Natural Language Processing [CS4120/CS6120] Assignment 2

Instructor: Professor Lu Wang

Deadline: November 14th at 11:59pm on Blackboard

For the programming questions, you can use Python (preferred), Java, or C/C++. You also need to include a README file with detailed instructions on how to run the code and commands. Failure to provide a README will result in a deduction of points. **Violation of the academic integrity policy is strictly prohibited, and any plausible case will be reported once found**. This assignment has to be done individually. If you discuss the solution with others, you should indicate their names in your submission.

1 Parsing [14 points]

Using Version 3.8.0 or higher of the Stanford parser `https://nlp.stanford.edu/software/lex-parser.shtml#History`, parse the following corpus, where each file represents one genre of text:
`https://www.dropbox.com/s/vezkx8znrheti90/Brown_tokenized_text.zip?dl=0`
For sentences containing 50 words or less (punctuation does not count), obtain the part-of-speech tags, the context-free phrase structure grammar representation, and the typed dependency representation as shown in the following sample output: `https://nlp.stanford.edu/software/lex-parser.shtml#Sample`

1.1 [2 points]

Using the part-of-speech tags that the parser has given you, report the number of verbs in each file. Also report the part-of-speech tags that you are using to identify the verbs.

1.2 [2 points]

Report the number of sentences parsed; do so by searching for ROOT in either the dependency representation or in the context-free phrase structure grammar representation

1.3 [4 points]

Using the dependency representation (or the context-free phrase structure grammar representation) that the parser has given you, report the total number of prepositions found in each file. In addition, report the most common three preposition overall.

1.4 [6 points]

Take a look at the constituent parsing and dependency parsing results. List out two common errors made in each type of parsing results, and briefly discuss potential methods to reduce each type of error.

2 CKY parser (10 points)

Using the rules and probabilities below, draw a probabilistic CKY chart for the following sentence: **"An army officer ordered the American troops"**:

1. S → DP VP 1.0
2. Det → an 0.5
3. Det → the 0.5
4. DP → Det NP 0.7
5. DP → Adj NP 0.2
6. DP→ N N 0.5
7. NP → N N 0.8
8. NP → Adj NP 0.6
9. NP → American 0.2
10. NP → army 0.2
11. NP → officer 0.5
12. NP → troops 0.1
13. VP → V DP 0.4
14. VP → ordered 0.4
15. VP → troops 0.2
16. Adj → American 0.1
17. Adj → army 0.3

18. N → American 0.9
19. N → army 0.3
20. N → officer 0.1
21. N → troops 0.2
22. V → troops 0.1
23. V → ordered 0.9

3 Sentiment Analysis [40 points]

The following training data contains 9484 sentences of movie reviews taken from Stanford Sentiment Analysis Tree Bank dataset `https://nlp.stanford.edu/sentiment/`, where each sentence is rated between 0 to 4 (0 - negative, 1 - somewhat negative, 2 - neutral, 3 - somewhat positive, 4 - positive). The sentiment score is appended at the end of each sentence, separated by | .

The training data is available at: `https://www.dropbox.com/s/4tcyb2iefdr2jaz/train_data.txt?dl=0`

You can use the following tools for implementation:

1. TensorFlow `https://www.tensorflow.org/`
2. Theano `http://deeplearning.net/software/theano/`
3. Scikit-learn `http://scikit-learn.org/stable/index.html`
4. PyTorch `http://pytorch.org/`

In addition to output and results, please submit your code along with a README.txt file explaining how to run your code.

3.1 [10 points]

Train a Multilayer Perceptron to predict sentiment score . Using unigram features as input, call the training and testing functions for the Multilayer Perceptron from the tool. You do not need to implement the learning (i.e., back-propagation) algorithm. You should have an input layer, two hidden layers, and an output layer; the second hidden layer should have 10 nodes. Use 10-fold cross-validation to optimize any parameters (e.g. activation function or number of nodes in the first hidden layer). Use accuracy as the metric for parameter selection. Describe your parameter optimization process, and report the parameters for your best model.

3.2 [5 points]

Using the parameters for the best performing model from 3.1, re-train it on the whole training set, and report the accuracy on the **training** set.

3.3 [10 points]

Use pre-trained word embeddings GoogleNews-vectors-negative300.bin.gz from Word2vec `https://code.google.com/archive/p/word2vec/`, and compute the review feature vector by using the average word embeddings. Do the same thing in 3.1: Use 10-fold cross-validation to optimize any parameters (e.g. activation function or number of nodes in the first hidden layer). Use accuracy as the metric for parameter selection. Report the parameters for your best model. Then re-train the best performing model on the whole training set, and report the accuracy on the **training** set.

3.4 [5 points]

In addition to the word embeddings, add one type of features by your own design (e.g. POS tags distribution) to the model in 3.3. Then re-train this model on the whole training set, and report the accuracy on the **training** set.

3.5 [10 points]

Using the best model from above (based on results from 3.2, 3.3., and 3.4), predict the sentiment scores for all sentences in this test set: `https://www.dropbox.com/s/jf8mr7kgt3hfv6y/test_data.txt?dl=0` (contains 2371 sentences, one sentence per line).
Append your predicted sentiment score by the end of each line, separated by |, as shown in the training data.
Submit this file and name it labels.txt.

4 Summary Evaluation [36 points]

In this question, you will design classifiers to evaluate summary quality based on its *non-redundancy* and *fluency*. You will be given a training data and a test data, both in the following format:

1. Column 1: summary for a news article
2. Column 2: non-redundancy score [a score that indicates the conciseness of the summary]
   Non-redundancy scores range from -1 [highly redundant] to 1 [no redundancy].

3. Column 3: fluency [a score that indicates whether a sentence is grammatically correct in the summary]
   Fluency can range from -1 [grammatically poor] to 1 [fluent and grammatically correct].

For further discussion on the two aspects, please refer to `https://duc.nist.gov/pubs/2006papers/duc2006.pdf`, see Section 3.1 on "Grammaticality" and "Non-Redundancy".

Your task is to build classifiers of your own choice (e.g. Support vector regression, linear regression, neural networks, or other classifiers) on the train dataset to predict the non-redundancy and fluency of the summaries in the test dataset.

Here's the datasets:
Training set: `https://drive.google.com/open?id=1Z_VsJTe5M4N3_Q0LXtGisSnFsP2JulE4`
Test set `https://drive.google.com/open?id=1M4P1PgWdQNUHKeHwzLJsybbEKIZUh8CV`

In addition to output and results, please submit your code along with a README.txt file explaining how to run your code.

4.1 Building Non-Redundancy Scorers

4.1.1 [12 points]

Train your classifier with the following three features on training data, with summary as input and "non-redundancy" as gold-standard scores, and report evaluation performance on test data. For evaluation, please use metrics of Mean Squared Error (MSE) and Pearson correlation, both calculated between your classifier's predictions and gold-standard scores of samples in the test data.
Each feature implementation worths 3 points, and building classifiers worths 3 points.

1. Maximum repetition of unigrams: calculate the frequencies of all unigrams (remove stop words), and use the maximum value as the feature.

2. Maximum repetition of bigrams: calculate the frequencies of all bigrams, and use the maximum value as the feature.

3. Maximum sentence similarity: each sentence is represented as average of word embeddings, then compute cosine similarity between pairwise sentences, use the maximum similarity as the features. Use word embeddings GoogleNews-vectors-negative300.bin.gz from Word2vec`https://code.google.com/archive/p/word2vec/` as input for each word. Words in a summary that are not covered by Word2vec should be discarded.

4.1.2 [6 points]

Design two new features for this task. Add each feature to the classifier built in 4.1.1, and report MSE and Pearson correlation. You will get 2 bonus points if any of your proposed feature gets better MSE AND Pearson. You will get 4 bonus points if both features improve previous classifier's performance.

4.2 Building Fluency Scorers

4.2.1 [12 points]

Train your classifier with the following three features on training data, with summary as input and "fluency" as gold-standard scores, and report evaluation performance on test data. For evaluation, please use metrics of Mean Squared Error (MSE) and Pearson correlation, both calculated between your classifier's predictions and gold-standard scores of samples in the test data.
Each feature implementation worths 3 points, and building classifiers worths 3 points.

1. Total number of repetitive unigrams: count how many unigrams are the same as the previous unigrams. For example, for a summary "**The the** article **talks talks** about language understanding", the value should be 2.

2. Total number of repetitive bigrams: count how many bigrams are the same as the previous bigrams. For example, for a summary "**The article the article** talks about about language understanding", the value should be 1.

3. Minimum Flesch reading-ease score: use tool from `https://pypi.org/project/readability/` to get readability score for each sentence, and use the minimum value as the feature.

4.2.2 [6 points]

Design two new features for this task. Add each feature to the classifier built in 4.2.1, and report MSE and Pearson correlation. You will get 2 bonus points if any of your proposed feature gets better MSE AND Pearson. You will get 4 bonus points if both features improve previous classifier's performance.

NOTE

1. Replace tab or new line characters with space

2. Remove extra spaces

3. Stop words can be found in NLTK.

4. Mean squared error is the error of the results obtained from your classifier compared to the true values of the labels.
   The output ranges from 0 to 1 where

   (a) 0 indicates that your predicted values match the gold-standards perfectly.

   (b) 1 indicates that your prediction are not very accurate.

   To know more, visit: `https://en.wikipedia.org/wiki/Mean_squared_error#Interpretation`
   You can use sklearn package to compute the mean squared error: `http://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html`

5. Pearson Correlation Coefficient calculates the correlation between the predicted values and the gold-standards.
   The result ranges from -1 to 1 where

   (a) 1 indicates that your predicted values positively correlate with the gold-standards perfectly.

   (b) 0 indicates that there is no correlation between predictions and the gold-standards.

   (c) -1 indicates that your predicted values negatively correlate with the gold-standards perfectly.

   To learn more, visit: `https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#Interpretation`
   You can use sklearn package to compute the pearson coefficient: `https://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.pearsonr.html`