[Salta al contenido principal](https://egela.ehu.eus/mod/page/view.php?id=4325208&inpopup=1#maincontent)

Description of NLP datasets

Datasets collected in this subdirectory are based on "natural language processing" tasks.   
They can be directly loaded in WEKA using the "Open URL" option in the "Preprocess" tab:

- the dataset **"spam versus no spam"** saves 1201 tweets collected by Oteara company.

This company is specialized in analysis of social networks, trends, opinions, comments... and offers its services to

companies which want to know the reputation of its product in the web. Each tweet is characterized by 17331 unigrams and bigrams,   
showing their presence/absence. The dataset was annotated by experts of the Oteara company, and currently shows   
1162 non-spam-tweets and 39 spam-tweets.

The dataset is spacially suited for cost-sensitive-learning scenarios, where the misclassification costs of both classes are not equal.

That is, in order to improve the detection ability of "spam" tweets, the missclasification cost of a spam as non-spam has to be   
bigger than a real non-spam missclassified as spam.

- **"spambase.arff"** is a classic, benchmark dataset in machine learning studies. The dataset is part of the UCI Machine Learning Repository.

http://www.ics.uci.edu/~mlearn/MLRepository.html

https://archive.ics.uci.edu/ml/datasets/Spambase

The classification problem is to determine whether a given email is spam or not. It covers the emails received by a single user.   
The set of attributes (unigrams) is limited by the creators. They are clearly defined in the head of the file.

- the dataset **"Asomo"** saves 150 posts collected by the Asomo-Socialware company.   
Posts were related to opinions about Telefonica company.

Each post is labeled in 3 different class variables: the willing to influence of the writter of the post (WI), sentiment of the post (S), subjective/objective (Sj).

The dataset was used in the following work:

J. Ortigosa-Hernández, J.D. Rodriguez, L. Alzate, M. Lucania I. Inza, J.A. Lozano (2012).

Approaching Sentiment Analysis by Using Semi-supervised Learning of Multi-dimensional Classifiers. Special Issue in Data Mining Applicacions and Case Studies. Neurocomputing journal, 92, 98-115

A link to the work can be found [here.](http://www.sciencedirect.com/science/article/pii/S0925231212001282)

In the original work, this dataset was used with a larger set of unlabeled samples in order to perform a semi-supervised learning approach.

An explanation about the construction and feature engineering process of the dataset is explained in the *"Asomo features - description.pdf"* file: it is really interesting.

Asomo company completed the preprocessing steps.

I have several doubts about the explained preprocessing steps, specially about the normalization of the features.   
That's why I upload two versions of the dataset.

- the **"Reuters"** subdirectory contains 4 of the most popular document-categorization problems of the well-known "Reuters" collection of documents.

A rich description about the "Reuters" document-categorization problems can be found [here](http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt).

In the original version, 12,902 stories had been classified into 118 categories (e.g., corporate acquisitions, earnings, money market).   
The stories average about 200 words length.

Each document in the dataset is finally represented by 24330 uni-grams (features), each of them showing the presence/absence of the   
word in the document.

We have selected 4 problems: the 4 most frequent categories.

Each of the problems is composed of two files: train (7770 documents)and test (3019 documents) archives, for the construction and   
validation of the classification models. These data splits are commonly used in the related literature.

In this way, each text-categorization problem is treated separately: a natural extension would be to deal with all the categories in a   
multi-label classification way, as each document can belong to various categories simultaneously.

- the **"Telecom\_Tweets**" dataset collects a set of tweets that discuss the opinion of the users about an anonymous telecom company.   
In the current form, each sample is characterized as an string-type feature which saves the tweet, coupled with the anotation of the   
opinion (class [variable](https://egela.ehu.eus/mod/glossary/showentry.php?eid=18228&displayformat=dictionary) to be predicted: positive or negative). In order to deal with this type of features, the WEKA software first needs   
to transform the string feature in the set of unigrams (further predictive features) of the samples-tweets.   
For this purpose, please consult the current WEKA filters:

Preprocess--> Filters--> unsupervised--> attribute--> StringToWordVector

and

Preprocess--> Filters--> unsupervised--> attribute--> NumericToBinary

- the **"IMDB"** dataset collects a set of movie-review data: 1000 positive and 1000 negative processed reviews. It is a popular   
and benchmarking dataset for sentimen analysis experiments. For more information, please visit:

http://www.cs.cornell.edu/people/pabo/movie-review-data/

In the current location, the 1000+1000 reviews are saved in 1000+1000 \*.txt format files. The current WEKA script, run from the   
CLI command line of the software when the files are saved in associated "neg" and "pos" subdirectories (in their root "directory-path"   
directory), converts the 1000+1000 reviews in a single WEKA format file:

> java weka.core.converters.TextDirectoryLoader "directory-path" > IMDB.arff

The output file needs further processing, as each sample is characterized as an string-type feature which saves the tweet, coupled   
with the anotation of the sentiment (class variable to be predicted: positive or negative). In order to deal with this type of features,   
the WEKA software first needs to transform the string feature in the set of unigrams (further predictive features) of the samples-tweets.   
For this purpose, please consult the current WEKA filters:

Preprocess--> Filters--> unsupervised--> attribute--> StringToWordVector

and

Preprocess--> Filters--> unsupervised--> attribute--> NumericToBinary  
The "IMDB.csv" file is the result of this preprocessing. Class variable appears in the first index position.   
  
- **"epinions4.arff.csv"** dataset contains the 1,382 posts in a csv format. It has two categories: Pos (reviews that express a positive or favorable sentiment) and Neg (reviews that express a negative or unfavorable sentiment). It has been preprocessed with WEKA's StringToWordVector filter to obtain the unigrams for each post. [Epinions.com](http://epinions.com/) is a website where people can post reviews of products and services. It covers a wide variety of topics. In this case, a set of 691 posts that recommend Ford automobiles, and 691 posts that do not recommend Ford automobiles. Although the text is several years old, it is similar to comments found on Epinions.com today. It has been obtained from a CMU University course: http://boston.lti.cs.cmu.edu/classes/95-865-K/HW/HW3/  
  
- **"UMIC-SA-training.csv.arff"** dataset can be opened directly in WEKA. It is based on the training dataset of the [following kaggle.com competition.](https://inclass.kaggle.com/c/si650winter11) It covers 7086 labeled sentences for a sentiment analysis task. You can preprocess it with WEKA's StringToWordVector filter to obtain the unigrams for each sentence. Then, by saving it in csv format, this can be opened by R for applying a supervised classification task for predicting the sentiment of the sentence (caret, h2o, etc. packages).   
  
**- "sts\_gold\_tweet.csv"** dataset was compiled and used in the following work:   
Saif, H., Fernandez, M., He, Y., Alani, H.: Evaluation datasets for twitter sentiment analysis. In: Proceedings, 1st Workshop on Emotion and Sentiment in Social and Expressive Media (ESSEM) in conjunction with AI\*IA Conference. Turin, Italy (2013).   
It can be found in the following link, with a brief explanation: https://github.com/pollockj/world\_mood/tree/master/sts\_gold\_v03  
It covers 2034 tweets annotated as positive (4) or negative (0). Three annotatoss agreed in the labels.   
You can open directly in WEKA as "csv". Take the following facts into account:   
- at first, the "polarity" feature is load as "numeric". It is recommended to apply the "NumericToNominal" filter over it (use properly the "attributeIndices" parameter of the filter).  
- at first, the "tweet" feature is load as "nominal". It is recommended to apply the "NominalToString" filter over it (use properly the "attributeIndices" parameter of the filter). After this, the "StringToWordVector" can be applied to transform the string feature in the set of unigrams.

- R's **"text2matrix.R"** function can help you building a document-term matrix, starting from raw text (.csv file format). The code of the function has a helpful set of comments. You will then need to append the label of each document as a vector (see the tutorial about the "tm-textmining" package). Its functionality is similar to WEKA's StringToWordVector function. The "tweets\_CEC2017.csv" file can be used as starting point.

- the **"Enron"** email dataset in the "multilabel" directory is a benchmark in *multilabel scenarios*. A complete explanation about it can be found in:

http://bailando.sims.berkeley.edu/enron\_email.html

The following link explains the different labels-annotations:

http://bailando.sims.berkeley.edu/enron/enron\_categories.txt

- **"Twitters-GreenPeace.RData"** R-format workspace saves a set of 81 tweets of @GreenPeace user (december 2014).   
They were collected using a procedure exposed in the related tutorial (ask Inaki Inza). The workspace saves other sets of tweets.   
Opening this workspace in R, the objects which save the sets of tweets are easily identifiable.   
  
 - **More NLP datasets** can be found in the following links:   
---- The SMS Spam Collection v.1 is a public set of SMS labeled messages that have been collected for mobile phone spam research.   
It has one collection composed by 5,574 English, real and non-enconded messages, tagged according being legitimate (ham) or spam.   
Detailed information: http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/ The WEKA file: fee.unicamp.br/~tiago/smsspamcollection/smsspamcollection.arff   
---- The "Sentiment140" dataset contains a huge collection of annotated tweets: https://www.kaggle.com/kazanova/sentiment140   
Sentiment140 is an exciting project: http://help.sentiment140.com/for-students  
---- The SFU review corpus on sentiment analysis. Detailed information: http://www.sfu.ca/~mtaboada/research/SFU\_Review\_Corpus.html The following link shows a complete exercise over it: http://jmgomezhidalgo.blogspot.com.es/2013/06/baseline-sentiment-analysis-with-weka.html A corpus in english: https://raw.githubusercontent.com/jmgomezh/tmweka/master/OpinionMining/SFU\_Review\_Corpus.arff A corpus in spanish: https://raw.githubusercontent.com/jmgomezh/tmweka/master/OpinionMining/SFU\_Spanish\_Review.arff

Última modificación: lunes, 14 de diciembre de 2020, 09:27