Web Intelligence

Assignment 2 – part 2 (of 3)

In this assignment you will be predicting, if people are going to buy “fine food” from Amazon.com based on sentiment analysis of reviews that have been submitted by other people in a social network. Some people in the network will submit a review and some will not. The challenge is to first predict the sentiment (that I will keep secret) from submitted reviews (summary and text), and then following predict if connected people that have not given a review are going to purchase “fine food” from Amazon.com.

You could represent your predictions as a file where each line contains three *tab-separated* fields (other representations are possible): the name of a person, the sentiment score (1-5), and the purchase decision (yes,no). Persons with a review should have * in the purchase field and, similarly, persons without a review should have a * in the sentiment field Example:

Peter	5	*
Jens	2	*
Tine	*	yes
Lone	*	no

Each week, I will supply a hint that will help you progress (and “inspire” your mini-project report).

1. **Part1 [Last week]**: The network has distinct communities. In one community, Amazon is running a very convincing mass-advertising campaign. In another community there is a very convincing person. In the remaining communities nothing special is happening (there will be somewhere between 2 and 10 communities in the network). It is important to identify communities in order to do a good job.

   How did you do the community detection? What were the results? Argue for you choice of algorithm or describe what you would have done, if you have had more time.

   A community “friendships.txt” file can be found in the Moodle resource folder. It should be self-explanatory (no reviews are added yet, but be ready to read in the reviews later).

2. **Part2 [This week]**: We have been fortunate enough to acquire some earlier fine food reviews from Amazon.com (I will distribute the file **SentimentTrainingData.txt** to you!). Build a sentiment classifier, where a score of 1.0 or 2.0 is a negative sentiment label and a score of 4.0 or 5.0 is a positive sentiment label.

The actual product is not important; all reviews are to be considered as reviews for the single category: “fine foods from Amazon”.

Use your sentiment classifier to evaluate the reviews in the friendships.reviews.txt file, which can be found in the Moodle resource folder. (Not all people in this file have given reviews!). Record the scores; It may be helpful to treat the scores of 1-2 as negative or 4-5 as positive! It may (or may not) be helpful to remember that there is one community where scores are particularly positive, because Amazon has run a very convincing mass-advertising campaign in that particular community.

What are the steps involved in constructing a sentiment classifier? What did you do; did you cut any corners? How good is your classifier, when you do cross-validation?