Big Data and ML Systems – Fall 2017
Assignment 2b

This assignment is open-ended where you will be given access to a popular text dataset and you will be given possible directions on types of tasks that can be performed on this dataset. You can perform any text analytics task on this dataset. Please explicitly document any code that you might use from online sources. The key aspect you will be evaluated on would be the quality of the results that you are able to derive on this dataset for the specific set of tasks you decide to pursue.

The text dataset is called the 20-newsgroups data.

Dataset: http://qwone.com/~jason/20Newsgroups/

Here is some more detailed information from the original data source.

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.

Your assignment goal(s):

You have complete freedom to determine the type of task you wish to perform on this dataset. Understand the dataset and determine which appropriate analytics tasks you wish to perform on this dataset. Corresponding to the task you can define your success metrics.

You will measured by three simple metrics: (a) type of task you wish to perform; (b) success of the task based on the task metrics; (c) your ability to extract interesting insight from this data based on this task.

The documents are categorized and there is separate training data. Labels are the classes

Here are some simple examples of tasks one can perform on this dataset:

(1) Document Classification using Bag of words and Naive Bayes

(2) Document Classification using word2vec and/or doc2vec

(3) Language modeling using RNNs (no need to build the RNN but just tweak parameters)

Feel free to post example tasks for others on Piazza.

Note: Please explicitly document any code that you might use from online sources.