IEMS 308 Homework 3

Text Analysis

Yiheng Pan

3051317

**Environment:**

Anaconda-Jupyter-python3.6

**Method:**

Sentence Segmentation: After downloading all the files, we use “nltk” package to take every sentence from the article.

Word Tokenization:

Take each word from the segmented sentences, and delete the useless stop word.

POS Tagging:

POS means part of speech. For each word, we want to know it is a noun or verb or something else. This is an important feature for future usep.

Create Training Set:

First, we need to preprocess the raw data by sentence segmentation and using regular expression, then delete all the stop words.

Next, we choose 60% data as training set to, then analyze their features and let our model to learn the features to catch the words by using “sklearn” package.

Then we use 20% validation data to evaluate the accuracy of our model. Last, we use the rest 20% data set as test data to catch the missing words which has the same feature as our model.

CEO:

Find name:

1. One capital word or two consecutive capital words: regular expression: ([A-Z][a-z]+) ([A-Z][a-z]+)?

Example: Peter Pan

1. No stop word.

Feature:

1. Part of speech: should be noun. (NNP)
2. The position of name: the first word’s index.
3. Each word is beginning from a capital letter.
4. The length of first word (Since some names only have a single first name, then the length of last name would be 0, we cannot set the length of last name as model feature)

Company:

Find company name:

1. One or more than one consecutive capital words: regular expression: [A-Z][a-z]+ ([A-Z][a-z]+)\*

Example: Panda Company

1. No stop word.

Feature:

1. The position of the company name: the index of first word.
2. The POS of the nearby words. (Front& back)
3. There are specific words in the sentence: “Company”, “Group”, “Co”, “Corp”, “Corporation”, “Inc”, “Ltd”.

Percentage:

Find:

1. Numerical number + percentage/ %

Example: 10%/10percentage/ten percent

Re: [0-9] %|[a-z]+ percent|[A-Z][a-z]\* percent|[0-9] percent

2. No stop word.

Feature:

1. The position of the percentage: the index of the first word.
2. The POS of nearby word. (Front& back)
3. There are specific words in the sentence: percent, %.

Model evaluate:

CEO:

precision recall f1-score support

avg / total 0.86 0.88 0.88 52483

Company:

precision recall f1-score support

avg / total 0.90 0.91 0.90 811349

percentage：

precision recall f1-score support

avg / total 0.96 0.97 0.97 1811