CS 6320 – Natural Language Processing
Fall 2018
Dr. Mithun Balakrishna
Course Project

A. Project Steps and Deadlines:

- **Project Group Formation**:
    - Due by **Tuesday, October 16th 2018, 11:59pm**
    - A maximum of two (2) 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 **Tuesday, October 16th 2018** at **8:15pm** in the class room (ECSS 2.203)
        - 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 an Information Extraction application using NLP features and techniques:

Input:

- Set of information templates

  Examples:

  - Template #1:
    BUY(Buyer, Item, Price, Quantity, Source)

  - Template #2:
    WORKS(Person, Organization, Position, Location)

- Set of natural language statements

  Example:

  - Statements #1:
    Amazon.com Inc. will acquire Whole Foods Market Inc. for $13.7 billion, a bombshell of a deal that catapults the e-commerce giant into hundreds of physical stores and fulfills a long-held goal of selling more groceries.

  - Statements #2:
    Jeff Bezos is best known as the founder, chairman, and chief executive officer of Amazon.

Output:

- Filled information templates

  Examples:

  - Template #1:

    BUY("Amazon.com Inc.", "Whole Foods Market Inc.", "$13.7 billion", "1", "Whole Foods Market Inc.")

  - Template #2:

    WORKS("Jeff Bezos", "Amazon", "founder; chairman; chief executive", "") The following are the tasks that need to be performed:

1. **Task 1**: Create a set of information templates:
   - At least 10 information templates
   - At least 40 information properties

2. **Task 2**: Create a corpus of natural language statements:
   - At least 50,000 words

3. **Task 3**: Implement a deeper NLP pipeline to extract **at least** the following NLP based features from the natural language statements:
   - Tokenize the FAQs and Answers into sentences and words
   - Lemmatize the words to extract lemmas as features
   - Part-of-speech (POS) tag the words to extract POS tag features
   - Perform dependency parsing or full-syntactic parsing to parse-tree based patterns as features
   - Using WordNet, extract hypernymns, hyponyms, meronyms, AND holonyms as features

   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**: Implement a machine-learning, statistical, or heuristic (or a combination) based approach to extract filled information templates from the corpus of natural language statements:
   - Run the above described deeper NLP on the corpus of natural language statements and extract NLP features
   - Implement a machine-learning, statistical, or heuristic (or a combination) based approach to extract filled information templates from the corpus of natural language statements
   - Evaluate the results of at least 10 filled information templates for each information templates

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: 5 points
        iii. Task 3: 40 points
        iv. Task 4: 35 points
    c. Project Report: 8 points