Proyecto 2:
Búsqueda Basado en el Contenido
Base de Datos Multimedia

1- Introducción

El logro del estudiante está enfocado a entender y aplicar los algoritmos de búsqueda y recuperación de la información basado en el contenido.

Este proyecto está dividido en dos partes, búsqueda y recuperación en documentos de texto, y búsqueda y recuperación en imágenes.

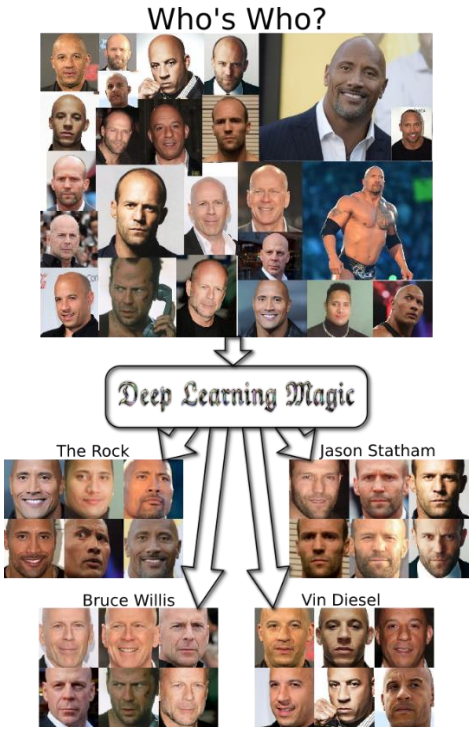
2- Recuperación de Tweets Basado en el Contenido (10 pt)

Implementar el índice invertido para recuperación de texto usando el modelo de recuperación por ranking para consultas de texto libre. Considere los siguientes pasos generales:

- Preprocesamiento:
  - Tokenization
  - Filtrar Stopwords
  - Reducción de palabras - *Stemming* [1]
- Construcción del Índice
  - Use la similitud de coseno sobre con el peso TF-IDF.
  - Debe poder guardarse en memoria secundaria
- Consulta
  - Proponga tres consultas y muestre el **top-10 de los tweets** que se aproximan a dicha consulta.
  - La consulta es solo una o más palabras en lenguaje natural.
- Para probar el desempeño de su implementación. Se proveerá una colección de aproximadamente 20mil tweets de Twitter.
  - El diccionario de términos se construye usando el contenido del atributo “text”. El docID vendría a ser el Id del tweet.
- Analice el performance de su implementación y proponga una la solución algorítmica para el uso de memoria secundaria ante grandes colecciones de datos.

3- Búsqueda Eficiente en Imágenes (10 pt)

Aplicar la búsqueda de los k vecinos más cercano en la detección de rostros usando la librería Face_Recognition. En dicha librería se encuentra implementado las técnicas de extracción de características para obtener de cada imagen una representación numérica y compacta (enconding). La eficacia del reconocimiento ha sido probada con modelos de búsqueda basados en *deep learning* (99.38% de precisión). En esta tarea vamos a usar las funciones básicas de dicha librería para una tarea de recuperación de rostros que responda a las consultas de tipo ¿Quiénes son cinco las personas más parecidas a Bruce Willis?



- Implementar en lenguaje Python el algoritmo búsqueda KNN, el cual recibe como parámetro el objeto de consulta y la cantidad de objetos a recuperar.
- La colección de fotos lo pueden descargar desde http://vis-www.cs.umass.edu/lfw/.
- Investigue y aplique el uso de un índice multidimensional para acelerar la búsqueda KNN.
    - Por ejemplo, R-Tree para Python http://toblerity.org/rtree/
- Use la distancia Euclidiana (ED) como medida de distancia.
- Muestre los resultados de búsqueda interactivamente.
    - El personaje de consulta y el valor de k debe ser un dato de entrada.
- Investigue o proponga una medida de evaluación de los resultados obtenidos.
- El entregable es un informe que evidencie el trabajo realizado (explicación del algoritmo de búsqueda, capturas de pantallas, el uso de otras librerías, etc).

4- Entregable

Los alumnos formaran grupos de máximo de tres integrantes. La entrega del proyecto se hará mediante el aula virtual. La carpeta zipeada debe contener dos elementos:

1- Informe del proyecto conjunto
2- Código fuente de cada aplicación.

La fecha límite de entrega es el 03/11/2019.

5- Lenguaje de programación:

Cualquier lenguaje de programación, pero se recomienda Python. Los resultados deben visualizarse de forma amigable e intuitiva.

6- Informe del proyecto

- Formato artículo de una sola columna.
- Máximo número de hojas: 10.
- Ortografía y consistencia en los párrafos.
- Trabajar de forma colaborativa en GitHub. Se considerará para su nota individual.