Assignment 1 Report

Pre-processing

1. Apostrophe Expansion: Replacement of contractions available at this **link**. (∼4.5k)
2. Tokenizer: reg expression r"[!?']+", text words and punctuations("!" & "?" only).(∼7k)
3. No stop words removal, no stemming, no lemmatization, no POS tagging. (∼23k).

TFidf Vectorizer

1. lowercase = true: (∼4k).
2. ngram = (1, 3): better than (1, 2) by (∼3k), no improvement for 3.
3. min_df, max_df = (5, 0.95): hyperparameter tuning, no reasoning, (∼0-2k).
4. Default values for other parameters.

Logistic Regression

1. Peer Interaction & self evaluations established that logistic regression is the only hope.
2. C = 5.0: better than C = 1.0 by (∼4-5k).
3. multi_class = "multinomial", solver = "sag": simple tuning, much better than others.
4. max_iters = 500, tolerance = 0.0000001: time limiting constraints.
5. Default values for other parameters.

Linear Regression

This can be considered as a post-processing step over the generated results. This idea is a hammer to the fact that we had to optimize the cost and not the accuracy by generating a floating point number.

Learning: Learning a linear model over the probability values for each class generated by the Logistic regression model to predict a number between 1 to 5. The fact that the cost function of the linear model is exactly equivalent to the cost that we had to optimize, it gave an improvement of about ∼26-27k.

Intuition: 1: 0.291, 2: 0.185, 3: 0.107, 4:0.241, 5:0.175
argmax gives 1 as the class but a linear model can generate 2.45 saying that a very bad sentiment and very good sentiment might turn out to be neutral.
Note that simply an expected value using the generated probabilities of the logistic model improves the cost by ∼19k.

Shashank Goel: 2016CS10332