari

Problema

- 520 cuentos de Casciari sin sistema de recomendación
- Usuarios navegan manualmente para encontrar contenido similar
- Baja retención y subutilización del catálogo
- Necesidad de automatizar descubrimiento de contenido

Solución

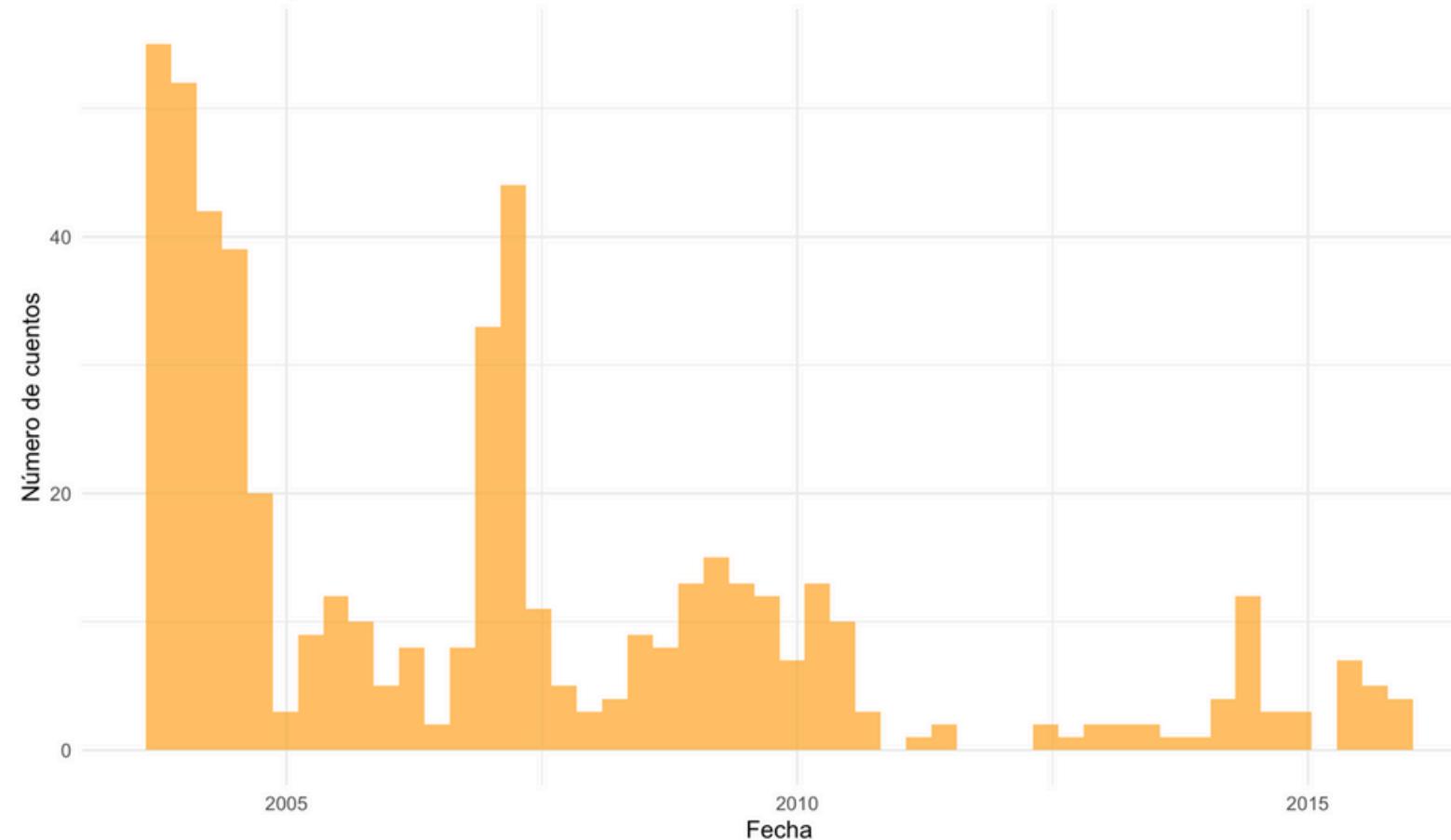
- Recomendador híbrido: TF-IDF (estilo) + LDA (temas)
- Top-5 por cuento con máxima similitud.
- Símbolos gráficos de afinidad léxica/temática
- Ajuste automático de pesos con A/B + Thompson sampling.
- Monitoreo constante para posible reentrenamiento
- Promoción de al menos un cuento nuevo

Hernán es un escritor argentino pionero del blog literario, y en esta transformación del mundo digital, la clave para las plataformas digitales está en la retención de audiencia.

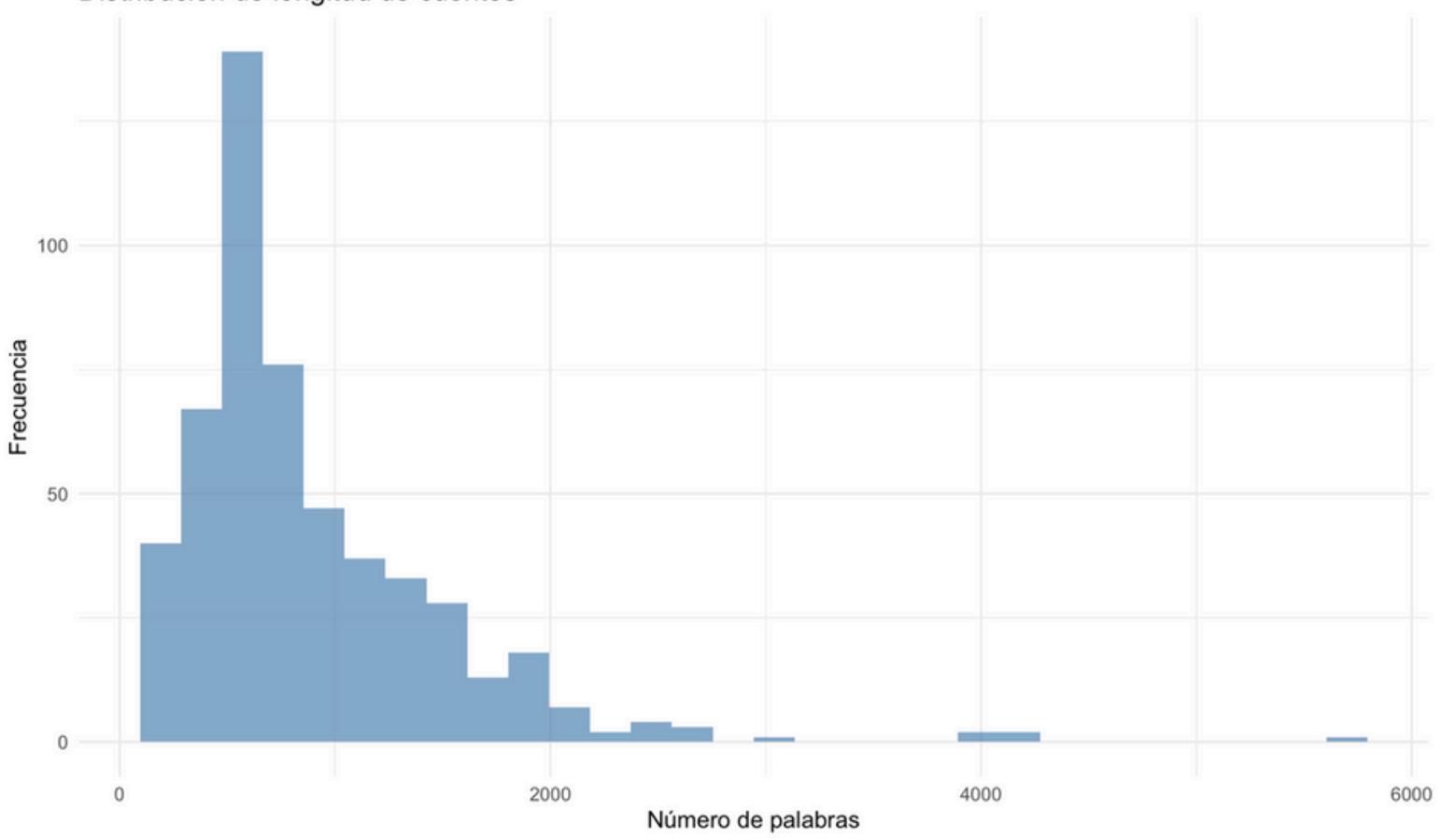
Limpieza de los datos

02/06

Distribución temporal de cuentos



Distribución de longitud de cuentos



Características

- 520 cuentos (2003-2015), promedio 882 palabras
- Distribución heterogénea: 95 - 5,603 palabras
- Boom inicial 2004-2005 (~200 cuentos), declive posterior

Pipeline

- Limpieza estándar + eliminación stopwords específicas argentinas
- Lematización (udpipe español)
- N-gramas combinados (uni + bi + trigramas)
- Filtro sparsity 90%

Desafíos

- 26 outliers largos (>4,000 palabras) pueden sesgar similitud
- Variabilidad temporal refleja evolución narrativa del autor

Resultado

- Matriz final: 520 documentos × 504 términos
- Reducción vocabulario: 97.7% ($22,232 \rightarrow 504$ términos)
- Términos frecuentes: "decir", "hacer", "él" → estilo oral característico

TF-IDF

Matriz de Similitud Coseno - TF-IDF

Messi es un perro y sus recomendaciones

| | Messi es un perro | Canelones | Mundos paralelos | Espectáculo volar | ¿Cuni qué? | Ni olvido perdón |
|-------------------|-------------------|-----------|------------------|-------------------|------------|------------------|
| Messi es un perro | 1 | 0.631 | 0.496 | 0.457 | 0.435 | 0.407 |
| Canelones | 0.631 | 1 | 0.485 | 0.428 | 0.395 | 0.227 |
| Mundos paralelos | 0.496 | 0.485 | 1 | 0.562 | 0.567 | 0.206 |
| Espectáculo volar | 0.457 | 0.428 | 0.562 | 1 | 0.508 | 0.304 |
| ¿Cuni qué? | 0.435 | 0.395 | 0.567 | 0.508 | 1 | 0.273 |
| Ni olvido perdón | 0.407 | 0.227 | 0.206 | 0.304 | 0.273 | 1 |

Metodología

- TF-IDF: Combina frecuencia del término (TF) + frecuencia inversa de documento (IDF)
- Penaliza términos ubicuos, prioriza palabras distintivas de cada documento
- Normalización L2: Controla efecto de longitud textual (documentos largos vs cortos)
- Similitud coseno: Mide ángulo entre vectores (0=diferentes, 1=idénticos)

Caso de Estudio: "Messi es un perro"

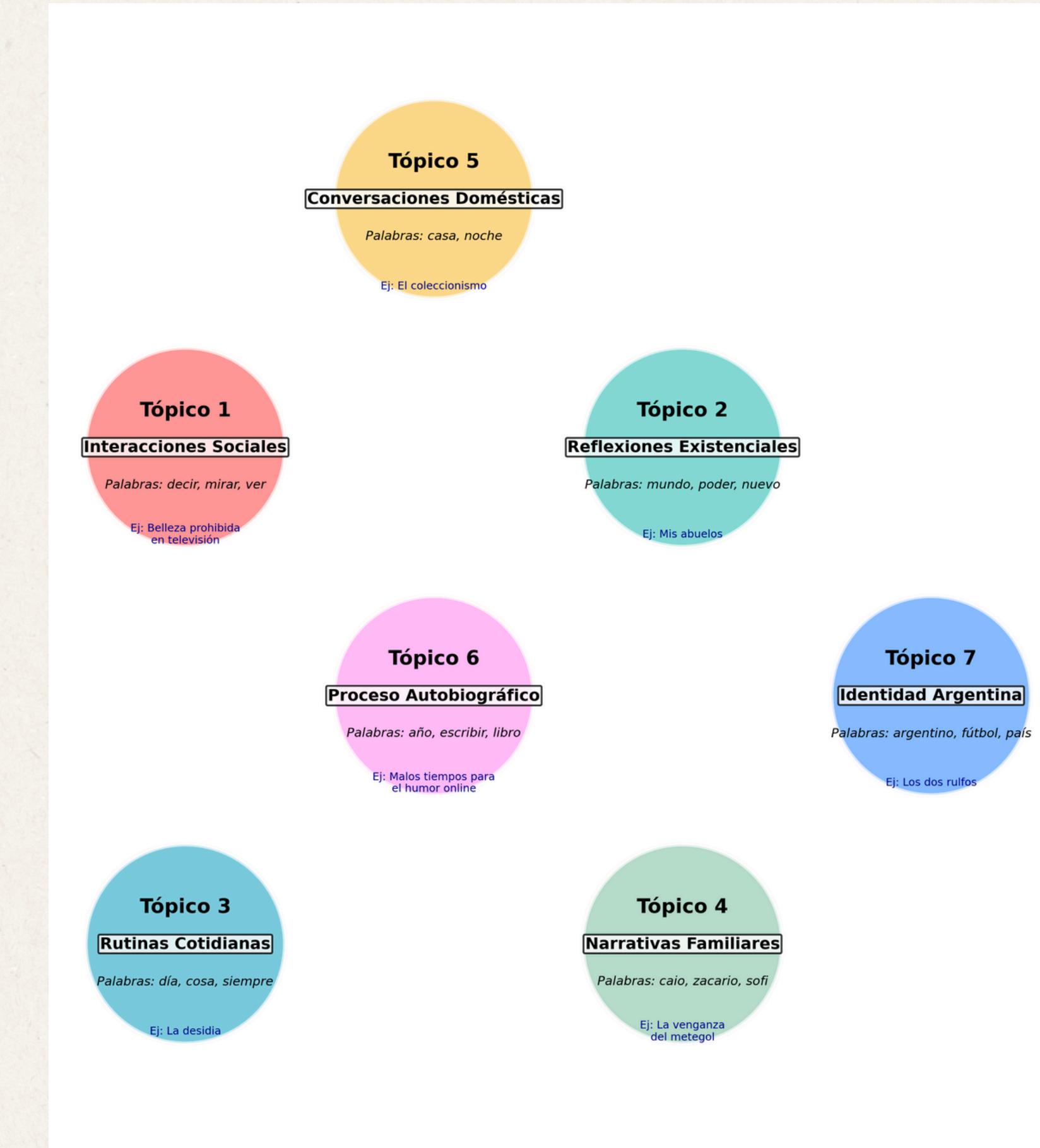
- Reflexión autobiográfica sobre permanecer en Barcelona por el fútbol
- Combina elementos familiares, futbolísticos y crisis económica
- Registro coloquial típico con argentinismos y referencias culturales específicas
- Narrativa introspectiva: decisiones personales vs circunstancias adversas

Metodología

- LDA: Modelo probabilístico que descubre tópicos latentes en documentos
- Cada documento = mezcla de tópicos subyacentes (distribuciones gamma)
- K=7 tópicos seleccionados por criterios de interpretabilidad
- Similitud coseno entre distribuciones de probabilidad (no palabras)

Tópicos

- **T1: Interacciones sociales** → "Belleza prohibida en televisión"
- **T2: Reflexiones existenciales** → "Mis abuelos"
- **T3: Rutinas cotidianas** → "La desidia",
- **T4: Narrativas familiares** → "La venganza del metegol"
- **T5: Conversaciones domésticas** → "El colecciónismo"
- **T6: Proceso autobiográfico-creativo** → "Malos tiempos para el humor online"
- **T7: Identidad argentina** → "Los dos rullos"



TABLA

04/06

| Posición | TF-IDF | | LDA | |
|----------|--|------------------|---------------------------------------|------------------|
| | Título del cuento | Similitud Coseno | Título del cuento | Similitud Coseno |
| 1 | <i>Messi es un perro</i> | 1 | <i>Messi es un perro</i> | 1 |
| 2 | <i>Canelones</i> | 0.63 | <i>Los dos comodines</i> | 0.96 |
| 3 | <i>Instrucciones para crear mundos paralelos</i> | 0.49 | <i>Donar los órganos</i> | 0.95 |
| 4 | <i>El espectáculo de volar</i> | 0.46 | <i>Futbol, fervor e independencia</i> | 0.95 |
| 5 | <i>¿Cuni... qué?</i> | 0.43 | <i>Los payasos</i> | 0.94 |
| 6 | <i>Ni olvido ni perdón</i> | 0.41 | <i>¿Cuni... qué?</i> | 0.94 |

Modelo híbrido

Al fusionar ambos modelos, el sistema logra encontrar recomendaciones que sean parecidas en la forma de escritura del autor y a la vez diversas en contenido, equilibrando forma y fondo para ofrecer resultados más naturales y representativos de su obra

Reglas de negocio

- No auto-recomendación
- Cobertura temática: en top-5, al menos 2 tópicos distintos.
- Descubrimiento guiado: reservar 1 lugar en top-5 para temas afines con léxico distinto.

Similitud léxica:

$$s_{\text{lex}}(i | q) = \cos(x_i, x_q),$$

Mide qué tan parecidas son las palabras usadas.

Similitud temática:

$$s_{\text{lda}}(i | q) = \cos(\gamma_i, \gamma_q)$$

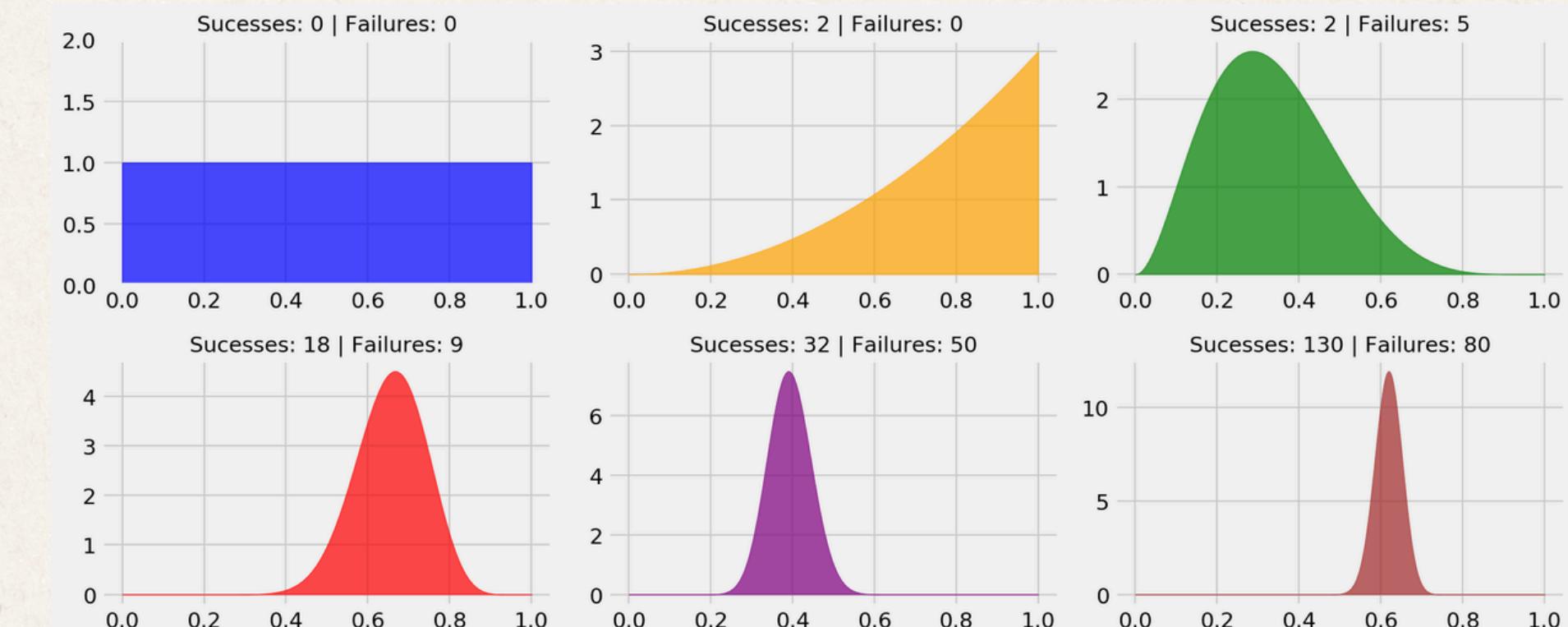
Mide qué tan parecidos son los temas o tópicos.

Mezcla ponderada

$$s(i | q) = \alpha s_{\text{lex}}(i | q) + (1 - \alpha) s_{\text{lda}}(i | q),$$
$$\alpha \in [0, 1]$$

Equilibra forma y fondo en la recomendación

- **KPI primario:** tasa de conversión de lectura
- **Unidad de análisis:** sesión
- **Diseño:** experimentos A/B controlados
- **Asignación de (α):** Thompson Sampling por sesión, actualización bayesiana con lecturas observadas
- **Criterio de éxito:** convergencia al (α) que maximiza
- **Validación cruzada** para calibrar (α) antes del experimento online



$$CTR_{\text{lectura}} := \frac{\text{Recomendaciones que el usuario efectivamente lee}}{\text{Recomendaciones totales hechas a un usuario}}$$