Universidad Peruana de Ciencias Aplicadas

INFORME DEL TRABAJO PARCIAL

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

▪ Descripción del conjunto de datos (dataset). Redactar las características y origen de los datos recolectados motivo de análisis y para su posterior analítica.

▪ Análisis exploratorio de los datos (EDA). Los datos recolectados deberán ser semi o no estructurados. Se debe incluir la descripción de las tareas de carga, inspección, preprocesamiento y visualización de los datos.

▪ Modelización. Comprende el seleccionar uno más algoritmos para ser entrenados a partir de los datos preparados. Comprobar el rendimiento de los modelos creados y experimentar con ellos con datos de prueba para realizar clasificaciones / predicciones.

▪ Publicación de los resultados. Comunicar los resultados obtenidos a partir de los experimentos realizados con los modelos creados (uso de métricas y tablas comparativas).

▪ Conclusiones. En un párrafo redactar las conclusiones del trabajo, especificando la(s) técnica(s) utilizadas, los resultados obtenidos (positivos o no), y de ser el caso, el trabajo a futuro.

▪ Referencias bibliográficas

1. Descripción del caso de uso

El presente caso de uso consiste en el desarrollo de un sistema de recomendación de código abierto para películas y series, utilizando técnicas de Minería de Textos y Procesamiento de Lenguaje Natural (NLP). El objetivo principal es construir un modelo que, a diferencia de los sistemas convencionales, entienda el ADN del contenido (género, reparto, director y, sobre todo, la semántica de la sinopsis) para predecir y clasificar títulos de manera inteligente.

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.

1.1. Innovación del Proyecto: Enfoque Semántico vs. Convencional

La principal innovación de nuestro proyecto radica en el método empleado. Mientras que las plataformas tradicionales se basan mayoritariamente en el **Filtrado Colaborativo** ("a otros usuarios les gustó..."), nuestro sistema se centra en el **Filtrado Basado en Contenido**.

- **Enfoque Convencional (Filtrado Colaborativo):** Recomienda basándose en la similitud de comportamiento entre usuarios. Su gran limitación es el "problema de arranque en frío" (no funciona para usuarios nuevos o títulos sin vistas) y tiende a crear "burbujas de filtro" que limitan el descubrimiento de contenido nuevo.
- **Nuestra Propuesta (Filtrado Basado en Contenido y NLP):** Analizamos la semántica del contenido para encontrar similitudes intrínsecas entre los ítems. Esto nos permite ofrecer recomendaciones personalizadas desde el primer momento,

descubrir "joyas ocultas" y resolver problemas de clasificación automática sin depender del historial de otros usuarios.

1.2. Preguntas de Investigación

Para guiar el desarrollo, nuestro proyecto busca responder dos preguntas clave:

1. **Pregunta de Predicción/Recomendación:** Dado un título específico, ¿cuáles son los 5 títulos del catálogo que nuestro modelo puede predecir como los más similares basándose en un análisis semántico de su contenido?
2. **Pregunta de Clasificación:** A partir de la sinopsis de un nuevo contenido, ¿podemos clasificarlo y asignarle automáticamente sus tres géneros más probables para facilitar su catalogación?

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.

```
1 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**".

Para las variables textuales (**description** y **listed_in**), se hicieron técnicas de

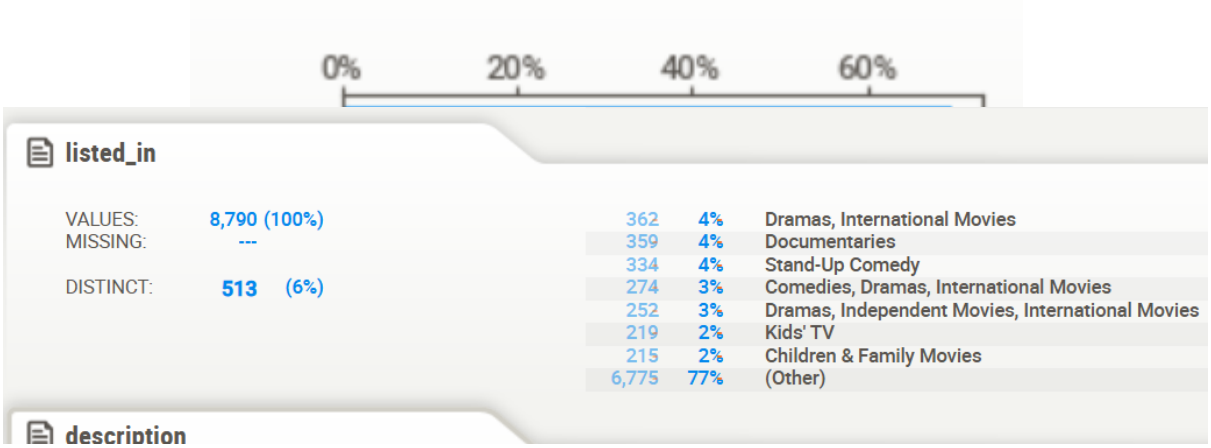
tokenización, se eliminaron stopwords y se lematizó la información, preparando así los datos para un análisis semántico posterior mediante técnicas **TF-IDF**.

- **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 Inicial.

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. Propuestas de Modelado Avanzadas

Para superar las limitaciones de TF-IDF y alcanzar un rendimiento superior, se proponen las siguientes arquitecturas avanzadas de NLP y Deep Learning:

5.1. Para Recomendación (Análisis de Similitud)

- **Modelo 1: Embeddings de Documentos (Doc2Vec):**
  - **Arquitectura:** Modelo de NLP que aprende una representación vectorial de dimensión fija para cada sinopsis, capturando su "esencia" temática.
  - **Respuesta a la Pregunta:** Se entrena un modelo Doc2Vec con todas las sinopsis. Luego, se calcula la similitud del coseno entre el vector del título elegido y todos los demás para encontrar el Top 5 de recomendaciones.

- **Modelo 2: State-of-the-Art (Sentence-Transformers):**
  - **Arquitectura:** Modelos Transformers como `SBERT` (Sentence-BERT), pre-entrenados y afinados para tareas de similitud semántica de alta precisión.
  - **Respuesta a la Pregunta:** Se vectoriza cada sinopsis con SBERT. La comparación mediante Similitud del Coseno sobre estos embeddings de alta calidad permite obtener un ranking de recomendaciones semánticamente muy superior.

5.2. Para Clasificación de Géneros (Multietiqueta)

- **Modelo 1: Redes Neuronales Recurrentes (LSTM/GRU):**

    - **Arquitectura:** Redes con celdas `LSTM` (Long Short-Term Memory) o `GRU` (Gated Recurrent Unit) que procesan el texto de forma secuencial, conservando una "memoria" del contexto.
    - **Respuesta a la Pregunta:** Se entrena una red LSTM/GRU usando las sinopsis como entrada y los vectores de géneros (multietiqueta) como salida. La capa final (sigmoide) predice una probabilidad para cada género, permitiendo seleccionar los 3 más probables.

- **Modelo 2: Fine-Tuning de un Modelo Transformer:**
    - **Arquitectura:** Utilizar un Transformer pre-entrenado (ej. `BETO`, la versión en español de BERT) y re-entrenar sus capas finales (fine-tuning) para la tarea específica de clasificación multietiqueta.
    - **Respuesta a la Pregunta:** Se especializa el Transformer para que aprenda a mapear el texto complejo de una sinopsis a las etiquetas de género. Al recibir un nuevo texto, el modelo predice las probabilidades de todos los géneros, permitiendo identificar los 3 principales.

Para determinar el mejor modelo de forma objetiva, se establece el siguiente marco de evaluación:

- **Para Recomendación:**
    - **Precisión / Recall:** Mide cuántos de los *k* ítems recomendados son relevantes.
    - **MAP (Mean Average Precision):** Evalúa la calidad del orden (ranking) de las recomendaciones.
- **Para Clasificación Multietiqueta:**
    - **F1-Score (Macro/Micro):** Métrica robusta que balancea precisión y recall, ideal para problemas con clases desbalanceadas.
    - **Hamming Loss:** Mide la fracción de etiquetas que fueron predichas incorrectamente.

5. 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.

6. 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