Project – COSC 6590_001

School of Engineering & Computer Sciences
A & M University at Corpus Christi
Spring 2024

Points: 20
Out: Mar 12, Due: Apr 22.

Objective: Build an NLP System to Identify the Sentiment of a Document by Document Scoring Method

It is a group (of two, or three, or four) project.

Given corpora **D** consisting of **TrainData.csv** and **TestData.csv**. **D** is defined as $\mathbf{D} = \{< \mathbf{x^i}, \mathbf{y^i} > |\mathbf{i} = \mathbf{1}...\mathbf{N}\}$, s.t. each $\mathbf{x^i}$ is the $\mathbf{i^{th}}$ document while each $\mathbf{y^i}$ is an associated class label in $\mathbf{y} = \{\textbf{positive}, \textbf{natural}, \textbf{negative}\}$ of the $\mathbf{i^{th}}$ document. Your task is to build a system by the document scoring method to identify the sentiment of a new document.

Build a System (8 points)

Design and Implementation

Text Preprocessing

- Delete punctuation marks
- Convert a letter from uppercase to lowercase
- Convert a verb into the base form, ...
- Determine a feature vector for the description of a document
  - The size of your feature vector should be great than 200

Build a Predicative Model (the System)

- Compute the tf-idf score for a feature **f** in the feature vector of a document **d**.

$$\mathbf{tf - idf(f, d) = tf_{f,d} * idf_f}$$

$$\mathbf{tf_{f,d} = log_{10}(count(f, d) + 1)}$$

$$\mathbf{idf_f = log_{10} \frac{N}{\sum_{d,f \in D} 1}}$$

  Where: **N** is the number of document in the training set.

- Compute the average $\mathbf{tf - idf}$ score for a feature **f** in a feature vector of a document in the same category $\mathbf{y_i}$ to form $\mathbf{v_{y_i}}$.

$$\mathbf{v_{y_i} = avg(tf - idf(f, d \in y_i)) = \frac{\sum_{f \in y_i} tf - idf(f, d \in y_i)}{\sum_{d \in y_i} 1}}$$

  Where: $\mathbf{y_i} \in \mathbf{y} = \{\textbf{positive}, \textbf{natural}, \textbf{negative}\}$

- The trained predicative model should be stored into a file.

Test the Predicative Model (Evaluation)

- Load your predicative model created on **TrainData.csv**.
- Run the predicative model on the test data **TestData.csv** for evaluation by:

$$\mathbf{score(q, v_{y_i})} = \frac{\sum_{\mathbf{f \in q}} \mathbf{tf - idf(f, v_{y_i})}}{\mathbf{||v_{y_i}||}}$$

$$\mathbf{\hat{y} = argmax_{y_i \in y} score(q, v_{y_i})}$$

- The system accuracy is obtained by the following accuracy measurement.

$$\mathbf{ACC_{D_{test}}} = \frac{\mathbf{1}}{\mathbf{|D_{test}|}} \sum_{\mathbf{i=1}}^{\mathbf{T}} \mathbf{L(\hat{y}^i, y^i)}$$ [1]

where $\mathbf{\hat{y}^i}$ is the assigned class label by the system (the predicative model) and $\mathbf{y^i}$ is the true class label of the document $\mathbf{x^i}$ in $\mathbf{D_{test}}$ and $\mathbf{T}$ is the number of documents in $\mathbf{D_{test}}$.

$$\mathbf{L(\hat{y}^i, y^i)} = \begin{cases} 1 & \text{if } \hat{y}^i = y^i \\ 0 & \text{if } \hat{y}^i \neq y^i \end{cases}$$

- The contingency table should be generated for the system evaluation.

Report (8 points)

- Write a 10 page of report on your project. The report should have
  - Each of group members as well the associated tasks
  - Introduction
  - Project description
    - ∗ The text preprocessing tasks and discussions
    - ∗ The creating of the feature vector – the dimension of the vector and the details of each of the features
    - ∗ The detailed predicative model (the trained parameters) discussions
    - ∗ The evaluation metrics
    - ∗ The contingency table and discussions
    - ∗ The system accuracy discussions
  - Experiments and results discussions
  - Further improvements

Oral Presentation (4 points)

- Project presentation
  - 20 slides of power point for a 20 minutes of oral presentation for each of the groups
  - Demonstrate your project in front of the class
- Presentation Date:
  - Apr 23. 3 - 4 groups
  - Apr 25. 3 - 4 groups

Submission

- Project submission files
  - ReadMe.txt – how to run your system
  - Project source codes
  - The screen shots of the data structure and the outputs of your system
  - Presentation Slides
  - Report

[1] Note, $\mathbf{D_{test}}$ = **TestData.csv**