IEMS 308 Homework 3

Zi Mei

Objectives:

We need to find out all the company names, percentage numbers and CEO’s names from BusinessInsider’s articles.

Methodology:

Sample names and percentage numbers are given; thus, supervised learning need to be employed.

First, load the raw data and split the whole txt into sentences. Then build the training dataset with part of the sentences that containing label words and take the other part as test dataset to evaluate model performance and the rest of the raw dataset will be used for my model to predict. Then employ tokenization to cut the sentence into single word, use stop words to filter out information less tokens. Conducting POS tagging of the tokens so we can know their properties.

Then, define the features: POS tagging of current, previous and next tokens; whether the token is uppercased or not; whether the first character is capitalized or not; whether contains numbers; the word’s postfix property, and whether the token is at the beginning or ending part of the sentence.

Next, build the CRF model with defined features to train my model on training dataset, and evaluate it on test dataset. Then use this model to predict and extract the target words on the rest of the raw data, and return the results based on the model’s own prediction. And when predict the percentage number, I add regex to better define the numbers with ‘%’ ,’percent’ and ‘percentage’ signs.

The model’s evaluation results are pretty good; however, percentage number prediction is not satisfying like the other name’s prediction. And probably more features need to be defined to get a more accurate result. The next step, I’ll try to find more features and maybe try different models to extract the results.

