Trabajo Final (propuesta)

Diana Valentina Castro Silva
614181013

Simulación

3 de octubre de 2020

Durante mi proceso como estudiante de matemáticas, una de las áreas en las que más me ha gustado experimentar ha sido en el análisis de datos. Me gustaría en esta ocación trabajar datos relacionados con Netflix. Mi objetivo en este trabajo, será responder la siguiente pregunta:

¿Qué película/serie debería ver después?

Me llama la atención responder esta pregunta porque se que la técnica usada actualmente tiene mucho que ver con machine learning y analitica de datos. Como lo mencione, se algunas cosas de análisis de datos pero realmente me gustaría ponerme el reto de aprender algunos conceptos y tecnicas de machine learning.

Para resolver esta pregunta, recolecte una serie de datos (adjuntos):

En Kaggle, encontré un conjunto de datos abiertos[1] el cual proporciona los siguientes datos en cada registro:

- Tipo (Pelicula o serie)
- Título
- Director
- Elenco
- País
- Fecha en la cual fue agregado
- Calificación
- Duración

Directamente de Netflix[2], la actividad de visualización con el permiso adecuado de una serie de personas.

El archivo que proporciona netflix, tiene únicamente dos datos en cada registro:

- Titulo.
- Fecha de visualización.

[1] https://www.kaggle.com/dearsirmehta/100-analysis-using-netflix-datasets
[2] https://help.netflix.com/es/node/101917

Como primera respuesta para resolver esta pregunta:

Generar consultas las cuales agrupen los datos de manera objetiva. Así, dada una entrada (peliculas y/o series, gustos), retornar una serie de sugerencias.

Estos agrupamientos si corresponden a un "buen agrupamiento" deberían concordar con los datos recopilados de la actividad de visualización, en cada uno de los casos o al menos en la mayoria de ellos. Esto como forma de verificar que nuestros agrupamientos quedaran bien hechos.

Como segunda y ultima respuesta

Aplicar tecnicas de Machine Learning. Esta parte aun esta en proceso de diseño ya que es algo nuevo para mi, hasta el momento de lo que he leido tengo algunas ideas al aire como por ejemplo, uso de grafos/redes y aprendizaje automatico, este ultimo en especial por los datos que tengo siento que seria un muy buen candidato ya que, con una gran cantidad de registros de visualizaciones (varias cuentas/personas distintas) es posible entrenar un modelo.

Para esto propongo dos herramientas:

- VectorFlow:

  En The Netflix Tech Blog[1], hay un articulo en el cual hablan sobre lo útil que ha sido para netflix hacer de técnicas de Machine Learning, en especial al momento de personalizar y recomendar. En el caso de Netflix, usan VectorFlow ya que algunas aplicaciones se necesitan en tiempo real y en una sola máquina, además, de redes sencillas y poco profundas. Y estas necesidades las cubre VectorFlow muy bien.

  Con esto en mente, sabiendo que hace uso de hace uso de redes sencillas, a demás de las otras ventajas, siento que es una buena manera de iniciar en el área de Machine Learning y sacar buenos resultados.

- TensorFlow

  Es una plataforma de código abierto, la cual compila y entrena modelos de Aprendizaje Automático con facilidad mediante APIs con ejecución inmediata y depuración fácil.

  Lo que me llama la atención de utilizar esta herramienta, es que, es más conocida, por lo cual tiene más repositorios, problemas resueltos, ejemplos, guias y demás, además de que al igual que VectorFlow es posible utilizar redes sencillas.

Referencias

[1] Faisal Siddiqi. Machine learning platform meetup. *The Netflix Tech Blog*, 2017.