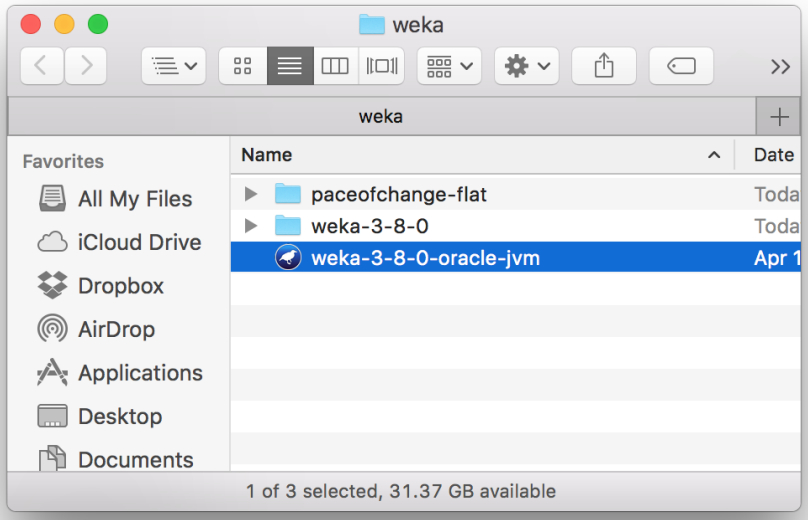
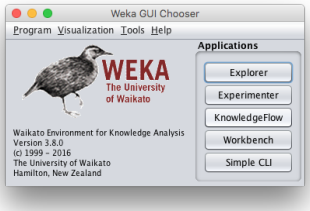
Open the Weka folder. Double-click on the “weka-3-8-0-oracle-jvm” file. (The numbers are as of 2016-06-13 and will probably change in the future.)

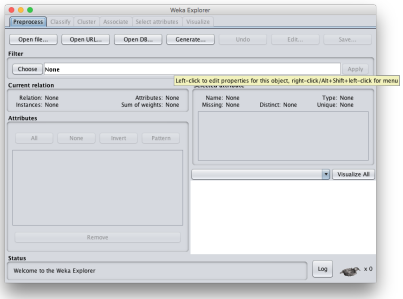


A splash screen like this will appear:

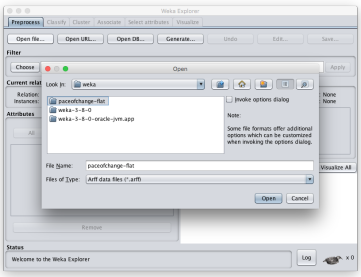
Then a GUI Chooser will appear. Click on **Explorer**.



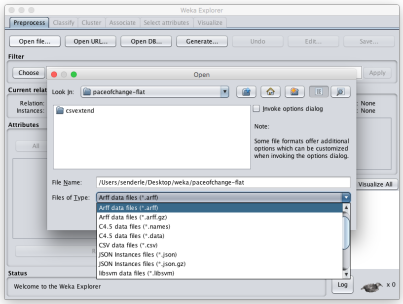
The Explorer interface will look like this:



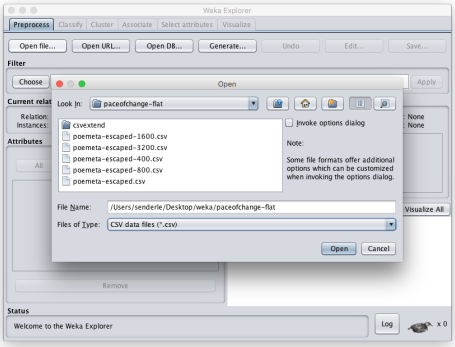
Select **Open File...** and double-click the “paceofchange-flat” directory:



The data to load will be in Comma Separated Values (CSV) format, which is conventionally denoted by a “.csv” extension. Under “Files of Type:” select “CSV data files (*.csv)”:



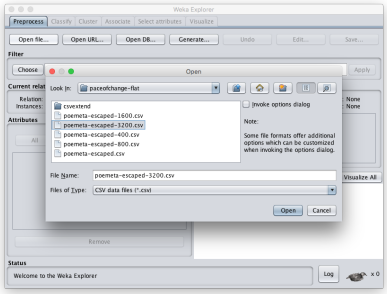
A set of CSV files will appear in the “paceofchange-flat” directory:



These files are based on data from Underwood and Sellers’ article “How Quickly Do Literary Standards Change?” (MLQ, forthcoming 2016 -- the data and a link to the article preprint are available at the [paceofchange github repository.](#)).

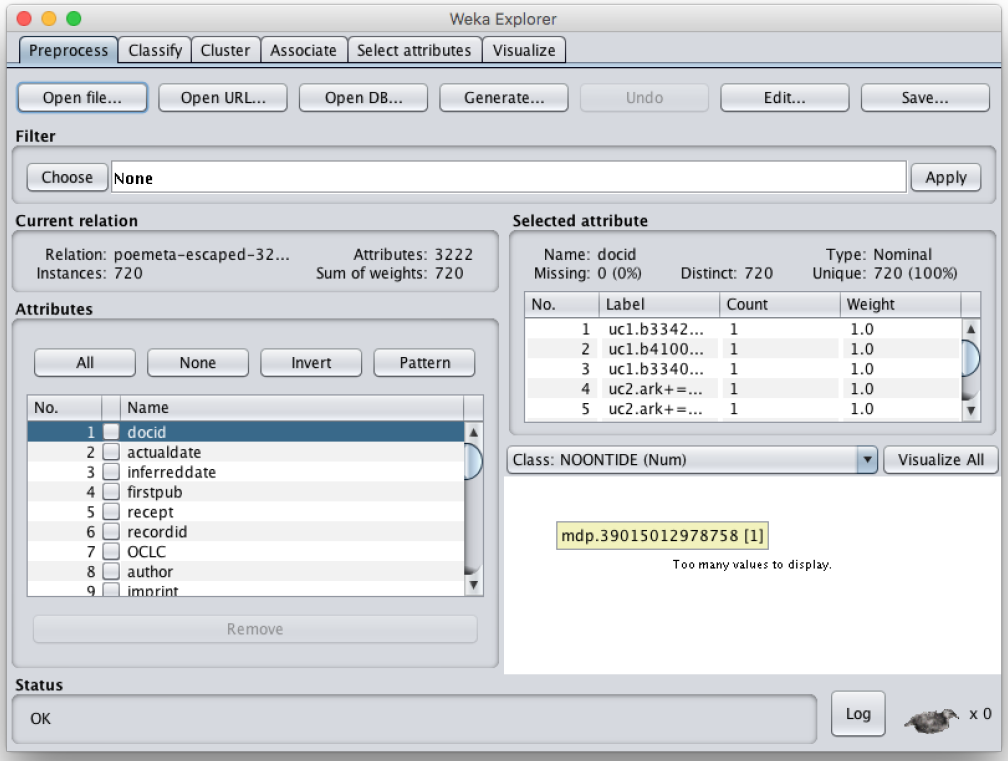
The data consists of word counts for each of 720 volumes of poetry. The numbers indicate the number of words used; “poemeta-escaped-3200.csv” contains counts of 3200 different words for each of the volumes, and we’ll start with that one.

Open the “poemeta-escaped-3200.csv” file:

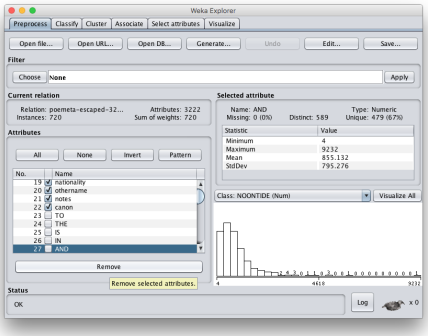


Now take a moment to look at the data preview window. Moving clockwise from the top, it contains the following frames:

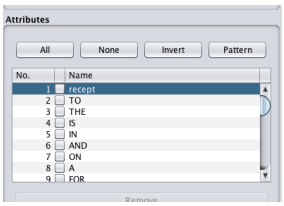
1) A **Filter** for selecting a subset of data points (i.e. rows or instances -- in this case, volumes of poetry).
2) A **Selected attribute** browser. This displays information about the selected attribute for the first few rows of data. It also allows you to examine the distribution of attribute labels over another set of labels. (In Weka, the second set of labels defaults to the last item; here, it's the word count for "NOONTIDE". That's not a very useful! If you'd like to tinker with this, try selecting "recept" -- that's the field indicating whether or not the given volume was reviewed in one of the publications selected by Underwood and Sellers).
3) A **Status** bar. This appears in all views (the tabs at the top, Preprocess, Classify, Cluster, and so on.) and will tell you if something has gone terribly wrong.
4) An **Attributes** selector. This allows you to select and discard attributes that you don't want to use for prediction. Because we have metadata and word counts stored together, we'll want to remove irrelevant metadata that might throw off our predictions; we'll see how to do that momentarily.
5) A **Current relation** information box. This just tells us about the data we're using. "Relation:" is just a fancy word for "table," and tells us which file we've loaded. "Instances:" tells us how many data points we have -- in this case, 720 volumes of poetry. "Attributes:" tells us how many items of information we have about each instance -- in this case, 3222, which includes all 3200 word counts, as well as 22 metadata fields. Ignore "Sum of weights:" for now.



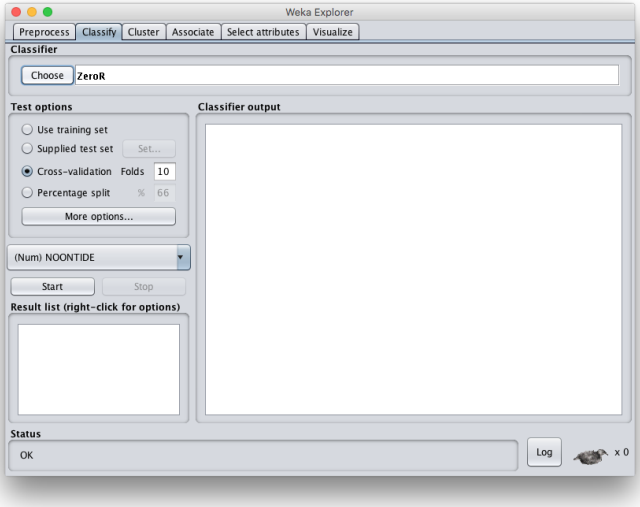
The classification algorithm we'll be using is called logistic regression, and Weka's implementation doesn't deal well with attributes that can't be represented as numbers. That means we have to remove all the metadata other than the one field we want to predict, "recept." Click the check box for each of the other 21 metadata fields. (You'll be able to tell the difference between metadata fields and word counts because in this data set, the names of the word count attributes are in ALL_CAPS.) To speed this up, you can highlight an attribute by clicking on it, and then use the arrow keys to move up and down. You can use the spacebar to check or uncheck the box for the highlighted attribute.



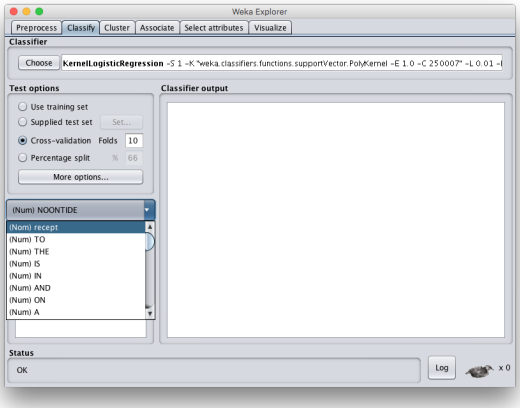
When you've selected the correct fields, click on the "Remove" button at the bottom. Afterwards, the **Attributes** frame should look like this:



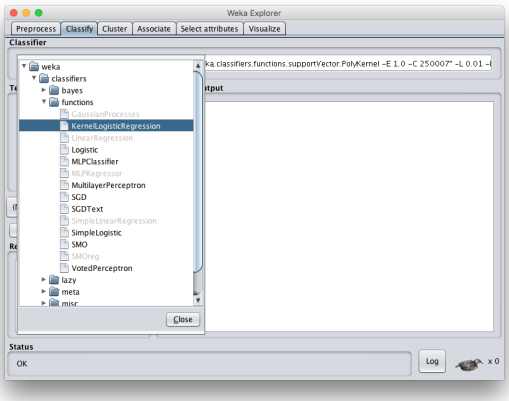
Now we can start classifying! Click on the **Classify** tab.



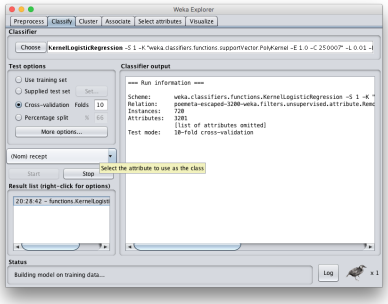
Select the “recept” field. This will be the field that the Logistic Regression algorithm will try to predict based on word counts.



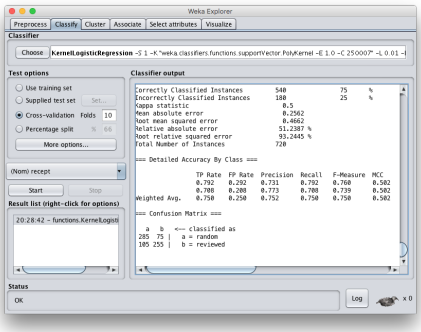
Click on the Choose button and select “KernelLogisticRegression” under “weka” > “classifiers” > “functions.” Don’t worry about what “KernelLogisticRegression” means for now.



And click start!



You'll see some activity, and eventually you'll get an accuracy readout.



Now try some other classifications. You might try predicting gender or nationality. Also try some other classifiers.