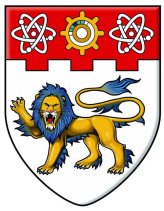

NANYANG
TECHNOLOGICAL
UNIVERSITY

ASSIGNMENT

API Recognition in Online Forum

CZ4045 Natural Language Processing

2016/2017 Semester 1

NANYANG TECHNOLOGICAL UNIVERSITY

1 Objective

The objective of this assignment is to let you getting familiar with the main components in an end-to-end NLP application, the challenges faced by each component and the solutions. Through this assignment, you shall also get hands on experiences on various packages available for NLP tasks.

2 Assignment Format

1. This is a group assignment. Each group has 4 to 5 students.
2. One report is to be submitted by each group and all members in the same group receive the same grade.
3. You may use ANY programming language of your choice, *e.g.,* Java, Python, C#, C++.
4. You may use any NLP and Machine Learning library/software as long as the license allows free use for education and/or research purpose. Example packages are listed below.
   - All-in-one library: NLTK (Python), LingPipe (Java), GATE (Java), Stanford NLP(Java), OpenNLP (Java)
   - HTML/XML processor: jsoup (Java), Python HTML/XML tools (Python)
   - Conditional Random Fields (CRF) library: CRF++ (C++), CRFsuite (C++, Python)
   - Machine learning library: Weka (Java), CMU Rainbow (C), SVM^{light} (C), libsvm (C++, Java), milk (Python)

3 Assignment (100 marks)

The task is to develop an application to recognize the APIs mentioned by users on Stack Overflow `http://stackoverflow.com`. Stack Overflow is a forum dedicated to programming related questions and answers.

The assignment consists of the following components: Dataset Collection (10 marks), Dataset Analysis and Annotation (30 marks), API Recognition (20 marks), Evaluation and Error Analysis (20 marks), and Application (20 marks)

3.1 Dataset Collection (10 marks)

For data collection, you may get a data dump from `https://archive.org/details/stackexchange` and then select a subset of questions (with answers), or write your own crawler to collect questions/answers from the website. Your data set shall contain at least 100 threads of discussion with at least 500 posts. A post can be either a question or an answer.

Your report shall show the statistics of the dataset used in the assignment, including number of questions, number of answers, and the distribution of the answers *e.g.,* what is the percentage of questions having one, two, or more answers, and the length of the posts in number of words.

3.2 Dataset Analysis and Annotation (30 marks)

Stemming. From the dataset, identify the top-20 most frequent words (excluding the stop words) before and after performing stemming. Stop words are the words that are commonly used but do not carry much semantic meaning such as *a*, *the*, *of*, *and*. For each of the top-20 most frequent stems, list their original words from which the stem is reached. Discuss your results.

POS Tagging. Randomly select 10 sentences from the dataset, and apply POS tagging. Show and discuss the tagging results.

API Annotation. In discussions on Stack Overflow, users often mention APIs. For example, in this sentence "*Will String.trim() remove all spaces on these sides or just one space on each?*" the API named String.trim() is mentioned. Select at least 100 posts, each of which contains at least one API mention, and then annotate all the API mentions in the selected posts. The annotated dataset will be used as ground truth in the next section. You may use `http://brat.nlplab.org/` for your annotation task. Note that, before the annotation, the group needs to clearly define "what is an API mention" in user posts.

3.3 API Recognition (20 marks)

Your task is to recognize all API mentions from given user posts obtained from stack overflow. Your group may design different techniques to recognize the API mentions. API mention recognition can be considered as a form of *named entity recognition*, and Conditional Random Field (CRF) is often used for this task. You may use the annotated data to train a CRF classifier to recognize the API mentions. Note that, there is NO need to develop your own CRF implementation. You may use any of the CRF libraries.

3.4 Evaluation and Error Analysis (20 marks)

To evaluate the effectiveness of a model for named entity recognition, a cross-validation is often applied. In K-fold cross validation, the annotated data is partitioned into k non-overlapping portions. Among them, $k-1$ portion are used to train the recognition model and the remaining one portion is used to evaluate the effectiveness of the model. For instance, if $k = 4$, then each time 3/4 of the annotated data are used in training, and the remaining 1/4 of the annotated data are used for evaluation. By using each of the 4 folds as evaluation data, the model is evaluated on all annotated instances. Then the performance of the model can be evaluated by using *Precision*, *Recall*, and F_1. Report the *Precision*, *Recall*, and F_1 of your model, and analysis the errors (*e.g.,* case studies on false positives and false negatives).

3.5 Application (20 marks)

Now your group has developed an API recognition tool using the annotated data. If you apply the model to all the posts in the dataset you have prepared earlier, all the API mentions can be recognized with certain level of accuracy. Your group shall brainstorm and suggest how the recognized results can be used in any applications to make programmers' life better.

4 Submission of Report and Source Code

4.1 *Progress Report in Softcopy (Not Evaluated)*

A progress report shall be submitted on or before 10 Oct 2016, 11:59pm. The purpose of the progress report is to give a clear picture to all group members on what have been achieved and what are planed to achieve. The progress report will NOT be evaluated. The progress report shall use the same template as the final report.

4.2 *Final Report in Hardcopy*

- The hardcopy report must be submitted on or before **31 Oct 2016** (Monday, Week 12). The report shall be formatted following the ACM SIG Proceedings Templates (either MS Word or Latex), **maximum 8 pages** including appendix if any.[1] DO NOT include in your report all the source code and complete results sets. However, you must include *code snippets* which are important for the main functions of your system. You should cite all third-part libraries used in your assignment.
- The report shall be printed in double-sided format whenever possible. A plastic cover or ring-binding leads to 2% penalty.

4.3 *Final Report in softcopy, Source code, and documentation*

- A CZ4045.zip file containing the following files and folder shall be submitted: Report.PDF, Readme.txt, SourceCode.
  - Report.PDF shall be the same as the hardcopy report submitted.
  - Readme.txt shall include
    - A link to download the third-party library if you used any in your assignment
    - A link to download the datasets used in your assignment, one for the 500 posts and the other for the 100 annotated posts.
    - An installation guide on how to setup your system, and how to use your system (*e.g.,* command lines, input format, parameters).
    - Explanations of sample output obtained from your system.
  - SourceCode folder shall contain all your source code. The dataset and the libraries shall **NOT** be included in the softcopy submission to minimize the file size.
- Softcopy submission deadline: ***31 Oct 2016 11:59PM***. Late submissions are allowed but will be penalized by 0.5% every calendar day (until zero). The softcopy can be submitted for at most three times, only the last submission will be graded and time-stamped.

[1] `http://www.acm.org/sigs/publications/proceedings-templates`