CS 6320 – Natural Language Processing
Fall 2017
Dr. Mithun Balakrishna
Course Project

A. Project Steps and Deadlines:

- **Project Group Formation**:
  - Due by **Sunday, October 22nd 2017, 11:59pm**
  - A maximum of three (3) students per project group
  - The group should decide on an appropriate group name
  - One group member should submit a document containing the group name and the group member information i.e. Group name and Group member names, via eLearning
    - Please name the document following the convention "ProjectGroupInfo-GROUPNAME.pdf", where GROUPNAME is your project group's name.
    - Submit the document to the "Group Information Submission" assignment inside the "Final Project" folder listed in the course home page on eLearning.
    - Students that want to work on the project individually should also submit this document
  - Students that need help to form a group should meet the Instructor on **Thursday, October 19th 2017** at **8:15pm** in the class room (CR 1.202)
    - Students that want to work on the project individually do NOT need to do this
- **Project Demo**:
  - Due date: **TBA**
  - Demo sign-up details: **TBA**
  - Submit your project source code and report via eLearning before your group's allocated demo session:
    - One group member should submit a single zip file containing the following via eLearning:
      - Project source code/script file(s)
      - A ReadMe file with instructions on how to access the project demo
      - Project report in PDF or MS Word document format.
    - Please name the zip archive document following the convention "ProjectFinalSubmission-GROUPNAME.zip", where GROUPNAME is your project group's name.
    - Submit the document to the "Project Final Submission" assignment inside the "Final Project" folder listed in the course home page on eLearning.

- Please hand over a hard copy of the project report before the start of your group's demo session with the TA

B. Project Report

Please write a project report (5 to 10 pages) with the following details:

- Problem description
- Proposed solution
- Full implementation details
  - Programming tools (including third party software tools used)
  - Architectural diagram
  - Results and error analysis (with appropriate examples)
  - A summary of the problems encountered during the project and how these issues were resolved
  - Pending issues
  - Potential improvements

C. Project Description:

For the project, you need to implement a semantic search application that will produce improved results using NLP features and techniques. Your project should implement a keyword-based strategy and an improved strategy using NLP feature and techniques. The following are the tasks that need to be performed:

1. **Task 1**: Create a corpus of News articles. Your corpus should contain at least:
    - 1,000 articles
    - 100,000 words

    Note: you are free to download freely and publicly available News articles corpora from public websites such as: http://www.nltk.org/nltk_data/

2. **Task 2**: Implement a shallow NLP pipeline to perform the following:
    - Keyword search index creation
        - Segment the News articles into sentences
        - Tokenize the sentences into words
        - Index the word vector per sentence into search index such as Lucene or SOLR
    - Natural language query parsing and search
        - Segment an user's input natural language query into sentences
        - Tokenize the sentences into words
        - Run a search/match with the search query word vector against the sentence word vector (present in the Lucene/SOLR search index) created from the corpus
    - Evaluate the results of at least 10 search queries for the top-10 returned sentence matches

3. **Task 3**: Implement a deeper NLP pipeline to perform the following:
    - Semantic search index creation
        - Segment the News articles into sentences
        - Tokenize the sentences into words
        - Lemmatize the words to extract lemmas as features

        - Stem the words to extract stemmed words as features
        - Part-of-speech (POS) tag the words to extract POS tag features
        - Syntactically parse the sentence and extract phrases, head words, OR dependency parse relations as features
        - Using WordNet, extract hypernymns, hyponyms, meronyms, AND holonyms as features
        - Index the various NLP features as separate search fields in a search index such as Lucene or SOLR
    - Natural language query parsing and search
        - Run the above described deeper NLP on an user's input natural language and extract search query features
        - Run a search/match against the separate or combination of search index fields created from the corpus
    - Evaluate the results of at least 10 search queries for the top-10 returned sentence matches

    Note: you are free to implement or use a third-party tool such as:

        1. NLTK: http://www.nltk.org/
        2. Stanford NLP: http://nlp.stanford.edu/software/corenlp.shtml
        3. Apache OpenNLP: http://opennlp.apache.org/

4. **Task 4**: Improve the shallow NLP pipeline results using a combination of deeper NLP pipeline features

D. Project Point Distribution

1. Max points available: 100 points
2. Division of points:
    a. Group information: 2 points
    b. Project implementation and demo: 90 points
        i. Task 1: 10 points
        ii. Task 2: 20 points
        iii. Task 3: 45 points
        iv. Task 4: 15 points
    c. Project Report: 8 points