PURDUE UNIVERSITY · CS 54701
INFORMATION RETRIEVAL

PROJECT PROPOSAL
Duc Viet Le
March 26, 2016

Proposal

1. What is the problem you are trying to solve? How is it done today? E.g., what is a good baseline (which could be done manually, but explain what this means.)
   Answer:
   Problem: To predict if movies review is either negative or positive.
   Some systems simply identify words in isolation, assign positive weight for words and sum up the weight to give prediction about the review. Other systems have human experts to classify the negative and positive feedback.

2. What is your approach, and why do you think it will be successful?
   Answer:
   a naive way to solve the problem is to use text categorization techniques (i.e. methods mentioned in classes: SVM, K-nearest neighbors, Naive Bayes) to predict outcome of a movie review (i.e. negative or positive.). We might add some
   Text categorization techniques are well-suited for this problem because the idea of text classifier is to process document, extract its feature, and tell us which class document belongs to.

3. How long will it take? Include a brief timeline for what you expect to have done when - at least include a mid-April “progress check". Also describe potential roadblocks/risks and how you will mitigate those risks.
   Answer: 4 weeks:

   - Week 1: Implement one or two of text classifier algorithms like Naive Bayes to test our dataset.
   - Week 2,3: finding different ways to improve our approach (i.e by using different online library like wordnet to identify negative/positive words.)
   - Week 4: Finish project, implement somekind of GUI for testing purpose.

   Potential risks: some reviews might have some hidden meaning that difficult to detect. Or review contains lots of positive words, but can be a negative review.
   To fix this, I can implement different approaches like text clurstering, or finds a more recent approaches to the problem (i.e. different learning techniques, using online library about sentiment lexicons.)

4. What is the measure for success? In other words, how will you measure how well you've solved the problem you defined?
   Answer:
   To measure how well the approach solves the problem:

   - *Training set: movie review dataset provided by Cornell University*
   - *Metrics: Precision/Recall/F1 Measure*
   - Cross-Validation: split training set into training and test set.
   - Choosing an unseen test set to test on our training set.

5. How will YOU be evaluated - what are the deliverables? Possibilities could include a web-accessible system that can be tested, an on-line live system demonstration, a written report, a (taped) oral presentation, or some combination of the above.
   Answer: To evaluate my approach, I will provide a written report of the result. Also, if possible, I will provide some access to test the perfomance of my implementation.