Raw Simulation Results for SVM-Modeled Text Classification

Akshay Agrawal, Shane Leonard

November 14, 2014

Overview

The following tables outline the training and testing results applied to the `medicine_gold` dataset, using classifier `logistic` and the `EdxConfusion` data cleaner. Each section reports the result of processing the raw document in a different way before classifying the processed text as a bag of words.

Original Bag-of-Words Approach

The raw input documents were passed directly to the classifiers with no preprocessing.

Classification results for file ../../Confusion Datasets/medicine_gold . . . ; using classifier LogisticRegression and data_cleaner EdxConfusion +————————+————————-+————————+ | Label | Train: F1 | Test: F1 | +========================+=========================+=========================+ | knowledgeable | 0.91259 | 0.420034 | +————————+————————-+————————+ | neutral | 0.974787 | 0.882031 | +————————+————————-+————————+ | confused | 0.911703 | 0.378374 | +————————+————————-+————————+

Full Classification results for file ../../Confusion Datasets/medicine_gold . . . ; using classifier LogisticRegression and data_cleaner EdxConfusion +————————+————————-+————————+ | Label | Train: F1 | Test: F1 | +========================+=========================+=========================+ | knowledgeable | 0.664368 | 0.421321 | +————————+————————-+————————+ | neutral | 0.918744 | 0.873467 | +————————+————————-+————————+ | confused | 0.675589 | 0.382527 | +————————+————————-+————————+

Full2 Classification results for file ../../Confusion Datasets/medicine_gold . . . ; using classifier LogisticRegression and data_cleaner EdxConfusion +————————+————————-+————————+ | Label | Train: F1 | Test: F1 |

+===============+==============+============+ |
knowledgeable | 0.688346 | 0.402335 | +————————+————-+————+ |
neutral | 0.922442 | 0.869801 | +————————+————-+————+ | confused |
0.692077 | 0.368105 | +————————+————-+————+