20 Newsgroups

The 20 Newsgroups data set

The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. To the best of my knowledge, it was originally collected by Ken Lang, probably for his *Newsweeder: Learning to filter netnews* paper, though he does not explicitly mention this collection. The 20 newsgroups collection has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering.

Organization

The data is organized into 20 different newsgroups, each corresponding to a different topic. Some of the newsgroups are very closely related to each other (e.g. **comp.sys.ibm.pc.hardware / comp.sys.mac.hardware**), while others are highly unrelated (e.g **misc.forsale / soc.religion.christian**). Here is a list of the 20 newsgroups, partitioned (more or less) according to subject matter:

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

Data

The data available here are in .tar.gz bundles. You will need tar and gunzip to open them. Each subdirectory in the bundle represents a newsgroup; each file in a subdirectory is the text of some newsgroup document that was posted to that newsgroup.

Below are three versions of the data set. The first ("19997") is the original, unmodified version. The second ("bydate") is sorted by date into training(60%) and test(40%) sets, does not include cross-posts (duplicates) and does not include newsgroup-identifying headers (Xref, Newsgroups, Path, Followup-To, Date). The third ("18828") does not include cross-posts and includes only the "From" and "Subject" headers.

- 20news-19997.tar.gz - Original 20 Newsgroups data set
- 20news-bydate.tar.gz - 20 Newsgroups sorted by date; duplicates and some headers removed (18846 documents)
- 20news-18828.tar.gz - 20 Newsgroups; duplicates removed, only "From" and "Subject" headers (18828 documents)

I recommend the "bydate" version since cross-experiment comparison is easier (no randomness in train/test set selection), newsgroup-identifying information has been removed and it's more realistic because the train and test sets are separated in time.

[7/3/07] I had originally listed the bydate version as containing 18941 documents. I've discovered that the correct count is 18846, of which rainbow skips 22. So the matlab version (below) represents 18824 documents. However, my rainbow2matlab.py script drops empty and single-word documents, of which there are 50 post-rainbow-processing, so you will find only 18774 total entries in the matlab/octave version.

Matlab/Octave

Below is a processed version of the 20news-bydate data set which is easy to read into Matlab/Octave as a sparse matrix:

- 20news-bydate-matlab.tgz

You'll find six files:

- train.data
- train.label
- train.map
- test.data
- test.label
- test.map

The .data files are formatted "docIdx wordIdx count". The .label files are simply a list of label id's. The .map files map from label id's to label names. Rainbow was used to lex the data files. I used the following two scripts to produce the data files:

- lexData.sh
- rainbow2matlab.py

[Added 1/14/08] The following file contains the vocabulary for the indexed data. The line number corresponds to the index number of the word---word on the first line is word #1, word on the second line is word #2, etc.

- vocabulary.txt

Other sources of information concerning this data set include

- Tom Mitchell's web supplement to his Machine Learning textbook.
- The CMU Text Learning group
- The UCI KDD 20 Newsgroups entry.

---

jrennie@gmail.com

Last modified: Mon Jan 14 13:38:35 2008