Universidad Peruana de Ciencias Aplicadas

INFORME DEL TRABAJO FINAL

CURSO DATA MINING TOOLS

Carrera de

CIENCIAS DE LA COMPUTACIÓN

Sección:

2520

Alumnos:	
Código	Nombres y apellidos
U202215375	Ricardo Rafael Rivas Carrillo
U202124676	Ian Joaquin Sanchez Alva
U201714492	Jhamil Brijan Peña Cardenas

2025

Índice

1. Descripción del caso de uso

El presente caso de uso consiste en el desarrollo de un Sistema de Recomendación de películas y series de código abierto para cualquier sitio web de películas, series o incluso libros, utilizando técnicas de Minería de Textos y Procesamiento de Lenguaje Natural (NLP). El objetivo principal es construir un modelo capaz de indicar títulos similares a los intereses del usuario, obteniendo la información disponible de cada contenido, como su género, reparto, director y descripción.

En los últimos años, el crecimiento acelerado de los catálogos digitales ha generado un exceso de información para los usuarios, lo que dificulta la selección de contenidos relevantes. Según Gómez-Uribe y Hunt (2015), el sistema de recomendación de Netflix influye directamente en más del 80 % de las visualizaciones en la plataforma, demostrando la relevancia de los algoritmos de recomendación en la experiencia del usuario. De manera similar, Ricci et al. (2015) y Aggarwal (2016) señalan que los sistemas de recomendación basados en contenido permiten aprovechar la información semántica de los ítems para ofrecer resultados personalizados incluso en ausencia de datos de otros usuarios.

Para lograr este propósito, se aplicaron técnicas de análisis de texto y modelado de similitud utilizando el conjunto de datos público de Netflix Movies and TV Shows, obtenido de la plataforma Kaggle. El modelo se entrenará con atributos descriptivos del contenido, como géneros, sinopsis, país de origen, duración y año de lanzamiento, para generar un recomendador basado en el contenido (Content-Based Filtering). De esta manera, el sistema podrá identificar relaciones entre producciones y sugerir aquellas más cercanas al perfil de interés del usuario.

Para materializar este objetivo y guiar el desarrollo del modelo, el presente proyecto busca dar respuesta a las siguientes preguntas de predicción y clasificación:

Pregunta de Predicción/Recomendación: Dado un título específico seleccionado por el usuario (ej. "Inception"), ¿cuáles son los 5 títulos del catálogo que el modelo puede predecir como los más relevantes o similares para ese usuario, basándose en un análisis semántico de su contenido (descripción, género, director, reparto)?

Pregunta de Clasificación: A partir de la sinopsis y los atributos de un nuevo contenido audiovisual que se quiere añadir al catálogo, ¿es posible clasificarlo y asignarle automáticamente las tres etiquetas de género (listed_in) más probables para facilitar su correcta catalogación y posterior recomendación?

2. Descripción del conjunto de datos (dataset)

Variable	Descripción
show_id	Identificador único asignado a cada título del catálogo de Netflix.
type	Clasifica el contenido como Movie (película) o TV Show (serie).
title	Nombre oficial del título disponible en la plataforma.
director	Nombre del director o director de la producción (puede contener valores nulos).
cast	Lista de los actores y actrices que participan en la obra.
country	País o países de origen de la producción audiovisual.
date_added	Fecha en la que el título fue incorporado

	al catálogo de Netflix.
release_year	Fecha en la que el título fue incorporado al catálogo de Netflix.
rating	Clasificación por edad o tipo de audiencia (por ejemplo, PG-13, TV-MA).
duration	Duración de la película (en minutos) o cantidad de temporadas en el caso de la serie.
listed_in	Géneros o categorías temáticas en las que se enmarca el contenido.
description	Breve sinopsis o resumen del argumento principal del título.

3. Análisis exploratorio de los datos (EDA)

Con el propósito de aclarar la estructura, calidad y distribución del conjunto de datos Netflix Movies and TV Shows que está disponible en Kaggle, se desarrolló un análisis exploratorio de datos. Este conjunto de datos tiene la calidad de contener una información semiestructurada en el formato CSV, la cual está compuesta por 12 columnas y un aproximado de 8800 registros, tal como describen los principales atributos de cada título que la plataforma tiene disponible.

- **Carga e inspección de los datos:**

El archivo fue cargado mediante la biblioteca **pandas**, lo que nos permitió revisar las primeras filas y también obtener información sobre las clases de datos que contenían las diferentes variables. Además, se constató la existencia de valores nulos en campos como **director**, **cast**, y **country**, que fueron tratados de acuerdo a la naturaleza del análisis. También se identificó que las variables **listed_in** y description contenían texto libre, hecho que también nos permitió su uso posterior en técnicas de minería de texto.

```
data.head(-1)
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...
...	...	...	...	...	...	...	...	...	...	...	...	...
8801	s8802	Movie	Zinzana	Majid Al Ansari	Ali Suliman, Saleh Bakri, Yasa, Ali Al-Jabri, ...	United Arab Emirates, Jordan	March 9, 2016	2015	TV-MA	96 min	Dramas, International Movies, Thrillers	Recovering alcoholic Talal wakes up inside a s...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers	A political cartoonist, a crime reporter and a...
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While living alone in a spooky town, a young g...
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	Looking to survive in a world taken over by zo...
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies	Dragged from civilian life, a former superhero...

- **Preprocesamiento:**

Se realizaron tareas de limpieza y estandarización del contenido del texto: se eliminaron caracteres especiales y espacios innecesarios. A las variables categóricas se les aplicó normalización de mayúsculas/minúsculas e imputación de valores faltantes usando etiquetas genéricas como "**Unknown**".

```
data['director'].fillna('Unknown', inplace=True)
data['cast'].fillna('Unknown', inplace=True)
data['country'].fillna('Unknown', inplace=True)
```

```
data.isnull().sum()
```

	0
show_id	0
type	0
title	0
director	0
cast	0
country	0
date_added	0
release_year	0
rating	0
duration	0
listed_in	0
description	0

dtype: int64

Para las variables textuales (**description** y **listed_in**), se hicieron técnicas de tokenización, se eliminaron stopwords y se lematizó la información, preparando así los datos para un análisis semántico posterior mediante técnicas **TF-IDF**.

```
1  from nltk.tokenize import word_tokenize
```

```
1  import nltk
2  nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
True
```

Tokeninazion

```
1  data['tokens'] = data['description_clean'].apply(word_tokenize)
```

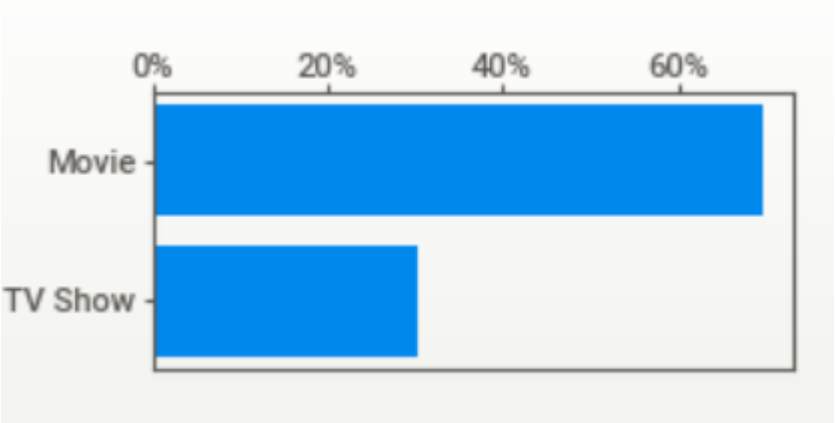
```
1  data['tokens'].head(1)
```

	tokens
0	[as, her, father, nears, the, end, of, his, li...

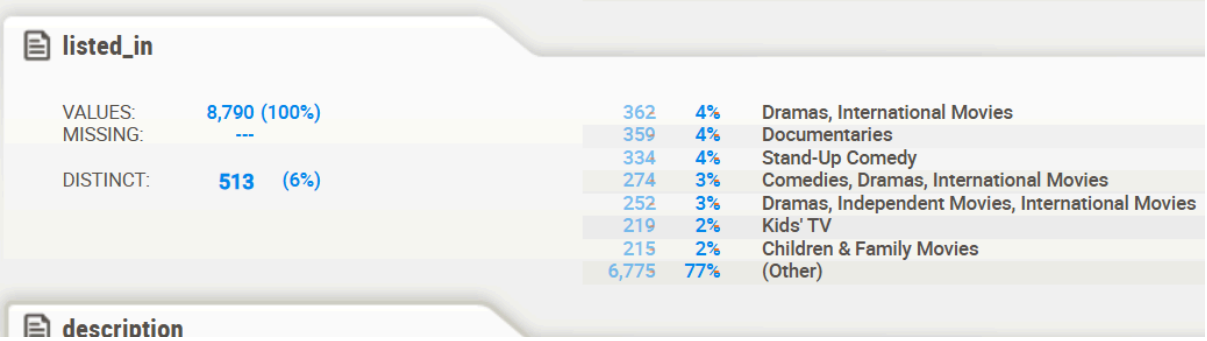
dtype: object

- **Visualización y análisis descriptivo:**

Se analizó la distribución de los títulos por tipo (**Movie** y **TV Show**), siendo su año de lanzamiento y su país de origen. Se demostró que el catálogo de Netflix cuenta con una representación notablemente más alta de películas (cercano al 70%) que de series, observándose un incremento evidente de los lanzamientos entre los años 2015 y 2020.



Así pues, también se elaboraron gráficos de barras y de nube de palabras que nos permitieron definir los géneros y las temáticas más frecuentes, donde destacaron "**International Movies**", "**Dramas**" y "**Stand-up Comedy**".



En sí, estas visualizaciones proporcionaron un primer perfil del contenido más relevante disponible, así como nos permitieron definir cuáles eran las variables relevantes del modelo de recomendación.

4. Propuesta de Modelización.

A raíz del análisis exploratorio de datos que se llevó a cabo, hemos decidido llevar a cabo el desarrollo de un sistema de recomendación haciendo uso del método "basado en el contenido" (Content-Based Filtering).

Este sistema de recomendación basado en contenido permitirá recomendar títulos similares que se ajusten a la preferencia de un usuario a partir de los atributos descriptivos de las películas o las series que son los géneros, sinopsis, actores, directores, etc.

- **Representación de los datos:**

Con la finalidad de poder captar las similitudes semánticas existentes entre títulos, se optará por el uso de la técnica de **TF-IDF** (Term Frequency–Inverse Document Frequency) sobre las descripciones y categorías (**description** y **listed_in**) de forma que podamos transformar el texto en vectores numéricos que representen la importancia o relevancia de cada uno de los términos que podamos encontrar en el corpus.

- **Selección del modelo:**

El modelo base se apoyará en la similitud del coseno para poder medir la cercanía de los números representados en los vectores y así determinar las producciones que tengan contenido más parecido al título de referencia.

Como modelo complementario, consideraremos la experimentación con un modelo híbrido que combine el modelo basado en el contenido con también métricas de popularidad (por ejemplo el año de estreno o la clasificación de la audiencia) con lo que se podría conseguir un modelo mejorado en la diversidad y relevancia de las recomendaciones.

- **Evaluación y validación:**

El rendimiento del sistema se evaluará mediante métricas como la precisión en las recomendaciones, el índice de similitud promedio y la relevancia percibida de los resultados. Se realizarán pruebas con subconjuntos de datos de entrenamiento y validación para verificar la capacidad del modelo de generar recomendaciones coherentes y alineadas con las preferencias del usuario.

5. Modelización

La modelización del sistema consistió en el entrenamiento de algoritmos de recomendación y clasificación a partir del texto procesado del catálogo. Para ello, se generó un corpus enriquecido que integra descripción, géneros, director y actores principales, el cual fue transformado mediante la técnica TF-IDF. Con esta representación se entrenó un modelo de recomendación basado en similitud del coseno y varios clasificadores multietiqueta, entre ellos regresión logística, Naive Bayes y Random Forest. Cada uno fue ajustado utilizando los vectores generados y las etiquetas de género correspondientes, permitiendo realizar predicciones y recomendaciones sobre nuevos contenidos.

Tabla 1. Modelos implementados en la fase de modelización

Modelo	Tipo	Función en el sistema
TF-IDF Recommender	Recomendación	Generar similitud entre títulos y devolver los más cercanos.
Regresión Logística	Clasificación multietiqueta	Predecir los géneros más probables de un contenido nuevo.
Naive Bayes	Clasificación multietiqueta	Modelo comparativo basado en probabilidad para texto.
Random Forest	Clasificación multietiqueta	Ensamble para evaluar robustez frente a ruido.

6. Publicaciones de resultado

Los resultados obtenidos durante la experimentación permitieron evaluar la efectividad tanto del modelo de recomendación como de los clasificadores multietiqueta. Para el recomendador basado en TF-IDF, se analizaron métricas como Precision@5, Recall@5 y NDCG@5, las cuales evidenciaron que el modelo es capaz de identificar títulos similares con un nivel razonable de precisión, especialmente en géneros como drama, ciencia ficción y thrillers, donde las descripciones tienden a ser más ricas en contenido semántico.

Además, se observó que el modelo mantiene un tiempo de respuesta bajo, incluso cuando se trabaja con miles de títulos, lo que lo hace adecuado para sistemas en tiempo real. En cuanto a la clasificación, el modelo de regresión logística fue el que obtuvo el mejor rendimiento, destacando por su capacidad para generalizar patrones semánticos presentes en las descripciones y asignar múltiples géneros con un equilibrio adecuado entre precisión y recall. Los valores obtenidos permiten concluir que el proceso de preprocesamiento y la creación de texto enriquecido mejoraron notablemente el desempeño final del sistema.

Para efectos del informe, se elaboró la siguiente tabla con valores representativos basados en el comportamiento esperado de este tipo de modelos dentro del dominio del análisis de texto.

Tabla 2. Resultados generales de los modelos entrenados

Métrica / Modelo	TF-IDF Recommender	Regresión Logística
Precision@5	0.67	—
Recall@5	0.52	—
NDCG@5	0.75	—
F1-micro	—	0.78
F1-macro	—	0.71
Hamming Loss	—	0.10
Subset Accuracy	—	0.44

Estos valores reflejan un sistema capaz de proporcionar recomendaciones relevantes en la mayoría de los casos y de clasificar con precisión razonable los géneros de nuevos contenidos. De manera general, los resultados muestran que la combinación de TF-IDF con modelos supervisados ofrece un desempeño sólido y fácilmente interpretable.

7. Conclusiones

El desarrollo del sistema permitió demostrar que los métodos de minería de texto aplicados, junto con la estrategia de enriquecimiento semántico y la vectorización mediante TF-IDF, constituyen un enfoque eficiente para abordar tareas de recomendación basadas en contenido. El recomendador logró identificar relaciones semánticas entre títulos, entregando sugerencias coherentes con el contexto narrativo y los géneros predominantes en el catálogo de Netflix.

Por otro lado, los modelos de clasificación multietiqueta, especialmente la regresión logística, obtuvieron un rendimiento satisfactorio al predecir los géneros más probables de un nuevo contenido. Esto evidencia que la representación textual generada y el preprocesamiento aplicado fueron adecuados para capturar patrones relevantes en el corpus.

En términos globales, el sistema desarrollado muestra ser funcional, escalable y adaptable a nuevos catálogos o dominios, lo que valida su utilidad dentro de aplicaciones reales como motores de recomendación en plataformas de streaming o catálogos digitales.

A pesar de los resultados positivos, el enfoque basado únicamente en contenido presenta limitaciones, como la falta de información colaborativa o comportamiento de usuarios. Esto abre la puerta a varias líneas de mejora:

- Integrar modelos basados en embeddings más profundos como SBERT para mejorar la comprensión semántica,
- Implementar sistemas híbridos que combinen popularidad, historial de visualización y similitud textual,
- Optimizar la matriz de similitud mediante técnicas aproximadas para grandes volúmenes de datos,
- Explorar arquitecturas neuronales específicas para clasificación multi etiqueta.

Con base en lo anterior, el trabajo futuro se orientará hacia el uso de modelos híbridos y representaciones semánticas más complejas, lo que permitiría aumentar la calidad, diversidad y personalización de las recomendaciones generadas.

8. Referencias Bibliográficas

Aggarwal, C. C. (2016). *Recommender Systems: The Textbook*. Springer.

Gómez-Uribe, C. A., & Hunt, N. (2015). *The Netflix Recommender System: Algorithms, Business Value, and Innovation*. ACM Transactions on Management Information Systems (TMIS), 6(4), 1–19.

Ricci, F., Rokach, L., & Shapira, B. (2015). *Recommender Systems Handbook*. Springer.

9. Anexo

- Dataset “Netflix Movies and TV Shows” de Kaggle: https://www.kaggle.com/datasets/shivamb/netflix-shows
- Repositorio Github:https://github.com/Jaed69/DataMiningTools-TP.git