Project 1 for CS421 – University of Illinois at Chicago

Name 1: vsakur2@uic.edu

Name 2: ufaiz2@uic.edu

----------Setup----------

---

1. Extract the .zip file
2. Install Maven using
    a. Ubuntu - “sudo apt install maven”
    b. Windows - https://maven.apache.org/install.html
3. Add all the files in input/testing/essays folder
4. Run ./run.sh in executable directory in terminal

----------Technique----------

---

1. Length of Essay: We use parse tree to obtain length of the essay. We use three cues for this:
    a. No. of root sentences ([ROOT [S) tags in the parse tree.
    b. No. of main verbs in the essay - We increment the sentence count if there more than one verb is used in the essay and the sentence does not have clause (SBAR tag).

   c. No. of conjunctions in the essay - If a sentence tag follows a conjunction tag ([CC … [S) then it implies that it connecting one sentence with another. We keep an account of this too.

2. Spelling Mistakes: For counting the number of spelling mistakes, a dictionary.txt file is created which has a large list of words.
   a. The word is considered as a spelling mistake if the POS tagging of the word comes out as NN and if the word is not present in the dictionary.txt file.
   b. An additional file called wordStopper.txt is introduced just to make sure that all the stop words are not considered as a spelling mistake.

3. Agreement: Agreement between the nouns and verb were calculated in the following way
   a. Using the Sentence Splitter from the StanfordParser, we are breaking the complete essay into sentences.
   b. We are then checking whether each sentence has the correct POS tags assigned to all it's words in that sentence so that they won't violate the Singular-Plural property between the Nouns and Verbs.
      i. For singular: {"VBZ", "NN", "NNP", "VBG", "VBN", "VB", "VBD"-"was"} and other tags can occur.
      ii. For Plural: {"VBP", "NNS", "NNPS", "VBG", "VBN", "VB", "VBD"-"were"} and other tags can occur

4. Missing Verb: Using the Sentence Splitter from the StanfordParser, we are breaking the complete essay into sentences.
    a. We are then checking whether each sentence has all the six main verb tag in them i.e. {“VBZ”, “VBP”, “VBG”, “VBN”, “VB”, “VBD”}
5. Sentence Formation:
    a. Getting count of fragments in an essay
    b. Patterns found from the training set like "the my" in sentence counted as an invalid sentence
    c. Counting sentences with only subordinate clause and no main clause
    d. SINV sentences checked for validity
    e. Identify run on sentence patterns
6. Text Coherency:
    a. Collecting all the pronouns from each sentence.
    b. Checking each pronoun in the sentence is available in the output of CoReferenceChain output.
7. Essay Validness - Topic Coherence: Using wordnet to identify synonyms of nouns used in essay and topic
    a. Get list of synonyms to the nouns in the topic sentence
    b. Get list of synonyms to the nouns in complete essay
    c. If the essay synonym intersects with noun then we say that the noun used is in line with the topic of essay and hence increment count

d. The score is mapped as ratio of nouns related to essay/ total nouns in essay

Error Patterns observed:

1. Some essays contained patterns of the form “ true.This ” i.e., the sentences did not have a space after a full stop. This gave rise to the problem of that pattern being tagged with a single POS tag. In order to overcome this problem, we preprocessed our file content to include a space after a full stop.
2. Some stop words are not recognized by the dictionary and hence a file name “wordStopper.txt” was created to ensure no stop word is being missed.
3. The verbs “was” and “were” come under the tag VBD and they are the only exceptions considered while looking for Noun-Verb agreement. The rest of the VBD tagged words do not violate the Noun-Verb agreement.
4. Some sentence had invalid word order of the form "the my"