DDL 2016.6.30 24:00

training set: pdtb-parses.json, pdtb-data.json(if need), raw (folder, if need)

testing set:pdtb-parses-dev.json

you should submit: your code, a final result file named “output.json” as well as a report showing your work

submit to: [rival2710@gmail.com](mailto:rival2710@gmail.com)

name of package(.rar or .zip): student ID\_name\_nlp.rar/zip

data format illustration:

There are constituent parse trees, dependency parse trees as well as words in the training file (pdtb-parses.json), and your task is to get the implicit senses as well as their precise argument spans.

Each line in the gold label file (pdtb-data.json) is a json line. In Python, you can turn it into a dictionary. (Similarly, you can turn it into HashMap in Java)

The dictionary describes the following component of a relation:

* Arg1 : the text span of Arg1 of the relation
* Arg2 : the text span of Arg2 of the relation
* Connective : the text span of the connective of the relation
* DocID : document id where the relation is in.
* ID : the relation id, which is unique across training, dev, and test sets.
* Sense : the sense of the relation
* Type : the type of relation (Explicit, Implicit, Entrel, AltLex, or NoRel)

The text span is in the same format for Arg1, Arg2, and Connective. A text span has the following fields:

* CharacterSpanList : the list of character offsets (beginning, end) in the raw untokenized data file.
* RawText : the raw untokenized text of the span
* TokenList : the list of the addresses of the tokens in the form of (character offset begin, character offset end, token offset within the document, sentence offset, token offset within the sentence)

What should the system output look like?

* The system output must be in json format. It is very similar to the training set except for the TokenList field. The TokenList field is now a list of document level token indices. If the relation is not explicit, Connective field must still be there, and its TokenList must be an empty list. You may however add whatever field into json to help yourself debug or develop the system. Below is an example of a relation given by a system.
* output\_file = open('pdtb\_trial\_system\_output.json')
* output\_relations = [json.loads(x) for x in output\_file]
* output\_relations[7]

reference:

[A PDTB-Styled End-to-End Discourse Parser](http://www.baidu.com/link?url=aVaJqjOenOD8DbfTyYs86kIqqDmrASrELxNTS2Fhf5RrAdkInbNm8adXFaRkhQ6yDpNpAMmp-wqbfRfvkAepoa)