In the name of Allah

Text Mining

Instructor: Dr. Fakhrahmad

Project: Semantic Textual Similarity Due Date: **1402/05/15**

Calculating the Semantic Textual Similarity (STS) is an important research area in natural language processing. Originally, the work on STS largely focused on similarity between short texts such as abstracts and product descriptions. It plays a significant role in many applications such as question answering, document summarization, information retrieval and information extraction. As an example, for certain types of question answering systems, having an accurate STS component is the key to success since the questions with similar meanings can be answered similarly. STS is also important in translation memories retrieval and matching. Translation memories help translators by finding in the database they maintain previously translated sentences, which are similar to the one to be translated, and retrieving their translations.

Given the growing importance of having a good STS metric, we decided to focus on this task as the final project. You are allowed to use any packages and data for preprocessing the texts and to obtain semantic similarity between individual words.

Experimental Setup Dataset

STS task has a long history at SemEval workshops and many datasets have been proposed. Evaluation data consist of pairs of sentences (**s**$_1$ and **s**$_2$) and the degree of their semantic similarity. For this assignment, you are provided a dataset with 4500 English sentence pairs as training data and 4927 English sentence pairs as test data. The sentence pairs were annotated by human assessors in ranges from 0 (on different topics) to 5 (completely similar). You can download the dataset from the following link:

- http://www.mediafire.com/file/7rknmz3hw0grej5/Dataset.zip/file

Evaluation Metric

Given two sentences s_1 and s_2, the task is to assess pairs of sentences in accordance with their degree of semantic similarity. Hence, the method should produce real-valued similarity score for each pair of sentences. The performance is evaluated in terms of Pearson correlation with human judgments. This evaluation metric has also been used in the SemEval competitions. Let X be the set of scores reported by the STS model and Y be the set of scores assigned by human assessors. The $x_i \in X$ and $y_i \in Y$ refer to the i^{th} elements in X and Y. The Pearson Correlation r is calculated as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where n refers to the number of sentence pairs, $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, respectively.

Note that you are asked to

- Prepare a report in PDF format about your model and the strategy you chose.
- Describe the implementation details carefully.
- Submit your answer before the deadline (1402/05/15).
- Zip all files and submit as TM-Project-(Your Student Number).zip

Hint

If you have any question about the project, please email to **dianati.shiraz@gmail.com.**