



Sistemas inteligentes y representación del conocimiento

Grado en Ingeniería
Informática

Borja Monsalve Piqueras

Proyecto

Recomendaciones basadas en el contenido




**Universidad
Europea**




1

Proyecto: Recomendaciones basadas en el contenido



Índice de contenidos

- Objetivo general del proyecto
- Contexto: Movielens
- Objetivos específicos
- Proceso
- Recursos



2

© Borja Monsalve Piqueras

2

Proyecto: Recomendaciones basadas en el contenido



Objetivo general del proyecto

- Desarrollar un sistema de recomendación basado en contenido a partir del *dataset* de películas de Movielens.
- Palabras clave:
 - Python
 - Interfaz gráfica
 - Web scraping
 - Similitudes
 - Recomendaciones
 - Trabajo en equipo
 - Documentación

3

© Borja Monsalve Piqueras

3

Proyecto: Recomendaciones basadas en el contenido



Contexto: Movielens

- **Movielens** es una comunidad online de usuarios que dan su opinión sobre las películas que han visto, valorándolas entre 1 y 5 estrellas: <https://movielens.org>
- **GroupLens Research** es un laboratorio de investigación de interacción hombre-máquina en el Departamento de Ciencias de la Computación e Ingeniería de la Universidad de Minnesota.
- Disponen de varios dataset para investigación y experimentación relacionados con Movielens.
- Trabajaremos con la **versión Small** ("ml-latest-small"):
 - 100.000 valoraciones
 - 3.600 etiquetas
 - +9.000 películas
 - +600 usuarios
 - Última actualización: 9/2018

4

© Borja Monsalve Piqueras

4

Proyecto: Recomendaciones basadas en el contenido

Contexto: MovieLens



- Dataset de MovieLens:

(<https://files.grouplens.org/datasets/movielens/ml-latest-small-README.html>)

- movies.csv

- movieId
- title
- genre

- tags.csv

- userId
- tag
- timestamp

- géneros

1. Action
2. Adventure
3. Animation
4. Children's
5. Comedy
6. Crime
7. Documentary
8. Drama
9. Fantasy
10. Film-Noir
11. Horror
12. Musical
13. Mystery
14. Romance
15. Sci-Fi
16. Thriller
17. War
18. Western
19. (no genres listed)

- ratings.csv

- userId
- movieId
- rating
- timestamp

- links.csv

- movieId
- imdbId
- tmdbId

5

© Borja Monsalve Piqueras

5

Proyecto: Recomendaciones basadas en el contenido

Contexto: MovieLens



- Ejemplo

movies.csv

movieId	title	genres
1	Toy Story (1995)	Adventure Animation Children Comedy Fantasy

ratings.csv

userId	movieId	rating	timestamp
1	1	4.0	964982703
1	3	4.0	964981247

tags.csv

userId	movieId	tag	timestamp
336	1	pixar	1139045764
474	1	pixar	1137206825
567	1	fun	1525286013

links.csv

movieId	imdbId	tmdbId
1	114709	862



© Borja Monsalve Piqueras

6

6

Proyecto: Recomendaciones basadas en el contenido



Objetivos específicos

- **OE1.** A partir de una película dada, devolver un ranking con las N películas más similares.
- **OE2.** A partir de un usuario dado y en función de sus valoraciones:
 - **OE2.1.** Devolver la predicción para una película P dada.
 - **OE2.2.** Devolver un ranking de N películas más recomendables.
- Tareas adicionales:
 - Web scraping
 - PLN básico
- Contenidos a utilizar:
 - Géneros
 - Tags
 - Sinopsis

7

© Borja Monsalve Piqueras

7

Proyecto: Recomendaciones basadas en el contenido



Proceso

- Obtener las sinopsis de las películas
 - IMDB: <https://www.imdb.com/>
 - TMDB: <https://www.themoviedb.org/>
- PLN sobre las sinopsis
- Crear un vector (TF-IDF) para cada película a partir del contenido disponible:
 - Sinopsis
 - Géneros
 - Tags
 - Etc.

8

© Borja Monsalve Piqueras

8

Proyecto: Recomendaciones basadas en el contenido

Proceso



- TF-IDF:

TF-IDF = peso de un término T en un documento D.

TF = Número de veces que aparece el término T en el documento D (frecuencia).

IDF = Frecuencia inversa del término T en todos los documentos D_i de la colección.

$$\text{TF-IDF} = \text{TF} \times \text{IDF}$$

TF = Frecuencia (T, D)

IDF = $\log(N / n_i)$

N = Número total de documentos de la colección

n_i = Número de documentos en los que aparece T

9

© Borja Monsalve Piqueras

9

Proyecto: Recomendaciones basadas en el contenido

Proceso



- Utilizar la similitud del coseno para crear una matriz de similitudes entre cada par de películas.
- Crear el perfil de usuario a partir de los contenidos considerados.
- Calcular la predicción de una película para un usuario a partir de sus valoraciones.
- Crear un ranking de N películas recomendadas para un usuario, a partir de las predicciones individuales de las películas que no ha visto.

10

© Borja Monsalve Piqueras

10

Proyecto: Recomendaciones basadas en el contenido



Recursos

- NLTK
- SciKit Learn
- Beautiful Soup
- Numpy
- Pandas
- PyQT
- Delphi (interfaz) --> Exportar a Python
- Etc.

11

© Borja Monsalve Piqueras