Paraphrase Identification

Revised 9/27/2022

Task:

given two sentences, build a (non-deep-learning) machine learning model to classify whether they express the same meaning or not.

Data and format:

Please download the data from GitHub: https://github.com/yinwenpeng/MLMidTerm

- train_with_label.txt (4077 instances)
- dev_with_label.txt (724 instances; 50% positive, 50% negative)
- test_without_label.txt (1000 instances; 50% positive, 50% negative)

Each row in the three files denotes one instance, i.e., the two sentences.
"train_with_label.txt" and "dev_with_label.txt" have four columns: instance_id, sentence_1, sentence_2, gold_label (0 or 1). Columns are separated by \tab.
"test_without_label.txt" has the same column order except that it has no the "gold_label" column

The required algorithms to use:

Logistic regression or Support Vector Machine (whichever you observe better performance)
Please note deep learning is not allowed in this midterm project.
You do not need to implement logistic regression or SVM; you can use public libraries, such as "sklearn", "LIBSVM", etc., to build a model.

The knowledge/skills this project looks for from you:

1) based on your understanding of this task, define features manually for this machine learning problem;
2) Potentially you may need to do feature scaling, etc., to improve the performance;
3) get familiar with using popular machine learning libraries;
4) get familiar with tuning the machine learning models in those libraries;
5) get hands-on experience with which model (e.g., LR or SVM) works better

Timeline:

- Starting date: 9/27/2022
- Submission deadline: 10/27/2022, 11:59pm (EST)

What and how to submit:

You are expected to submit **three** things:

1) "**YourFullName_test_result.txt**": Use your best model on "dev_with_label.txt" to test on the "test_without_label.txt" file and generate a new file named "YourFullName_test_result.txt" with two columns separated by \tab: instance_id, predicted_label (0 or 1)
   Each row in "YourFullName_test_result.txt" corresponds to one test instance.
2) "**YourFullName_system_description.pdf**": A pdf file that describes what you did, including i) what features you designed; ii) data preprocessing and feature preprocessing if applicable; iii) the algorithms&libararies you used; iv) the experience&lessons you learned in this project….(you are welcome to write other topics)
3) **Code**: Commit the code to your github and let me know the repository URL

Commit "**YourFullName_test_result.txt**" and "**YourFullName_system_description.pdf**" to the same github repository with the code and send the **repository URL** to wenpeng.yin@temple.edu by the deadline.

How will your submission be scored?

The system will be evaluated by accuracy on the test file (the test file with gold labels is not given to you). Your score will be

$$100\times\frac{your\ accuracy}{max\ accuracy\ in\ the\ class}$$

Please note

- The accuracy of a random guess baseline in this task is 50%, so a system with an accuracy lower than 50% will be scored 0;
- You will only get the score if all three required submissions are available. A result file with no system description or code URL will be scored 0.

Some hints:

1) you may read some papers (note that non-deep-learning papers were published at least 6 years ago) about paraphrase identification to check how they solved this problem;
2) Potential features such as "sentences of similar lengths tend to be paraphrasing", "sentences with more overlapping words tend to be paraphrasing", etc.