Espero que investigaron el corpus de criticas cine. Las críticas son opiniones de personas acerca de películas. Cada persona escribió una crítica sobre una película y además seleccionó la calificación para esta película, la calificación está representada como el número de las estrellas (de 1 a 5).

Ahora vamos a tomar las críticas, es decir sus textos, normalizarlos y vectorizar. Los vectores como antes metemos en la matriz X. A cada texto, es decir su vector, le va a corresponder una etiqueta que es el número de estrellas (de 1 a 5), las etiquetas metemos en el vector y. La matriz X y el vector y será nuestro conjunto de datos.

Como antes, lo vamos a mezclar aleatoriamente, luego dividir en el conjunto de entrenamiento (70%) y el conjunto de prueba (30%). Vectorizar los textos pueden con el CountVectorizer de scikit-learn o también con TfidfVectorizer, pueden probar cual funcionaría mejor.

Con el conjunto de entrenamiento van a entrenar el clasificador de la Regresión logística ordinal implementado en “mord” (se llama LogisticIT) y luego lo aplicarán al conjunto de prueba y obtendrán la matriz de confusión y los valores de precisión, recall y F1 ahora para cinco clases (de 1 a 5).

Ahora vamos a hablar sobre la normalización de textos. De hecho los textos están preprocesados: están en minúsculas y lematizados: estos textos están el en archivo <número de crítica.review>.pos, aquí cada palabra de texto está en una línea, entonces de cada línea hay que extraer el segundo token que es la palabra del texto lematizada.

Por ejemplo, aquí hay unas primeras líneas del archivo 2.review.pos:

Cada cada DI0CS0 1
vez vez NCFS000 1 05449233
me me PP1CS000 0.889706
gusta gustar VMIP3S0 0.928571 01213391:01241953:01244897
menos menos RG 1
el el DA0MS0 1
cine cine NCMS000 1 02442077:04735661
de de SPS00 0.999919
masas masa NCFP000 1 03923435:03977751:05240866:05671312:05889686:06080290:06081960:06082537:06674595:09920569
. . Fp 1

Extrayendo el segundo token de cada línea vamos a tener:
cada vez me gustar menos el cine de masa .

Ustedes solo necesitan eliminar los stopwords y caracteres especiales y ya van a tener el texto normalizado.

En el archivo <número de crítica>.xml se encuentra el valor de la calificación en el tag “rank”, éste hay que meter en el vector y, por ejemplo, en el archivo 2.xml en la primera línea se encuentra

rank="1"

“1” es la calificación de la película, entonces el 1 hay que meter en el vector y.

La tarea es hacer el programa para entrenar y probar el clasificador de la Regresión logística ordinal y obtener los valores en la matriz de confusión, los valores de precisión, recall y F1 sobre el conjunto de prueba.

Por favor envíenme el programa y la captura de pantalla con los valores mencionados hasta el martes 19 de mayo.