Project 1 for CS421 – University of Illinois at Chicago

Sachin G. Biradar: sbirad2@uic.edu
Chirag Naresh Soni: csoni3@uic.edu

SETUP:

1. Kindly run the Stanford Core NLP server on default port : 9000
   java -mx4g -cp "*" edu.stanford.nlp.pipeline.StanfordCoreNLPServer 9000

2. In the *execution* folder execute the following command
   bash run.sh
   After the script finishes running, please find the results in the output folder.

TECHNIQUES:

1. For **Length**, we separated the sentences using nltk's sentence tokenizer. Then the sentence was split with "," with respect to constraint of minimum 5 words (we had checked for various length and 4 was the best fit).

2. For **Spelling Mistakes**, we counted the number of words that do not appear in the enchant dictionary of "en_US" and "en_UK". Also, since we observed that enchant outputs some correct spelled words as incorrect, we used GloVe to be doubly sure about the spell check of the words in the training corpus. For example, if there are certain words that were marked incorrect by enchant and correct by GloVe, then they were stored in the separate list so as to compare these words with the words that occur later in the essays of test corpus. If yes (i.e they occur in the test corpus), then they won't be considered as incorrect. This helped us to prevent from the detection of incorrect spelling mistakes. Few other words like e.g. , Mr. , etc. are correct and shouldn't be marked as errors, thus we have separately handled the cases where if either of such words occur, they are left untouched. *Using GloVe ensures correct spelling mistakes to be encountered and not incorrect ones.*
   We also gathered all possible lists of ProperNouns such as Countries, Cities, Brands, Companies, Languages, Nationalities, Days and Months.

3. For **Subject Verb** agreement, we used the parse tree given by Stanford Core NLP. Parse Tree helped us to map a list of all violating rules after traversing the tree to the pairs of incorrect POS tags in each sentence. We then traversed the parse tree

to find the head noun and its corresponding verb to check if they disagree with the rules mentioned for subject and verb agreement. *The most important feature of using a ParseTree for this part is that it not only find the set of subject and the verb that are adjacent to each other (and violates the rule) but also those that are separated by many POS tags and violates the subject-verb agreement policy. This increases the chances of finding errors that are hard to catch otherwise.*

4. For **Verb agreement**, we used POS tagging technique to create bigrams and trigrams of only those words that matches the Verb list (verb list includes POS tags like "VB", "VBZhas", "VBZis", "MD", "VBD", "VBG", "VBN" and "VBP"). The resulted chunks of bigrams and trigrams are checked against the rules (hand crafted verb disagreement criteria). If either of the two (bigrams or trigrams) matches any rule defined for verb disagreement then, that accounts for an error. Also, we have checked for the verb tense agreement by creating chunks of the sentences with each chunk having a verb associated with it. We created set of rules that check for the chunks that have occurrences of improper tense.

Pattern of Errors

subject-verb disagreement rules such as ['NNP', 'VBP'], ['NNP', 'VB'], ['NN', 'VB'] etc. These sets of POS tags indicate that none of the two words having those tags are semantically correct if they are in the same sentence and thus accounts for an error.

verb rules include bigrams and trigrams such as ["VBD, VBG", "MD, VBZis", "MD, VBZhas", "MD, VB, VBD"] etc. Presence of any such rules denotes the occurrences of incorrectly framed sentences with the improper usage of verbs. Verb Tense agreement includes certain pattern of errors such as ["MD,VB,VB", "VBP,VB", "VBZ,VBG,VBZ"].

You can find the complete set of rules in autograde.py file.

Please find below the logic for finding scores from number of errors :

Calculation of the sub scores:

a. Number of sentences and length
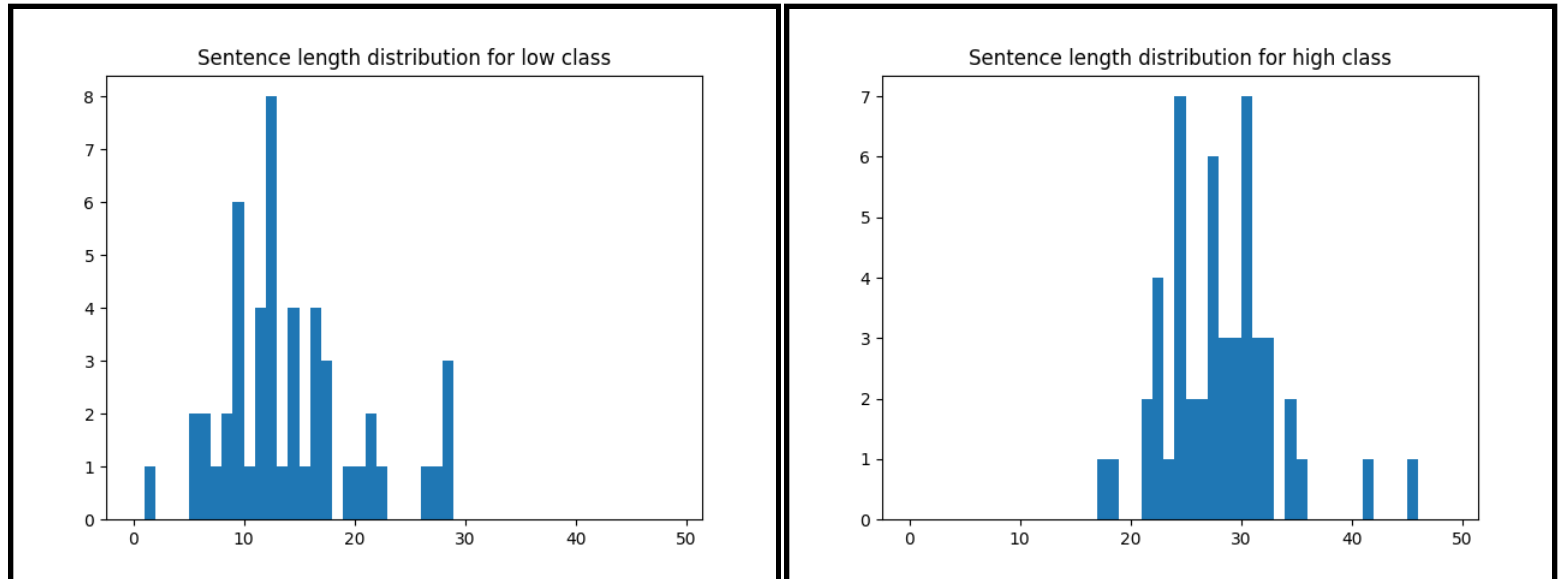
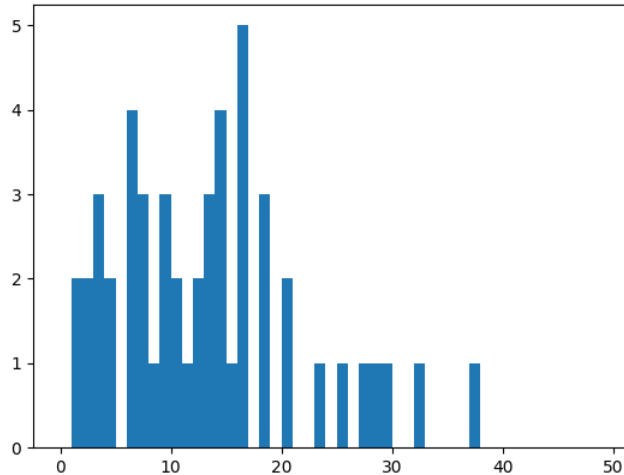**Histogram Distribution For The Length Of The Essays**

- The X-axis of the above histograms represents the possible length of the essays and the Y-axis represents the number of essays with a particular length.

- From the above distribution of the low and high essay types, following range of scores were inferred to score each essay on the basis of the appropriate length requirements.

Length (No of Sentences)	0 - 5	6 - 12	13 - 18	19 - 22	≥ 23
Score	1	2	3	4	5

Score Distribution Table For The Length Of The Essays

b. Spelling Mistakes
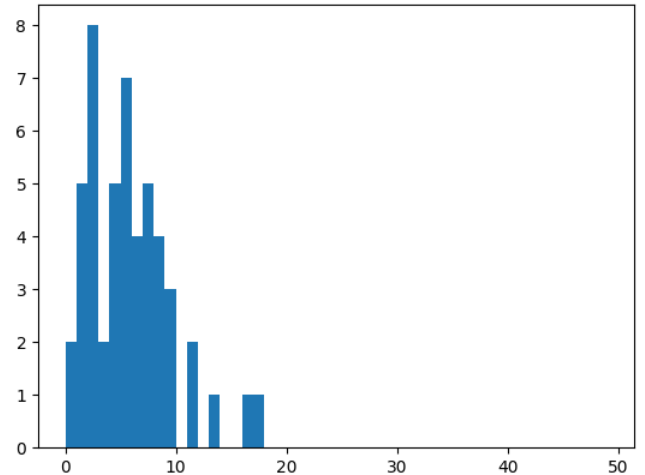
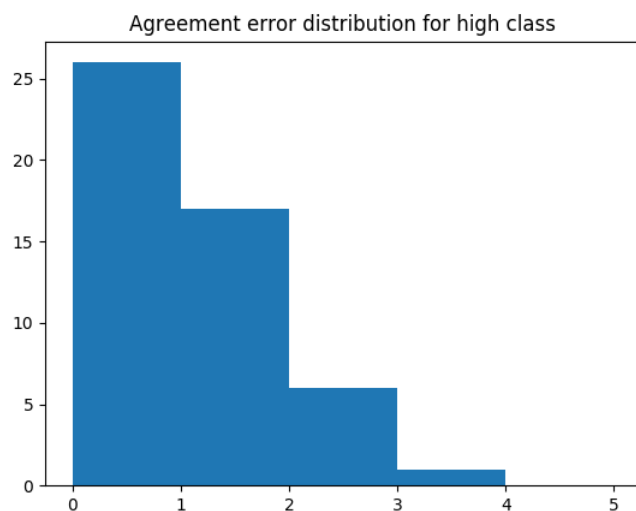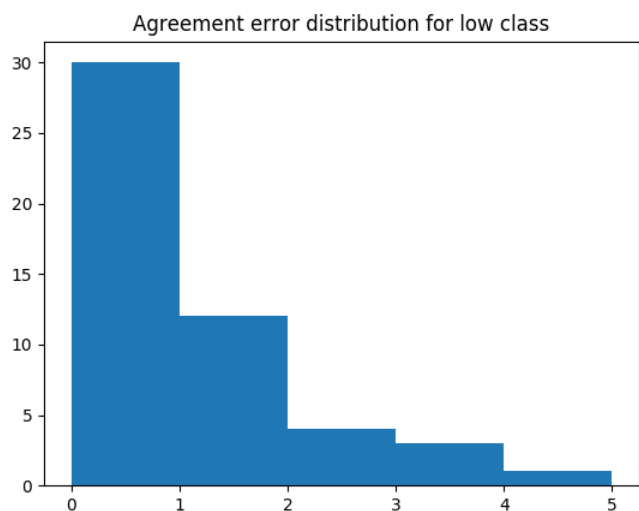**Histogram Distribution For the Spelling Mistakes In The Essays**

- The X-axis of the above histograms represents the possible number of spelling errors in the essays and the Y-axis represents the number of essays for a particular number of spelling mistakes.

- From the above distribution of the low and high essay types, following range of scores were inferred to score each essay on the basis of the occurrences of spelling mistakes in each essay.

No of Spelling mistakes	0 - 5	6 - 10	11 - 15	16 - 20	≥ 21
Score	0	1	2	3	4

Score Distribution Table For The Spelling Mistakes In The Essays

c. Syntax and Grammar
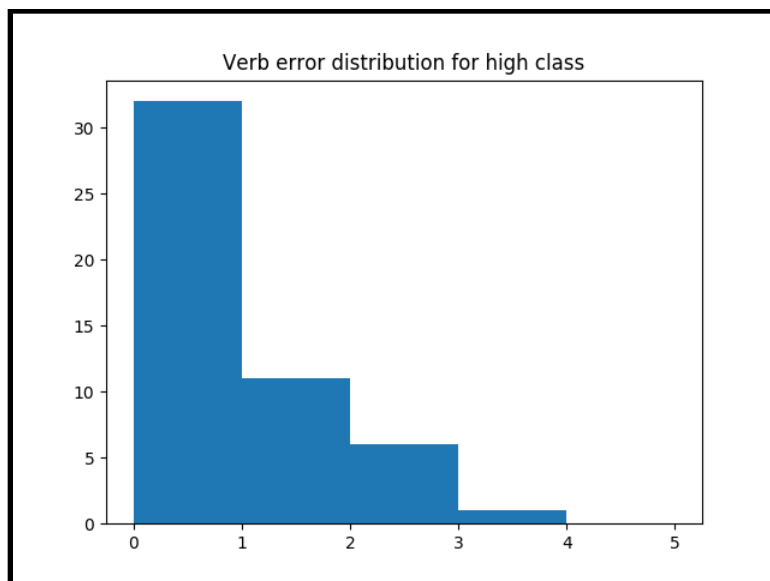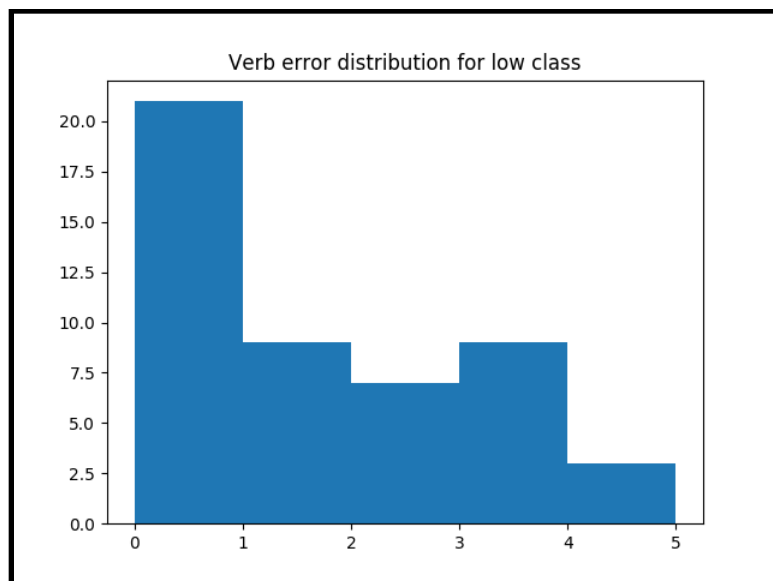## c - i. Subject - Verb Agreement

Histogram Distribution For The Subject - Verb Agreement Errors In The Essays

- The X-axis of the above histograms represents the possible number of subject - verb agreement errors in the essays and the Y-axis represents the number of essays for a particular number of agreement errors.

- From the above distribution of the low and high essay types, following range of scores were inferred to score each essay on the basis of the occurrences of subject - verb agreement errors in each essay.

No of Agreement Errors	1	2	3	4	>4
Score	5	4	3	2	1

Score Distribution Table For The Subject - Verb agreement Errors In The Essays

c - ii. Verbs

Histogram Distribution For The Verb Agreement Errors In The Essays

- The X-axis of the above histograms represents the possible number of verb errors in the essays and the Y-axis represents the number of essays for a particular number of verb errors.

- From the above distribution of the low and high essay types, following range of scores were inferred to score each essay on the basis of the occurrences of verb errors (including auxillary verbs, verb tenses) in each essay.

No of Verb Errors	1	2	3	4	>4
Score	5	4	3	2	1

Score Distribution Table For The Verb Errors In The Essays