Análisis de Reseñas de Letterboxd

Topic Modeling y Exploración de Datos No Estructurados

Alfredo Martín López · Marcos Fernández Arévalo · Pedro Pérez Blanco

Máster en Business Analytics (MUBA) — Universidad Pontificia Comillas ICADE
Datos No Estructurados — Curso 2024-2025

Agenda

01 Contexto: ¿Qué es Letterboxd?

02 Problema de negocio y objetivos

03 Dataset: descripción y filtrado

04 Librerías y herramientas utilizadas

05 Pipeline de preprocesamiento

06 Análisis exploratorio (Word Cloud, N-gramas, TTR)

07 Modelo LDA — Topic Modeling

08 Resultados y visualizaciones

09 Insights de negocio

10 Conclusiones y trabajo futuro

¿Qué es Letterboxd?

Red social de cine

Usuarios escriben reseñas y puntúan (0.5–5 ★)

Texto libre + Rating

Combina opinión escrita con valoración numérica

Comunidad única

Tono informal, ironía, memes y jerga cinéfila

→ Oportunidad: cruzar lo que el usuario ESCRIBE con lo que PUNTÚA

Problema de Negocio

El problema

- Plataformas y distribuidoras necesitan monitorizar la recepción de contenidos
- Leer miles de reseñas manualmente no escala
- Las estrellas no capturan el "por qué" de la opinión

Nuestra propuesta

- Aplicar NLP para extraer temas y patrones automáticamente
- Topic modeling para descubrir de qué habla la audiencia
- Generar insights accionables para marketing y contenido

Objetivos del Proyecto

1

Preprocesar y explorar

Limpiar y transformar un corpus real y multilingüe de reseñas

2

Identificar tópicos latentes

Aplicar LDA para descubrir de qué habla la audiencia de forma automática

3

Evaluar calidad

Medir la calidad del preprocesamiento con métricas como TTR y coherence score

El Dataset — De Crudo a Filtrado

Comparación del dataset

	Dataset original (HuggingFace)	Dataset filtrado (estrenos de 2024)
Fuente	HuggingFace (scraping Letterboxd)	Mismo, filtrado por año de estreno
Nº reseñas	hasta 10 por película (no contabilizadas)	119.096 → 111.798 (dedup.) muestra analizada: 11.968
Nº películas	847.209	8.470 en la muestra (40.486 estrenadas en 2024)
Idioma	Multilingüe	48 idiomas · inglés 62,4 %
Variables	title, year, reviews, average_rating, …	title, year, average_rating, review_text, likes
Mediana palabras	no medida	16 (en la muestra)

Desafíos: reseñas cortas (mediana 16 palabras) · 48 idiomas · ironía y memes. «2024» es el año de ESTRENO de la película, no la fecha de la reseña.

Librerías y Herramientas

Preprocesamiento

NLTK · spaCy (en_core_web_sm) · regex

Modelado de tópicos

Gensim (LdaModel, determinista) · pyLDAvis

Visualización

Matplotlib · WordCloud · NetworkX

NLP avanzado

Flair (biLSTM POS tagger) · scikit-learn (TF-IDF)

Datos

Pandas · NumPy · HuggingFace Datasets · Parquet

Entorno

Python 3.13 · Jupyter · requirements.txt con versiones fijadas

Pipeline de Preprocesamiento

1

Normalización

Minúsculas, eliminar números, emojis, puntuación

→

2

Tokenización

NLTK word_tokenize

→

3

Stopwords

Inglés + industria + multilingüe

→

4

Lematización

spaCy + correcciones custom

→

5

Corpus limpio

Formato Parquet 11.968 docs (muestra)

Limpieza del texto paso a paso

- Normalización: convertimos todo a minúsculas y eliminamos números, emojis, puntuación y caracteres especiales
- Tokenización: separamos el texto en palabras individuales con NLTK word_tokenize
- Lematización con spaCy + correcciones manuales para agrupar sinónimos (scary/spooky → horror, funny/hilarious → comedy)

Selección de Stopwords Personalizadas

Stopwords estándar (inglés)

the, is, at, which, on, a, an, and, or, but, in, with, to, for, of, it, this, that...

Palabras funcionales del idioma inglés sin valor semántico

Stopwords de industria (cine)

film, movie, show, watch, see, cinema, series, episode, scene, actor, director, screen, review, cast, feature, documentary, short

Términos del sector que aparecían en casi todas las reseñas sin aportar distinción

Stopwords multilingües

que, una, del, con, por, para, pero, más, todo, pelicula, le, la, une, pour, est, dan, der, das, die, bir, pas, non, mai, son, qui, uma

Palabras funcionales en español, francés, alemán, turco, portugués, italiano… que se colaban en el dataset

Part-of-Speech Tagging — 3 Métodos

Etiquetamos gramaticalmente cada palabra (sustantivo, verbo, adjetivo…) para entender la estructura del lenguaje y extraer adjetivos relevantes.

spaCy — MLP

Red neuronal multicapa

Rápido, buena precisión general. Usa el modelo en_core_web_sm preentrenado.

NLTK — Perceptron

Perceptrón promediado

Ligero y eficiente. Ideal para extracción masiva de adjetivos con paralelización (joblib).

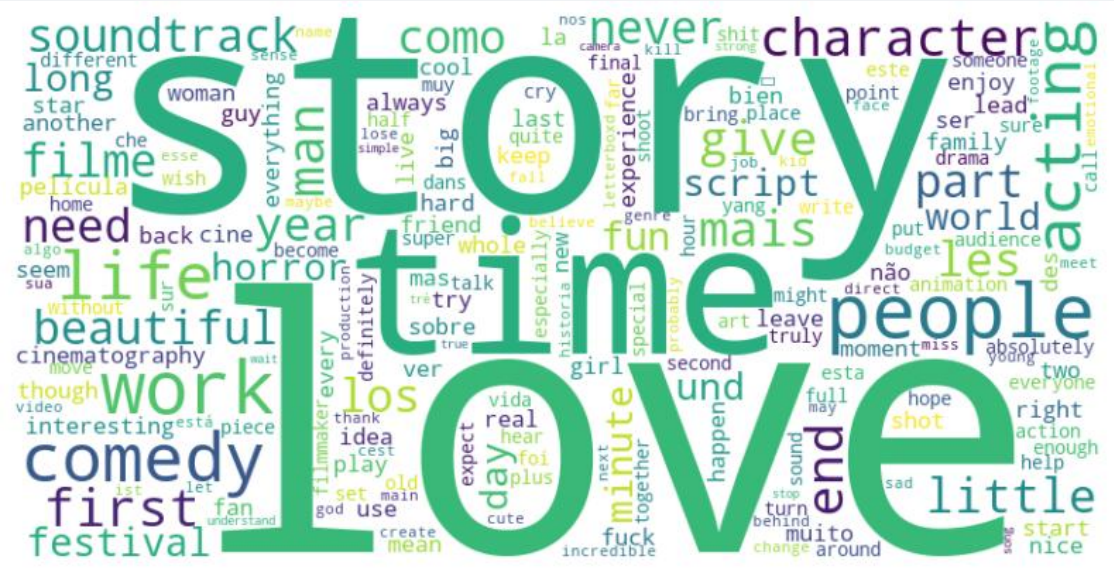
Flair — biLSTM

Embeddings contextuales char-level

Mayor precisión con ambigüedad semántica. Usa contexto bidireccional para resolver polisemia.

Resultado: extrajimos adjetivos con Perceptron → Top: good, great, beautiful, interesting, short, little, big, real

Nube de Palabras — Post-preprocesamiento

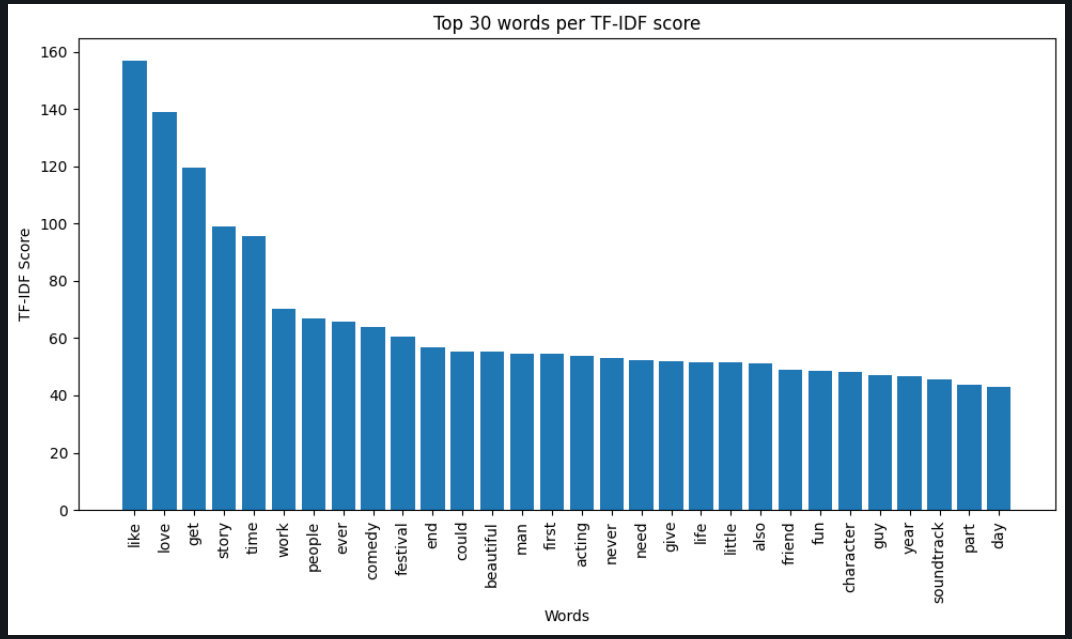


Hallazgos

- story, character, acting, script, soundtrack → la gente habla del trabajo en la película
- A simple vista se ven palabras que no son inglesas: como, mais, les, filme, película, não, muito, und, los. El 37,6 % del corpus no está en inglés, y se nota aqui antes que en ninguna métrica
- people, man, girl → hablan de personajes, con "man" mucho más grande que "woman"
- beautiful, interesting, real, special → adjetivos que van más allá de bueno/malo

Palabras como actor, director, film fueron eliminadas como stopwords de industria — su gran tamaño previo confirmó la decisión

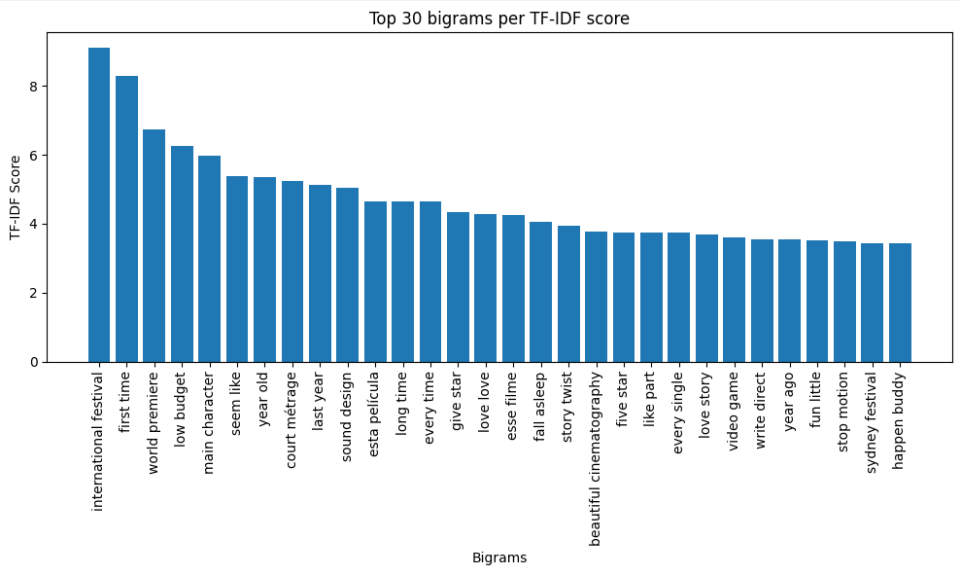
Análisis de Unigramas — TF-IDF



Top palabras por TF-IDF

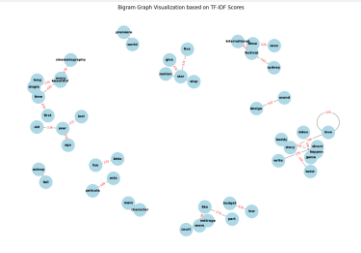
- "like" domina pero puede indicar semejanza, no solo gusto
- "love" → películas románticas o cosas que la gente amó
- Se comenta la historia más que los personajes o la actuación
- Se menciona más a hombres que a mujeres: man (14) y guy (26) frente a woman (48), girl (63)
- El tono informal sigue ahí, pero fuera del top-30: fuck en el puesto 40 y shit en el 69. No se filtra nada, simplemente no destacan
- De los géneros solo comedy destaca (puesto 9); horror queda en el 32, drama en el 170 y action en el 180

Análisis de Bigramas — TF-IDF



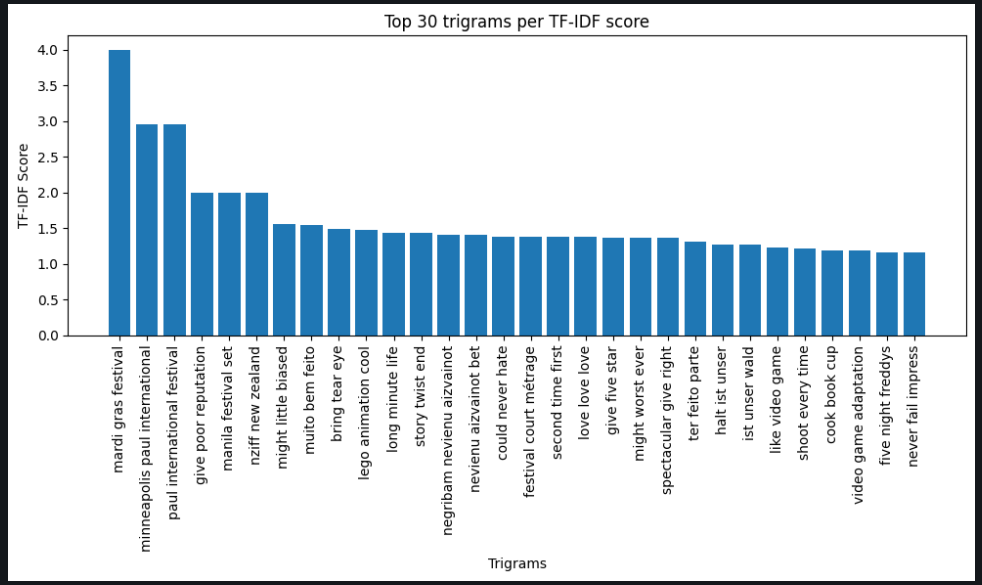
Bigramas más relevantes

- "international festival" y "world premiere" encabezan ahora la lista
- "year old" → edad de la película o nostalgia
- "main character" → el protagonista domina
- "love story" (23) sí; "true story" cae al puesto 84
- "low budget" (4) y "sound design" (10): aspectos técnicos, muy arriba
- Los tres primeros son de festival: "international festival" (1), "world premiere" (3), "low budget" (4)



Red de bigramas (NetworkX) → clusters temáticos emergentes: género, experiencia personal, dónde/cómo se vio la película

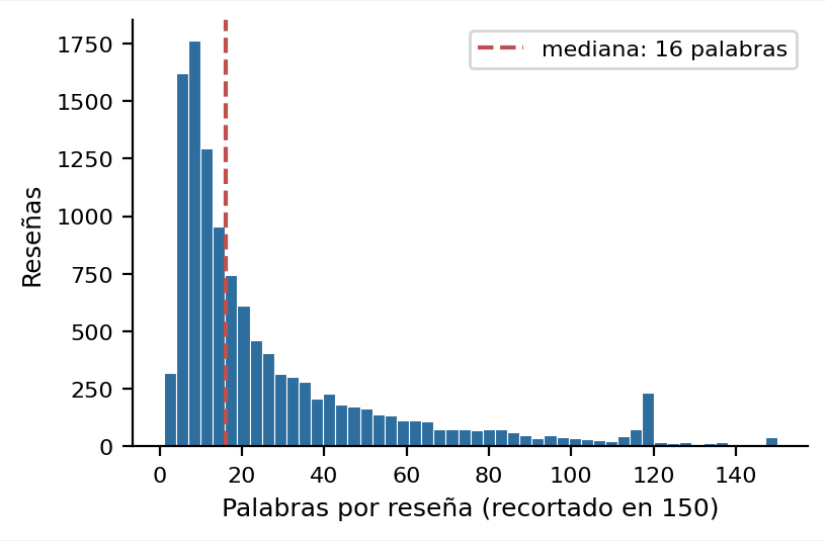
Análisis de Trigramas — TF-IDF

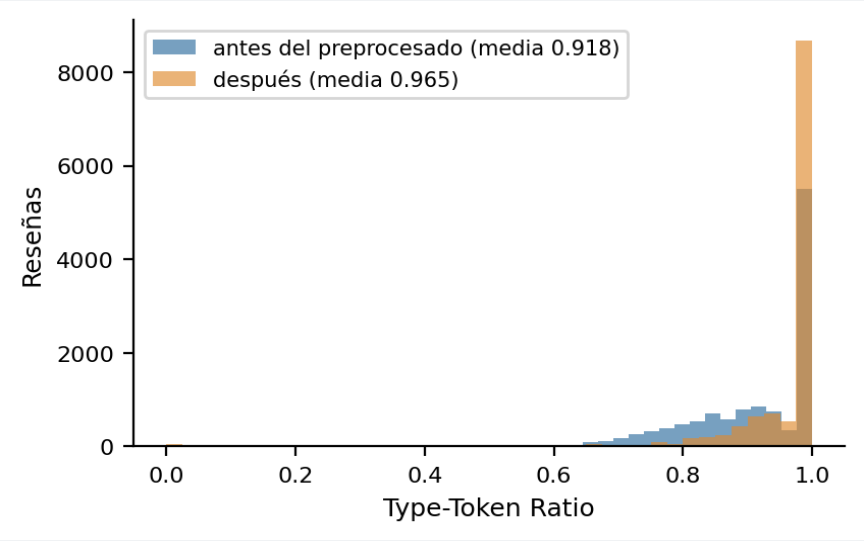


Hallazgos en trigramas

- En esta muestra los trigramas los dominan nombres de festivales («minneapolis paul international», «mardi gras festival»).
- Los memes que aparecían antes («close enough welcome», «fuck ass bob») siguen en el corpus, pero fuera del top: puestos 6.927 y 17.559. No se censura nada; simplemente no destacan en esta muestra.
- Festivales: "international festival", "north american premiere"
- Características de película: "base true story"
- Curiosidad: "dead poet society" entre los más frecuentes

Análisis Adicional — Longitud y Diversidad Léxica





Distribución de longitud de reseñas

- Fuerte asimetría derecha: el 83 % tiene menos de 50 palabras
- Repunte entre 110 y 120 palabras: 367 reseñas frente a ~105 y ~204 en los tramos vecinos. Parece un corte de la plataforma, no un perfil de crítico
- Mediana 16 palabras · media 27,9 · cuartiles 8 y 35

Type-Token Ratio (TTR) post-preprocesamiento

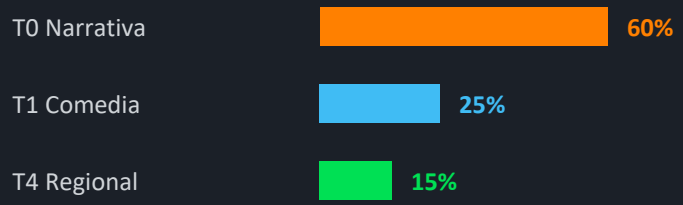
- TTR medio: 0,918 antes del preprocesado y 0,965 después (+0,047)
- Ojo: el TTR depende de la longitud. Las reseñas pasan de 27,7 a 14,5 palabras, y en textos más cortos el TTR sube casi solo.
- Así que la subida NO demuestra por sí sola más riqueza léxica.

Conclusión: el preprocesado quita ruido, pero el TTR no es la prueba de ello: sube en parte porque los textos se acortan.

¿Qué es LDA? — Latent Dirichlet Allocation

Cada documento es una mezcla de tópicos

Ejemplo — Una reseña puede ser:



Cada tópico es una distribución de palabras

Ejemplo — Tópico "Narrativa":

¿Cómo funciona?

- LDA es un modelo generativo probabilístico no supervisado — no necesita etiquetas
- Asume que los documentos fueron generados eligiendo tópicos y luego palabras de esos tópicos
- El algoritmo "invierte" ese proceso: a partir de los textos, infiere qué tópicos existen y cómo se distribuyen
- Hiperparámetro clave: K (número de tópicos) → se selecciona maximizando el coherence score

Modelo: LDA Topic Modeling

Latent Dirichlet Allocation

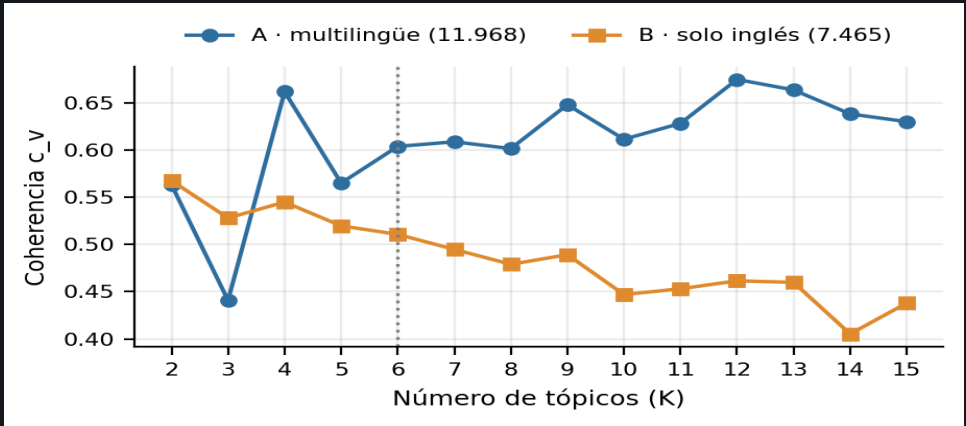
- Gensim LdaModel (determinista)
- Representación: Bag of Words (id2word + corpus)
- Barrido de K: 2 a 15 tópicos
- Métrica: Coherence Score (c_v) · passes=10, random_state=21

Resultado

K = 6

Coherence c_v: 0,6038

vs. K=11 → 0,6282



Coherence vs. K

- El máximo del barrido está en K=12 (0,6747), no en K=6 (0,6038)
- Mantenemos K=6 por interpretabilidad y para comparar con el trabajo original; ahora barrido y modelo final son el MISMO modelo
- c_v premia tópicos de palabras que co-ocurren, y las palabras de un mismo idioma co-ocurren siempre: no puede ser el único criterio

Resultados K=6 — Los 6 Tópicos

T0 Cine en inglés — comedia

like, love, get, time, comedy,
story, first, year, could, work

T1 IDIOMA — alemán / italiano

und, yang, che, ein, ist,
nicht, ich, mit, ini, eine

T2 Festivales y estrenos

festival, beyond, meat, ang, girl,
walk, remember, paul, possible, get

T3 IDIOMA — ruso / persa / turco

что, فيلم, det, это, çok,
این, как, term, фильм, ama

T4 IDIOMA — francés / portugués

mais, les, como, filme, los,
não, des, com, las, muito

T5 Cine en inglés — narrativa

like, story, get, love, people,
time, beautiful, life, work, end

3 de los 6 tópicos son de idioma, no de tema (T1, T3 y T4): se llevan el 32,3 % de las reseñas. El criterio está medido: más del 50 % de sus reseñas no están en inglés.

Comparación K=11 — más tópicos, más idiomas

T0 Cine en inglés · 5.868
like, love, get, story, time...

T1 IDIOMA — alemán · 302
und, ein, ist, nicht, ich...

T2 Festivales · 647
festival, beyond, least, youtube...

T3 IDIOMA — mezcla · 403
letterboxd, aussi, comme, что...

T4 IDIOMA — francés · 789
les, des, mais, sur, plus...

T5 Animación y estilo · 722
fan, animation, big, beautiful...

T6 IDIOMA — español/port. · 1.415
como, filme, los, não, las...

T7 IDIOMA — indonesio · 673
yang, ini, terrible, dude...

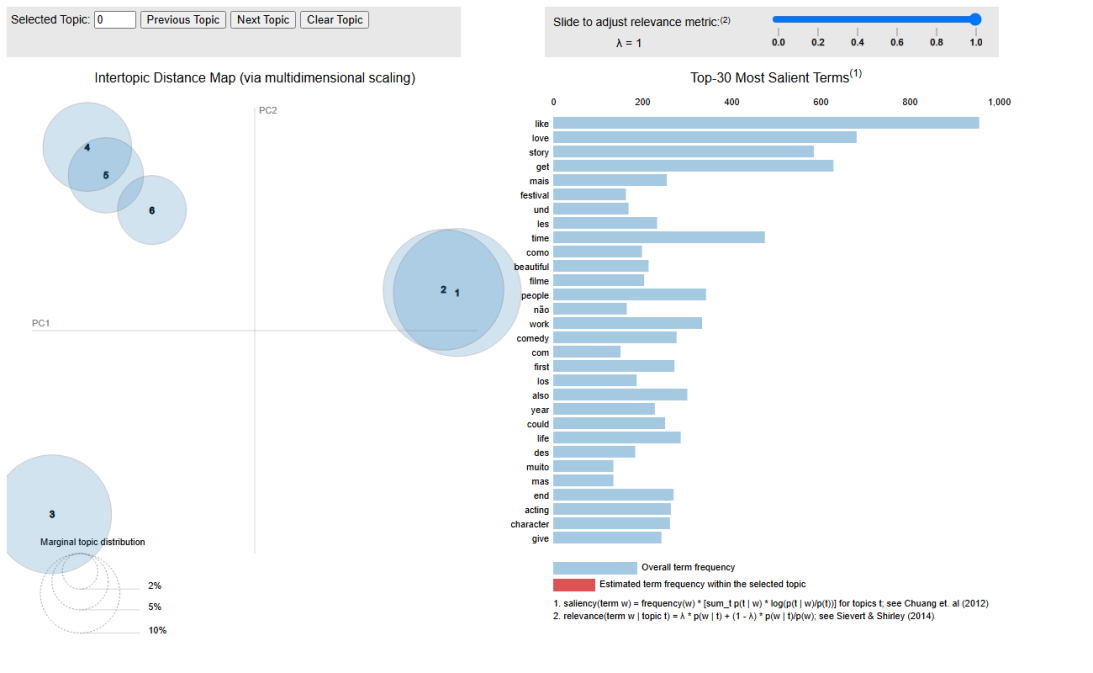
T8 IDIOMA — mezcla · 353
final, worth, vous, charming...

T9 IDIOMA — italiano/neerl. · 476
che, een, van, het, per...

T10 Meta / plataforma · 320
log, obra, det, ela, company...

K=11 → coherencia 0,6282 · 7 de los 11 tópicos son de idioma (36,9 % de las reseñas). Más tópicos no dan más temas: dan más idiomas.

Visualización — pyLDAvis

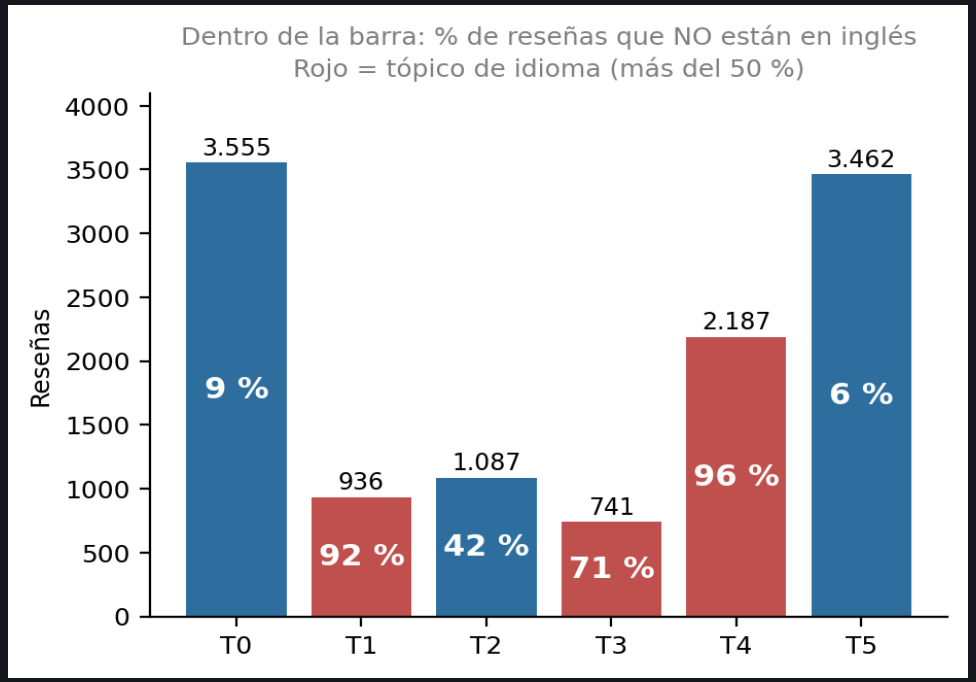


Interpretación

- OJO: pyLDAvis renumera los tópicos de 1 a 6 por tamaño, así que sus números NO son los T0…T5 de la diapositiva anterior
- Los dos círculos grandes que se solapan son los dos tópicos de cine en inglés: 58,6 % de las reseñas. Los de idioma quedan aparte, arriba a la izquierda y abajo
- Los salient terms confirman la mezcla: like, story, get junto a como, les, und, filme
- Tamaño de burbuja ∝ % de documentos asignados

Entre los 30 términos más salientes aparecen mais, und, les, como, filme, não, com, los, muito: el corpus multilingüe se ve a simple vista. Bajando λ se ven las palabras exclusivas de cada tópico.

Distribución de Reseñas por Tópico



Análisis de la distribución

- T0 (3.555) y T5 (3.462) son los mayores: cine en inglés
- T4 (2.187) es el tercero y es un tópico de idioma: 96 % no inglés
- T1 (936) y T3 (741) son los otros dos tópicos de idioma
- T2 (1.087) recoge festivales y estrenos, con un 42 % de reseñas no inglesas
- 58,6 % de las reseñas en tópicos de cine en inglés · 32,3 % en tópicos de idioma

Cada reseña se asigna al tópico con mayor probabilidad. Se generaron features LDA (vector de probabilidades por tópico) que podrían usarse como input para modelos downstream.

¿Y si filtramos por idioma?

A **Corpus multilingüe**

11.968 reseñas · 48 idiomas
coherencia c_v (K=6) = 0,6038

B **Solo inglés**

7.465 reseñas (62,4 % de A)
coherencia c_v (K=6) = 0,5107

12 **B pierde casi siempre**

B queda por debajo de A en 12 de los 14 valores de K del barrido

≠ **No es el tamaño**

Submuestreando A al tamaño de B, doce veces: B queda por debajo de las doce

3 **Tópicos de idioma**

En A, 3 de 6 (32,3 % de las reseñas).
En B ninguno, por construcción

c_v **No premia el ruido**

Los tópicos de idioma puntúan 0,5409 y los de tema 0,6666: tiraban hacia abajo

Mismo preprocesado, mismo barrido, misma métrica, mismos hiperparámetros: lo único que cambia es el corpus. Filtrar NO mejora la coherencia, pero deja de gastar 3 de 6 tópicos en separar idiomas.

Una cifra que nadie puede reproducir

A1 **LdaMulticore, 1.ª vez**

c_v = 0,5633 · random_state=21, workers=2 sobre el mismo corpus

A2 **LdaMulticore, 2.ª vez**

c_v = 0,5404 · mismo código y semilla, mismo ordenador. Otro resultado.

A3 **…y según los núcleos**

Con workers 2, 3 y 4 salen 0,5559, 0,5609 y 0,5378: depende del ordenador

B1 **LdaModel, 1.ª vez**

c_v = 0,5890 · un solo proceso, sin reparto asíncrono entre workers

B2 **LdaModel, 2.ª vez**

c_v = 0,5890 · idéntico, dígito a dígito. Esto sí se puede comprobar

! **La cifra que entregamos**

c_v = 0,678, generada con LdaMulticore: no es reproducible. Tampoco por nosotros.

Por eso todo el trabajo pasa a LdaModel. Entre una coherencia alta y una coherencia que otra persona puede comprobar, elegimos la segunda.

Insights de Negocio

Monitorización automática

Saber en tiempo real de qué habla el público sobre un estreno sin leer miles de reseñas

Segmentación de audiencia

Los tópicos revelan perfiles: cinéfilos técnicos vs. casuales vs. fans de género

Detección de señales

Bigramas como "low budget" o trigramas negativos como alertas tempranas

Estrategia de contenido

"story" y "character" dominan → orientar comunicación de marketing

Conclusiones y Trabajo Futuro

Lo que aprendimos

- El preprocesamiento consume la mayor parte del esfuerzo pero es donde se gana la partida
- Con K=6 la coherencia es 0,6038; el máximo del barrido está en K=12 (0,6747)
- Muestra de 11.968 reseñas de las ~112.000 de 2024. El trabajo original decía 38.531; hoy el dataset da 119.096 y no sabemos por qué
- El multilingüismo contamina: 3 de los 6 tópicos son de idioma (32,3 % de las reseñas), y con K=11 son 7 de 11

Próximos pasos

- Análisis de sentimiento (VADER, transformers) cruzado con ratings
- BERTopic para semántica contextual
- Filtrar por idioma ya está probado (ver experimento): no mejora c_v
- Evolución temporal de tópicos por meses

¿Preguntas?

Alfredo Martín López · Marcos Fernández Arévalo · Pedro Pérez Blanco

Máster en Business Analytics — ICADE — 2024-2025