Sentiment Lab: Training a sentiment classifier for tweets using active learning

Summer School 2012

Overview: In this lab, you will be using a tool called Dualist to help you train a sentiment classifier for Twitter data. Dualist asks you to label a few documents and/or features for a text classification task. It then tries to learn from your labels, and identifies new documents and features for you to label so as to help the system learn. This is called *active learning*.

Warm up: To start, open up a web browser and go to: http://www.tweetfeel.com/ There are any number of web sites and companies that are trying to analyze sentiment on twitter, and this is one of them.

Try searching on different terms that you think people might be opinionated about: politicians, other famous persons, products, companies, sports teams, etc. Look at the results that you get and think about the following questions.

Q1. What sentiments is it getting *wrong*?
Q2. Why is it getting them wrong?
Q3. Is the sentiment always about the search term?
Q4. What would you guess tweetfeel is doing to classify sentiment?
Q5. How would you make it better?

Software: You will need to download and install two tools on your computer:

1. Dualist (http://code.google.com/p/dualist/)

Download and unzip:
http://code.google.com/p/dualist/downloads/detail?name=dualist-0.3.zip

B. Settles. Closing the Loop: Fast, Interactive Semi-Supervised Annotation With Queries on Features and Instances. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2011.

2. Play! web framework for java

Download and unzip:
http://download.playframework.org/releases/play-1.1.zip

Add the Play! directory to your path:
Mac: Open a terminal. From the command line:
> export PATH=$PATH:/Users/twilson/play-1.1
substituting the path to where you installed play-1.1

Windows: Instructions for editing your path can be found here:
http://www.computerhope.com/issues/ch000549.htm

Files: Create a new directory and ftp the following file from my home directory on the CLSP machines: ~twilson/tweets-summerschool-10k.zip

Twitter Sentiment:
Your goal is to train a system to classify the sentiment of tweets. A tweet expressing a sentiment is one in which **the person tweeting** is expressing a ***positive or negative opinion, emotion, evaluation, or judgment***. Usually a sentiment is about something (e.g., "*I love having a pet again"* is expressing a positive sentiment towards having a pet), but it is also possible for a sentiment to just express a feeling or emotion (e.g., "*I'm feeling blue today"*).

Below are some examples of Twitter messages that express a sentiment, as well as a couple of ***neutral*** messages (which don't express a sentiment):

EXAMPLE 1: Why I don't like Obama... http://t.co/6zBUcn9 #vrwc #tcot

- ***Negative sentiment*** toward Obama

EXAMPLE 2: President Obama ROCKED it tonight! Pass the jobs bill, Congress! #jobspeech

- ***Positive sentiment*** toward Obama

EXAMPLE 3: @Chika0dinaka: The republicans started this mess and its gonna take a republican to end it. Obama is great but not seasoned enough

- ***Both a positive*** (that he's great) ***and negative sentiment*** (not seasoned enough) toward Obama.
- ***Negative sentiment*** toward republicans.

EXAMPLE 4: What the Sun-Times has to say about Obama's jobs plan… http://t.co/wuaTqp1

- ***NEUTRAL*** (NO sentiment is being expressed)

EXAMPLE 5: New Poll: Obama Beats Perry and Romney http://t.co/SJUeS9J

- ***NEUTRAL*** (NO sentiment is being expressed)

You will be training a system to perform 3-way classification, classifying tweets as positive, negative, and neutral (no sentiment).

Running dualist:

1. From the command line, change directories into the previously installed dualist directory.
2. ./dualist gui
3. Open a web browser and go to the following URL: http://localhost:8080/
4. Select the Explore scenario
5. Enter the following information to start:
    a. Your name
    b. Data set: find the tweets-summerschool-5k.zip file
    c. Class labels: positive,negative,neutral
    d. Data Type: Tweets
    e. Instance queries: 15
6. Click Start Exploring
7. Dualist will now allow you to label tweets (on the left) and/or label features (on the right) for each class, submit your labels and retrain (long button at the top).
8. If you want to see how well the system currently is doing, click the "make predictions" button on the lower left.

Spend at least 30-45 minutes working on your classifier, or until you are satisfied. When everyone is done, I'll have instructions for running your best model on a test set, and we'll compare results.