Project Proposal

Deirdre Quillen

My plan for the project is to enter a competition on Kaggle to infer weather conditions from weather-related tweets on twitter. My plan is to use a hidden Markov model to represent sentences, where weather states are the unobserved variables, and the words are observed. Here are the basic facts about the problem statement of the competition:

* A training set and a test set are provided. This training set is the average of labels provided by humans taking a survey.
* There are 24 labels in total, falling under three categories: the user’s sentiment on the weather (positive, negative, neutral), the “when” category (past, present, future), and the weather type.

Here is my broad plan for this project. I plan to use either a hidden Markov model, or a conditional random field to model the data, after determining which model would be more appropriate (I think a conditional random field might be more appropriate to model dependencies throughout an entire sentence, but I’m not sure yet). However, because I have more familiarity right now with the hidden Markov model, I will describe a plan assuming I use a hidden Markov model.

1. Create a dictionary of common weather words and label them with a type, and a probability of being positively and negatively associated.
2. Include in this dictionary other common words describing positive or negative sentiment –e.g. “happy”, “lovely“, ”Ugh”, etc.
3. Given these word “states”: which could be for example, “positive”, “rainy”, “neutral”, “unknown”, I would use maximum likelihood and EM to train transition probabilities and parameters.
4. Test these parameters on the test data set.

There are many places where I could make this very simple model more complex, but this is my simple sketch. This is the link for the competition: <https://www.kaggle.com/c/crowdflower-weather-twitter>.