Natural Language Processing – Assignment #3

Assignment description

This assignment includes a single task involving syntactic parsing and linguistic investigation. Please read the assignment description carefully and write the required answers in a document.

Task #1: using "that" as a subordinate conjunction (connecting sentence clauses)

In English, the word "that" can be used to mark various functions in a sentence. It is often used as a connector between two sentence clauses (פסוקיות) – main clause (פסוקית ראשית) and subordinate clause (פסוקית משנה). Consider the following example: "the authors thought [that] they should continue working on the project". Note that in this case, the conjunction "that" is optional: it can be omitted, while preserving the sentence meaning.

We are studying the phenomenon of "that" omission – for a fixed register (informal writing), in what cases people tend to use "that" as a connector, and when do they omit it? Is that an arbitrary decision (absolutely no rules can be found), or can we learn something about speakers' tendency?

Given a dataset of sentences collected from Reddit, your task is to (a) identify sentences where "that" conjunction connects a main clause to subordinate, and (b) identify sentences where it could be used for the same function but was omitted. We only consider the cases where the potential "that" follows a verb. As an example, sentences in (a) would include:

when someone in the audience asked about it, he explained **that** it was his day job

guess it took till now to realize **that** i don't value the org so much as I value the players

we don't know **that** he actually died though

And sentences in (b) would include:

she hopes [that] surgery will help her mental health...

guess [that] it took till now to realize that I don't value the org so much as I value the players

doesn't seem like he knows [that] it's wrong

We say that sentences in (a) are "explicit" usages of "that", and (b) are "implicit".

Note that there are additional syntactic functions "that" can be responsible for -- those are not in the focus of our work, and should not be included in neither (a) nor (b), as an example:

I hypothetically have never heard of **that** fallacy before

the book **that** my sister recommended, was excellent (here "that" does not follow a verb)

that would be a replacement card

You are given two files: `config.json` and `main.py`, where the config files contains input and output files locations. Do not make any changes to the code in the `main()`.

Your task is to implement the function:

```
def identify_explicit_and_implicit_that_clauses(filename):
    …

    return set(), set()
```

Implementation details and comments:

1. For simplicity, we are only considering cases where "that" is directly preceded by a single-word verb. Other (relatively rare) cases – like "it turns out that our work was ranked very high" ("turns out" is a phrasal verb) – can be ignored.

2. Multiple occurrences of explicit or implicit "that" can occur in the same sentence, in that case we only consider the same sentence once; note that sets are returned by the function.

3. We make use of one of the best contemporary syntactic parsers -- https://github.com/nikitakit/self-attentive-parser. Install the parser and use its `benepar_en3` English model. Read through tutorial and examples and learn how to work with it. A quick comment: some people do not work with the version integrated into spacy (as they recommend), but directly with the NLTK Tree objects, as in the notebook attached to this assignment. Feel free to use whatever you find more convenient.

4. Identifying roughly 1000 explicit and 1500 implicit sentences out of 10K input dataset, can be considered a reasonably good achievement.

5. Exploration: make use of the `SelectKBest` function you used in assignment #2 to find words that distinguish between the two sets of identified sentences. What can you say about these markers and peoples' tendency to use (or omit) "that" in certain cases? Explore the data, take some counts from the extracted datasets, and write your reflections in the document you attach to submission. Include your code in the version you submit, but don't invoke it in the final version.

6. Make sure your module runtime does not exceed 10 minutes.

Submission

Submit a single zip file – assignment3_ xxxxxxxxx_xxxxxxxxx.zip , where “xxxxxxxxx” stands for a student id. Please specify two student ids (your and your partner’s). It should include two files:

1. Your implementation for task #1 -- `main.py`.

2. A document with the results of your exploration in task #1.

Grading criteria include: correctness (the major part), code design, readability and documentation.

Good Luck!