Teammate 1 NAME: Shenfan Feng netid: sfeng37

Teammate 2 NAME: Yiming Wang netid:ywang626

TA: Barbara Di Eugenio

Section: CS 421

Final project part one

(a) Length of the Essay: Counting Sentences

- Methodology: To determine the essay's length, SpaCy's language model (en_core_web_sm) is utilized for its ability to effectively segment text into sentences. This is achieved by loading the essay text into SpaCy's NLP pipeline and using its doc.sents attribute to extract and count sentences.
- Technical Implementation: The number of sentences (num_sentences) is compared against predetermined thresholds for low and high scores (10 and 20 sentences, respectively). This count helps scale the score linearly between these thresholds to provide a nuanced assessment of essay length.

(b) Spelling Mistakes: Spell Checking

- Methodology: The SpellChecker library is used for identifying spelling errors. This library provides a list of words it considers as misspelled by comparing each word against a large English dictionary.
- Technical Implementation: Words from the essay are split and passed to the spell.unknown() method which checks their correctness. The count of these unrecognized words forms the basis of the spelling score. Note: The scoring logic here deviates from typical expectations, as more spelling errors uncharacteristically result in a higher score.

(c) Syntax/Grammar: POS Patterns and Agreement

c.i Subject-Verb Agreement:

- Methodology: Utilizes SpaCy's dependency parsing capabilities to identify verbs and their subjects within sentences. It checks agreement in number (singular vs plural).
- Technical Implementation: For each verb identified by the POS tag VERB, its related subjects (children with nsubj dependency) are examined. The agreement is validated based on POS tags (NN, NNP for singular and NNS, NNPS for plural) and corresponding verb forms (VBZ for singular, VBP for plural). Errors in agreement reduce the score.

c.ii Verb Tense/Usage:

- Methodology: Also leveraging SpaCy's parsing features, this component focuses on verb structures involving auxiliary verbs and main verb completeness.
- Technical Implementation: Each sentence is checked for the presence of auxiliary verbs (MD) and main verbs. The script ensures that auxiliary verbs are paired with appropriate main verbs or vice versa. Incorrect or missing verb pairings contribute to lower scores based on the prevalence of such errors relative to the total sentence count.