Universidad de los Andes
Facultad de Economía

Ciencia de Datos y Econometría Aplicada
ECON-3785
Créditos: 3
2025-2

Taller 3:
Recomendador editorial para el Blog de Hernán Casciari

Contexto

Hernán Casciari escribe historias breves que cruzan fútbol, infancia y familia con humor y ficción. Ese sello le valió en 2005 el premio de Deutsche Welle al "Mejor blog del mundo".

Orsai Media nos convoca como equipo consultor para decidir cómo **recomendar el siguiente cuento** al lector. La meta es simple y medible: que más personas sigan leyendo y vuelvan.

Qué esperan de nosotros

- Mapear **temas** del corpus y explicar el catálogo a un público no técnico.
- Probar dos estrategias de **recomendación por contenido** (léxica vs. temática) y **recomendar** cuál implementar en producción, justificando la decisión.
- Proponer **criterios operativos** para mejorar la experiencia del lector (diversidad, balance temático, etc.).
- Proponer un **mecanismo de evaluación** para comparar ambas formas de recomendación (KPI, unidad de análisis y criterio de decisión).
- Entregar **informe ejecutivo** y **slides** para la mesa.

Disponemos del archivo `blog_casciari.csv` con `titulo`, `fecha` y `texto` completo de cada cuento.

Entrega final

- **Informe escrito** (máx. **6 páginas**, sin anexos): describir el pipeline elegido, los dos enfoques de recomendación (léxico vs. temático), resultados clave y la **propuesta de mecanismo de evaluación** (KPI, unidad de análisis y criterio de decisión).
- **Una sola presentación por equipo** (máx. **6 slides**): narrativa ejecutiva para la mesa directiva, con (i) problema y métrica de éxito, (ii) mapa/etiquetas de temas con 1–2 ejemplos, (iii) *tabla única* Top-5 (léxico vs. temas) y decisión para producción, (iv) *mockup* del módulo recomendado, (v) **mecanismo de evaluación** propuesto (KPI, unidad de análisis y criterio de decisión).
- **Presentación en clase**: 15 minutos por equipo.

Formato de entrega:

Universidad de los Andes
Facultad de Economía

Ciencia de Datos y Econometría Aplicada
ECON-3785
Créditos: 3
2025-2

- `informe_equipo_XX.pdf` (reemplazar `XX` por el número del equipo con dos dígitos)
- `slides_equipo_XX.pdf`

Guía de trabajo (orientativa)

Esta guía sugiere buenas prácticas para que el resultado sea sólido y claro. No indica un único camino ni fija decisiones técnicas por ustedes.

1) Datos y preparación

- Documenten criterios de limpieza de texto: textos vacíos, duplicados, outliers de longitud; dejen constancia de cuántos casos afectan.
- Identifiquen patrones específicos del corpus: ¿hay palabras muy frecuentes propias de Casciari que no aportan semánticamente? ¿expresiones o nombres que se repiten?
- Elijan **un cuento** como *consulta* para comparar recomendadores; justifiquen brevemente por qué es representativo.

2) Pipeline

- Mantengan el pipeline **lo más simple posible** que aún permita buen desempeño (minúsculas, stopwords ES, lematización; n-gramas; manejo de acentos/tildes y caracteres especiales del español; creación de stopwords específicas del dominio (nombres propios recurrentes, muletillas del autor)).
- Si crean *stopwords de dominio* (nombres muy frecuentes, muletillas), dejen constancia.

3) Dos enfoques de recomendación

- **Léxico**: vectorizar los documentos (mediante una matriz de frecuencia de palabras por documento o TF-IDF) y aplicar la similitud coseno.
- **Temático**: LDA $\rightarrow$ representación documento-tema + similitud coseno.
- Presenten una **única tabla** con Top-5 de ambos enfoques para el mismo cuento; agreguen **1–2 frases** sobre diferencias relevantes (diversidad, tono, contexto).

Universidad de los Andes
Facultad de Economía

Ciencia de Datos y Econometría Aplicada
ECON-3785
Créditos: 3
2025-2

4) Elección de k (LDA)

- Exploren pocos valores de k (p. ej., 3, 5, 7, 10) y muestren un gráfico sencillo (coherencia u otro criterio razonable).
- Expliquen la decisión final con **interpretabilidad**: etiquetas claras y ejemplos breves de fragmentos.

5) Recomendación para producción

- Con base en los resultados obtenidos, **recomienden cuál enfoque implementar** (léxico, temático o híbrido) para el sistema de recomendaciones del blog.
- Justifiquen la decisión considerando: calidad y diversidad de las recomendaciones, facilidad para explicar los resultados al equipo de Orsai, y qué tan bien captura el estilo narrativo de Casciari.
- Propongan 2–3 **reglas de negocio** para la implementación (p. ej., diversidad de temas, evitar duplicados, antigüedad de los cuentos).
- Identifiquen **riesgos principales** del enfoque elegido (temas dominantes, *cold start*, sesgos por longitud) y sugieran cómo mitigarlos.

6) Mecanismo de evaluación

- Definan **un KPI primario** y la **unidad de análisis** (sesión/usuario) con breve justificación. Sugerencias:
  - **CTR del módulo**: $\text{CTR} = \frac{\text{clics en recomendaciones}}{\text{impresiones del módulo}}$ (unidad: sesión).
  - **Profundidad de lectura**: número de cuentos adicionales leídos tras ver el módulo (unidad: sesión).
  - **Tiempo en página** del primer cuento recomendado (reporten mediana por sesión).
- ¿Cómo validarían en la práctica cuál recomendador funciona mejor? Describan su enfoque de manera formalmente precisa y completa.