**Assignment on TOC**

Based on the theory outcome, construct a DAG based on the following scenario. A Graph G(V,E) with certain vertices(V) and edge(E) should be constructed such that **vertex denotes** **persons** in the dataset provided and **edge denotes all the relations** associated with that person. To be simpler, a Graph should extract all the valid relations of a person.

Step by step procedure to construct DAG:-

1. **Lexical Analyzer:**

This phase also called as tokenization which is used to scan all the literals and store each and every word as a single token.

For ex before LA:- Lexical analyzer is the first phase in compiler design

After LA:- ‘Lexical’, ‘analyzer’, ‘is’, ‘the’, ‘first’, ‘phase’, ‘in’, ‘compiler’, ‘design’

This can be done with the help of Natural Language Processing(NLP) in python, where students are been taught in their first year.( <https://www.geeksforgeeks.org/tokenize-text-using-nltk-python/>)

Using NLP or regular expressions(regex), student needs to remove all the hyperlinks, html tags or tag related. Second, student should import stopwords package from python nlp and remove all the stop words for further process. (stop words are common English words which are irrelevant to construct any DAG. (<https://www.geeksforgeeks.org/removing-stop-words-nltk-python/>)

1. **Syntax Analyzer and Semantics:**

Student has to validate the resultant output from previous phase and needs to perform the text classification. In this assignment, user needs to find the files which are cyber security related and which are unrealted to cyber security. For this student has to perform **text classification using Support Vector Machine(SVM)**

(<https://medium.com/@bedigunjit/simple-guide-to-text-classification-nlp-using-svm-and-naive-bayes-with-python-421db3a72d34>)

Once text classification is done, student needs to consider the files which are only cyber security related and discard all the files which are not related. For SVM training, student needs to convert data into vectors with the help of **doc2vec model** i.e,. vector representation. To train, student needs to consider 100 files related to cyber security and 100 files unrelated to cyber security and train the data and calculate precision and recall to validate results.

1. **Graph Constructor:**

The output files from the above phase are considered as input to this phase. Inorder to construct a DAG, following three modules are important.

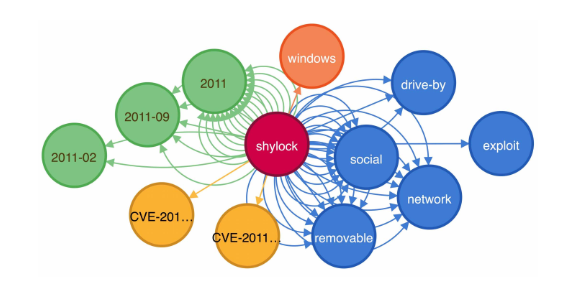
* + 1. ***Entity Tagger***:- this phase tags all the data given as input and classifies according to their category. (<https://www.kdnuggets.com/2018/08/named-entity-recognition-practitioners-guide-nlp-4.html>). This can be done by **NER(named entity recognition) model and coreference resolution** specific to cyber security domain.

For ex:- in normal context, generic NER detects ransomware as Noun., but in cyber security context, ransomware should be detected as malware.

* + 1. ***Relation Builder/ Symbol Table Generator***: - This phase helps to identify all the relations associated with an entity (person or object). Student needs to form a table based on the following relation types and identify the same relations throughout the dataset provided.

|  |  |  |  |
| --- | --- | --- | --- |
| **Entity** | **Entity** | **Relation** | **Label** |
| person | Organization | Work | WOR |
| person | Country/location | Address | ADD |
| person | Date | Detection | DET |
| person | Mobile/email | Communication | COM |
| person | Website/internet | Browsing | BRO |
| Person | Money | Finance | FIN |

* + 1. ***Relation Extraction:*** - It is the process of extracting semantic relationships between entities. Since multiple relations might exist in a document, relations become edge in a graph and entities shall become vertex. If there is more than one relation with an entity, multiple edges are created between single pair of nodes. For ex:- in the following graph, shylock is a malware and the relations with shylock malware is given in detail like shylock targeted windows os, infected through removable disk and its common vulnerability(CVE) exploits with detection date etc.



**Note: - Dataset and code will be provided to the students, where students need to optimise the code and refine the results in upcoming days**

**For Clarifications**,

D Paul Joseph,

Senior Research Fellow, VIT

Ph: +91 9394341910

[pauljoseph.d@vit.ac.in](mailto:pauljoseph.d@vit.ac.in)