מסקנות - מה צריך לעשות

1. Debate’s Features and not paragraph features

Features:

1. Quotes:
    a. Length of quotes (in words, in letters) in proportion to the debate’s length (in the same variable).
    b. Num of paragraphs include quote(s), divide by the number of paragraphs.
    c. Maximal number of quotes in one paragraph.
    d. Length of quotes, in proportion to the paragraph’s time span.
2. Function words\stop words:
    a. Number of “real” words (not stop words) divide by the number of total words in the debate.
    b. “Powerfull” words\Active words like “should”, “would”, “might”, etc. TODO: make a list of at least 30 of these words. The feature is the number of these words in proportion to the number of words in debate.
3. Names:
    a. Names of people. TODO: find source that can extract names from paragraph. If can’t find, Implement or find in NLTK.
    b. Is the debater mentioned the names of the opposite side’s debaters. Try find their personal name, Last name.
    c. Same as above, but your partner name (supporting him\her).
4. Personal Experiences: TODO: list of words, or sentences to find personal experiences in
5. the paragraph. For example, “In the last few *, I saw”, “my”, “mine”, “I noticed”. TODO: at least 30 examples and implement a function that find these in the paragraph.
6. Make least of X most common words in english (for example, 50K), and count the number of words not include there. Motivation: is the debater use “fancy” words.
7. Science:
    a. Numbers, Statistics and Concrete information: find concrete data in the debate, like umbers, sentences like “in the last decade”, “grown”, “decline”, “exactly”. TODO: at least 30, function that implements that.
    b. Famous Universities: Make list of 30 most famous universities, and find them in the paragraph. The debater use researches and the reputation of the university.
    c. Science related words: lab, research, science, physics. TODO: make list of 30-50 words.
8. Distinct words in the debate in proportion to total words.
    a. STD of repeated word.
    b. Try ignore “stopwords”.
9. Question marks, exclamation marks. In proportion to number of sentences in the debate.
10. Talk to the audience: find it by sentences like “how many of you”, or in general “you” as referred to the audience.
11. Proportion in speaking time between the same side’s debaters:
    a. In sentences

    b. In time.
    c. In paragraphs.
12. Proportion in speaking time by the sides (same sub-features as above).
13. Famous: when find name (full name), try to search it in the web (for example: Wikipedia):
    a. If find it (it is famous person)
    b. Try find in the page words that describe the person, like “researcher”, “prof.”, “dr”.
14. Length of sentence and number of sentences in paragraph
    a. The std of word in the sentence
15. Crowd reactions, count them by reaction.
    a. Appluses
    b. Laugh
    c. Boo
    d. Shocked

Code fix:

1. Unify seperate p tags when parsing, when its the same speaker, and following sentences who were divided.

Research:

1. Search for list of english words (or using NLTK) and find all the tokens in some number of transcripts that not “valid” in english. Try to find misspelled words.

Questions:

1. Should we test the features for each side (unify the debaters in each side), or can we test somehow on each debater separately, so they won’t distruct the features of each other.
2. מילות ציווי באנגלית?
3. Should we test the features on each paragraph, or on the whole debate as input.
4. Numbers: how can we catch sentences like “twice the number” etc. Some pos tagging?
5. איך אפשר למצוא ״העלבה״ של דובר אחר, ציניות?
6. דיבור בבטחון על פי POS TAGGING או על פי שיטה אחרת?
7. האם יש טעם לנסות למצוא שגיאות כתיב בtranscript או אפשר להניח שהוא הועתק נכון, ו/או תוקן על ידי הקלדן.

חלוקת עבודה:

שלב ראשון. עד 21 למאי.

1. חלוקת פיצ׳רים:
    a. אביב:
        i. מדע

ii. פאנקשן וורד
iii. מילים נפוצות
iv. מילים שונות וסטיית תקן

b. ליאור:

i. שמות
ii. ציטוטים
iii. ניסיון אישי
iv. סימני שאלה, קריאה