CS 406-01: Unstructured Information Processing

HW 3 (Given Sept. 18, 2019; Due Sept. 25, 2019)

Email your answers to the TA before midnight on the day it is due

Your answers must be entered in Google Classroom by midnight of the day it is due. If the question required a textual response, you can create a PDF and upload that. The PDF might be generated from MS-WORD, LaTeX, the image of a handwritten response, or using any other mechanism. Numbers in the parentheses indicate points allocated to the question.

---

1. In this assignment you will use the "Large Movie Review Dataset" available from `http://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz`. The dataset consists of reviews from IMDB. There are no more than 30 reviews per movie and the number of positive and negative reviews are equal (negative reviews have scores less or equal than 4 out of 10 while a positive review have score greater or equal than 7 out of 10. Neutral reviews are not included).

   The 50,000 reviews are divided evenly into the training and test set.

   Your task is to predict the sentiment of the test data. You may use any method you choose and all or part of the features in the data. Report your answer in terms of true positive, false negative and overall accuracy.

   Describe the method that you use in detail, the tools that you use and the above goodness figures of your classifier. You will be asked by the TA to demonstrate the working of your method and the numbers you report. (**100 points**)

2. Consider the weather (HOT – H or Cold – C) recorded over several sets of 10 days.

   H, H, C, C, H, H, C, C, H, H

   C, C, C, C, C, C, C, C, C, C

   H, H, H, H, H, H, H, H, H, H

   C, C, H, H, H, H, H, H, H, H

   Your friend's ice cream consumption on different weather days wee observed as follows,

0 cones, H (x 1 time)

1 cone, H (x 2 times)

2 cones, H (x 4 times)

0 cones, C (x 5 times)

1 cone, C (x 3 times)

2 cones, C (x 1 time)

Determine fully the Hidden Markov Model (i.e. start, transition and emission probabilities). (**50 points**)