TA has a hidden test set to make sure the algorithm is general enough.

Self-annotate 100 sentences, but show in .csv which 80 sentences used for training

CS372, Spring Semester, **2020**

School of Computing, KAIST

Homework #4

The U.S. National Library of Medicine (NLM) maintains a database called MEDLINE that contains more than 25 million references to journal articles in biomedicine, whose access is mediated by PUBMED. In this homework assignment, you are asked to go through the following steps for biomedical relation extraction: (1) Search the MEDLINE abstracts to collect 100 sentences, each of which contains at least one of the five verbs below (including their inflected forms), (2) provide annotations to these sentences for triples <X, ACTION, Y> to extract, (3) develop a relation extraction module based on a randomly selected set of 80 sentences and assess its performance, and (4) apply the module to the remaining (annotated but unseen) 20 sentences and assess its performance.
Additional constraints are shown below:

(1) There is no need for an automated method to collect such sentences from MEDLINE. However, you must include the verbs *activate*, *inhibit*, and *bind*, together with another verb with positive actions, such as accelerate, augment, induce, stimulate, require, and up-regulate, and another verb with negative actions, such as abolish, block, down-regulate, and prevent, with 20 sentences for each of the five distinct verbs. You should also prefer recency of publication, starting with the year 2020, limiting up to 30 sentences per year, up to 10 sentences per journal, and up to two sentences per organization, as identified by the affiliation of the corresponding author.

(2) The following shows some guidelines for your annotation of expected triples.

A. Inorganic phosphate inhibited HPr kinase but activated HPR phosphatase.
<Inorganic phosphate, inhibited, HPr kinase>,
<Inorganic phosphate, activated, HPR phosphatase>

B. All vasodilators activated K-Cl cotransport in LK SRBCs and HYZ in VSMCs, and this activation was inhibited by calyculin and genistein, two inhibitors of K-Cl cotransport.
<All vasodilators, activated, K-Cl cotransport>
<All vasodilators, activated, HYZ>

relax a bit
for grammar

<this activation, was inhibited by, calyculin> **OR** <calyculin, inhibited, this activation>
<this activation, was inhibited by, genistein> **OR** <genistein, inhibited, this activation>
<this activation, was inhibited by, two inhibitors> **OR** <two inhibitors, inhibited, this activation>

(3) In developing a relation extraction module, you should not use any of the third-party modules for coreference resolution, NER, relation extraction, event extraction, or parsers specifically made available for biomedicine, except for the NER module in NLTK. You should not use any of the external corpora for training. Report the performance in terms of Precision/Recall/F-score for (3) and (4).

As before, you should use techniques that can be implemented in Python and NLTK.

A Write a Python code for relation extraction.
B Show 100 sentences, annotated with expected triples, together with tags for 80 sentences.
C Discuss your results, to explain how you addressed the goal, and to suggest how you can improve the quality of the results further. This document must be in English.

All the requirements as underlined above must be composed by yourself and without help from anyone else. Any similarity of the results will be flagged for plagiarism and, if found sufficiently similar, penalized, up to, but not limited to, a failure to this homework.

Deadline for uploading your homework at KLMS: 3 June (11:59pm, STRICT)

Homework Submission Guidelines

1. Submission files
   - A CS372_HW4_code_[your ID].py
   - B CS372_HW4_output_[your ID].csv for the 100 sentences cited for their sources (PMID, year, journal title, organization) and with annotations for expected triples
   - C CS372_HW4_report_[your ID].docx and CS372_HW4_report_[your ID].pdf

2. Remarks
   - Use a maximum of 2 pages for your discussion C.
   - The code should include comments about your implemented idea.
   - For the implementation, you should not use external models.
   - Your code should be runnable in our environment.
   - Use 11pt font size and default margin/line spacing for your report.