75.06 Organización de Datos
Trabajo Práctico: Primer Cuatrimestre de 2015

"When Bag of Words meets bags of Popcorn"

Link a Kaggle: http://www.kaggle.com/c/word2vec-nlp-tutorial

El objetivo de este TP es desarrollar una solución para la competencia de Kaggle "When Bag of Words meets bags of Popcorn". En ésta competencia se pide desarrollar un algoritmo que permita predecir el sentimiento de reviews de películas tomados de IMDB. El set de entrenamiento cuenta con 25000 reviews y un indicador de si el review es positivo o no (1=positivo 0=negativo). El set de test tiene otros 25000 reviews cuyo sentimiento desconocemos. El TP tiene que ser capaz de predecir el sentimiento de estos reviews.

Los sets de datos se pueden bajar de Kaggle a partir del link indicado. El formato del archivo a generar también está indicado en Kaggle. Por favor notar que por cada review en el archivo de salida se pide la PROBABILIDAD de que el mismo sea positivo, una probabilidad cercana a 1 indica un review positivo y una probabilidad cercana a 0 indica un review negativo. El archivo de submission tiene que seguir el mismo orden que el archivo de kaggle.

MUY IMPORTANTE: Además del archivo de submission pedido por kaggle generar otro que solo tenga una única columna con las probabilidades.Es decir un archivo igual que el de submission pero sin la columna de ids.

Condiciones Necesarias De Aprobación

- El TP se tiene que desarrollar en C o C++ exclusivamente.
- El TP se tiene que desarrollar bajo Linux y será probado bajo Linux.
- El Grupo debe tener al menos un submission subido en Kaggle antes de la fecha de prueba del TP. (condición absolutamente excluyente)

El Grupo que mejor score alcance en Kaggle obtiene 10 puntos que pueden usarse en el examen por promoción o el segundo recuperatorio.

No es necesario usar Word2vec, puede plantearse cualquier solución, pretendemos que los alumnos investiguen y aporten una solución creativa al problema.

La entrega de diseño consiste en un informe en el cual cada grupo describe la solución que va a utilizar para el TP, bilbiografía y fuentes consultadas, ideas propias aportadas por el grupo, algoritmos que van a implementar y que uso les van a dar. Etc.

Los grupos que no respeten el código de honor explicado en la primera clase de la materia automaticamente quedarán libres todos sus integrantes y se iniciará un sumario administrativo.