Trabajo 14
Análisis de Sentimientos en Reseñas

Facultad De Ingeniería, Universidad De Cuenca
TEXT MINING
Freddy L. Abad L.
{freddy.abadl}@ucuenca.edu.ec

Esta práctica tiene como objetivo analizar el sentimiento de dos reseñas de película una vez entrenado un modelo con una base de datos etiquetada usando aprendizaje automático.
La idea es poder clasificar las reseñas de películas en un sentimiento positivo frente al negativo.
El modelo debe ser entrenado con datos que sean:

- a) relacionados con el tema, es decir, reseñas de películas; y
- b) revele información sobre el sentimiento de cada revisión, es decir, positiva frente a negativa.

Un conjunto de datos que cumple con estas expectativas se puede identificar en Polarity Dataset v2.0 es Movie Review Data disponible en: http://www.cs.cornell.edu/people/pabo/movie-review-data/

Movie Review Data

This page is a distribution site for movie-review data for use in sentiment-analysis experiments. Available are collections of movie-review documents labeled with respect to their overall *sentiment polarity* (positive or negative) or *subjective rating* (e.g., "two and a half stars") and sentences labeled with respect to their *subjectivity status* (subjective or objective) or *polarity*. These data sets were introduced in the following papers:

- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques, *Proceedings of EMNLP 2002.*
- Bo Pang and Lillian Lee, A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, *Proceedings of ACL 2004.*
- Bo Pang and Lillian Lee, Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, *Proceedings of ACL 2005.*

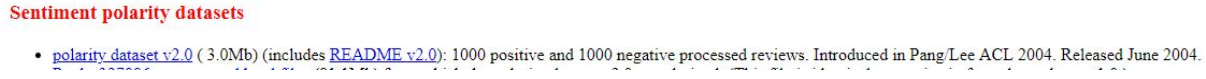
Until April 2012 (but no longer), we maintained a list for of other papers using our data the purposes of facilitating comparison of results.

Please cite the version number of the dataset you used in any publications, in order to facilitate comparison of results. Thank you.

Sentiment polarity datasets

- polarity dataset v2.0 (3.0Mb) (includes README v2.0): 1000 positive and 1000 negative processed reviews. Introduced in Pang/Lee ACL 2004. Released June 2004.
- Pool of 27886 unprocessed html files (81.1Mb) from which the polarity dataset v2.0 was derived. (This file is identical to movie.zip from data release v1.0.)
- sentence polarity dataset v1.0 (includes sentence polarity dataset README v1.0: 5331 positive and 5331 negative processed sentences / snippets. Introduced in Pang/Lee ACL 2005. Released July 2005.
- archive:
  - polarity dataset v1.0 (2.8Mb) (includes README): 700 positive and 700 negative processed reviews. Released July 2002.
  - polarity dataset v1.1 (2.2Mb) (includes README.1.1): approximately 700 positive and 700 negative processed reviews. Released November 2002. This alternative version was created by Nathan Treloar, who removed a few non-English/incomplete reviews and changing some of the labels (judging some polarities to be different from the original author's rating). The complete list of changes made to v1.1 can be found in diff.txt.
  - polarity dataset v0.9 (2.8Mb) (includes a README):. 700 positive and 700 negative processed reviews. Introduced in Pang/Lee/Vaithyanathan EMNLP 2002. Released July 2002. Please read the "Rating Information - WARNING" section of the README.
  - movie.zip (81.1Mb): all html files we collected from the IMDb archive.

Figura 1: *Página Web con la información relevante para la práctica*



Sentiment polarity datasets

- polarity dataset v2.0 (3.0Mb) (includes README v2.0): 1000 positive and 1000 negative processed reviews. Introduced in Pang/Lee ACL 2004. Released June 2004.

Figura 2: *Opción seleccionada para los dataset de prueba*

Este conjunto de datos fue generado dentro de un proyecto de investigación en la Universidad de Cornell. Este repositorio contiene 1000 revisiones positivas y 1000 críticas negativas.

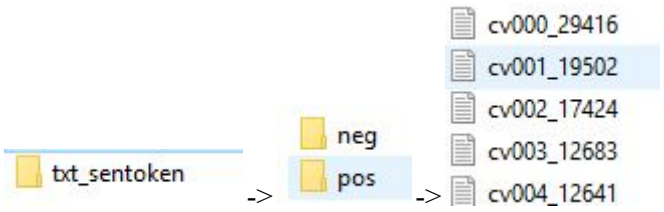


Figura 3: *Reseñas descargadas para el proceso de análisis de sentimientos*

***Reseña**: El proceso generado que soluciona los requerimientos de esta práctica se muestra en su totalidad en la Figura 4.*

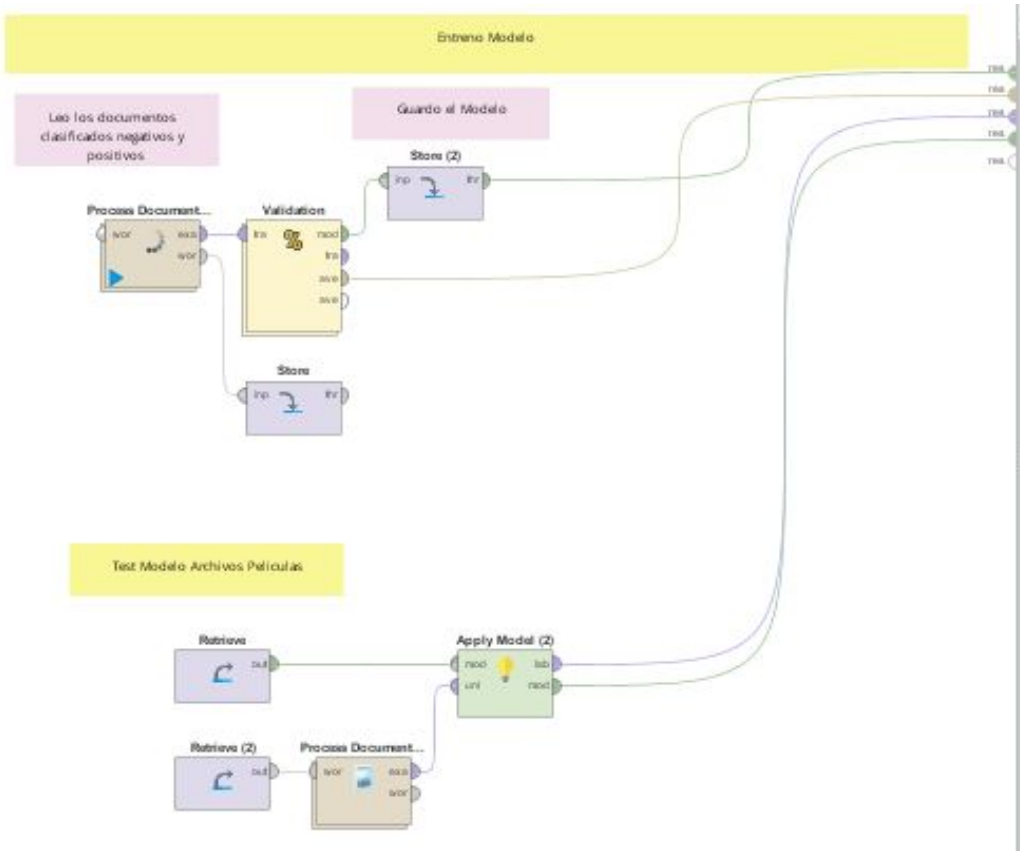


***Figura 4:** Proceso general para la clasificación en base de SVM.*

Después de descargar los datos, procese cada uno de ellos para extraer los términos relacionados utilizando la opción de creación de vectores **TF-IDF.**

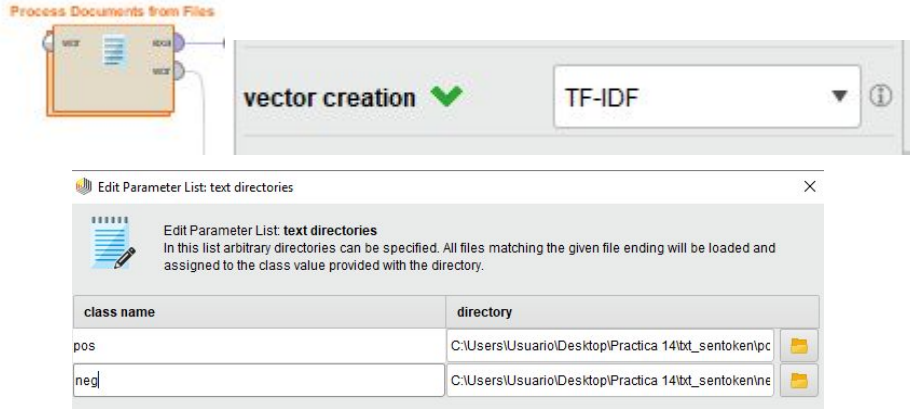


***Figura 5 :** Configuración del Proceso "Process Documents from files", con el respectivo path para las reseñas positivas y negativas, y la creación de vectores segun TF-IDF*

Se recomienda ejecutar un proceso previo de pre-procesamiento donde al menos aplique **tokenización, stem y eliminación de stop-words**. Divida el conjunto de datos en **70% para entrenamiento** y use los mismos para entrenar un clasificador **SVM** dentro del

operador de validación (validation). Almacene tanto el modelo entrenado como los términos procesados en un repositorio

Figura 6: *Procesamiento del texto segun los procesos de Tokenize, Stem, Filter Stopwords, todas en idioma Inglés.*

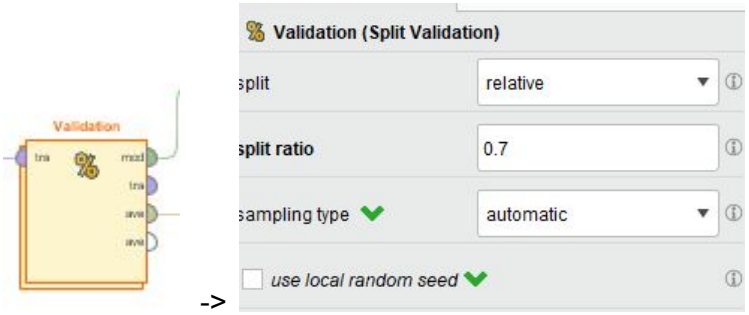


Figura 7: *Proceso de Validación de texto, con un porcentaje de 70% del texto para el aprendizaje y un 30% para el test.*

Figura 8: *Es importante guardar el modelo generado para utilizarlo con textos de pruebas*

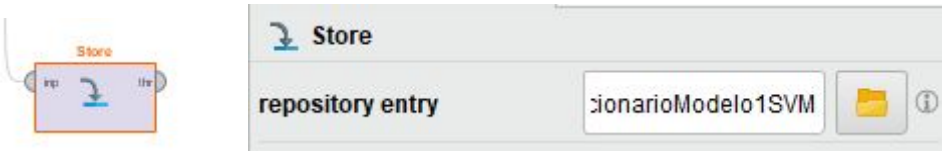


Figura 9: *Además del modelo generado, se necesita guardar el diccionario resultante del proceso "Process Documents from Files"*

Finalmente, se pide aplicar el modelo entrenado a dos revisiones de películas que no formaban parte de los datos de entrenamiento. Para la aplicación del modelo use la siguiente información

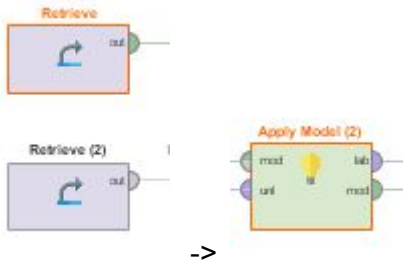


Figura 10: *El proceso de aplicar el modelo generado con anterioridad, se necesita nuevamente cargar el proceso y el diccionario generado en la Figura 8 y Figura 9*

Los términos de la fuente original que fueron pre-procesados y luego agregue las dos revisiones nuevas a las cuales se debe aplicar las mismas transformaciones (preprocesamiento) que en el caso de los datos entrenados anteriormente.

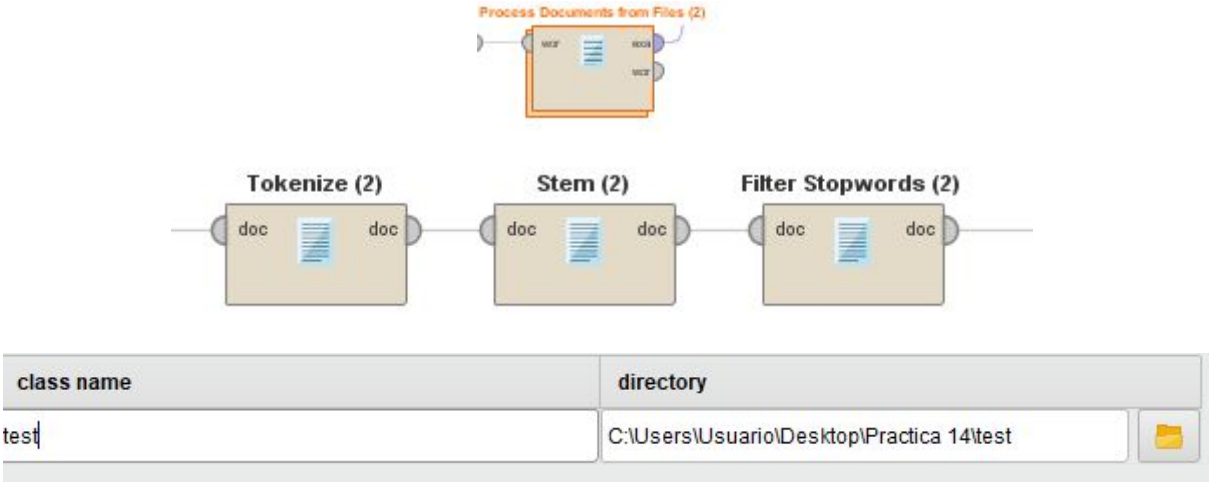


Figura 11: *Para validar la nueva entrada de prueba se genera un proceso "Process Documents from files" con los mismos subprocesos de preprocesamiento de texto.*

Finalmente, se aplica el modelo entrenado a la matriz de documento de término resultante y se obtiene un puntaje sentimiento para cada revisión de las películas ejemplo.

Resultados

Row No.	label	prediction(la...	confidence(pos)	confidence(neg)	metadata_file	metadata_d...
1	test	neg	0.453	0.547	hanselygretel...	Jul 1, 2019 1..
2	test	pos	0.528	0.472	themeg.txt	Jul 1, 2019 1..

Figura 12: *Resultados de Análisis de Sentimientos, obteniendo resultado negativo para "Hansel & Gretel" con una confidence negativa de 0.547 y positivo para "Themeg" con una confidence positiva de 0.528*

accuracy: 80.17%

	true pos	true neg	class precision
pred. pos	225	44	83.64%
pred. neg	75	256	77.34%
class recall	75.00%	85.33%	

Figura 13: *Resultado del modelo generado con las reseñas positivas y negativas de las figuras 3 a la figura 9, se tiene un accuracy del 80.17%*