CS 6320 – Natural Language Processing
Fall 2016
Dr. Mithun Balakrishna
Course Project

A. Project Steps and Deadlines:

- **Project Group Formation**:
  - Due by **Sunday, October 23rd 2016, 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 **Saturday, October 22nd 2016** at **9:45am** in the class room (ECSS 2.206)
    - Students that want to work on the project individually do NOT need to do this
- **Project's Short Description**:
  - Due by **Saturday, October 29th 2016, 11:59pm**
  - Please write a short description (1 to 2 pages) of the project, providing the following information:
    - Problem Description
    - Proposed Solution and Implementation Details
      - Baseline system
      - Improvement strategy
      - Examples
      - Programming tools (including third party software tools to be used)
      - Architectural diagram
  - Please submit the document to the "Project Short Description Submission" assignment inside the "Final Project" folder listed in the course home page on eLearning.

- **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
  - Baseline system
  - Improvement strategy
  - Examples
  - Programming tools (including third party software tools used)
  - Architectural diagram
  - Results
  - 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 application (of your choice) that will produce improved results using NLP features and techniques. Your project should implement a naïve strategy and an improved strategy using NLP feature and techniques. The following are the tasks that need to be performed:

1. Create a corpus of tagged examples that will be used for evaluating the naïve implementation and your improvement strategy. Examples:
    - For an email-spam detection system, collect 100 to 1000 emails and manually tag them as “spam” or “not-spam”.
    - For a sentiment detection system, collect 100 to 1000 sentences and manually assign each sentence to sentiment tag (e.g. positive, negative, and neutral)
2. Implement a naïve strategy to produce some preliminary results. Your naïve strategy implementation can be an implementation of very simple algorithm to produce the desired output. This system should produce a measurable result and work end-to-end. However, the performance of this simple baseline system need not be very high. It should just prove the feasibility of your NLP application. Example:
    - For an email-spam detection system, the simple baseline implementation can look for certain keywords that are indicators of spam and categorize emails as spam or not spam. Similar approach can also be applied for Sentiment Detection, Topic Detection, etc.
3. Finally, implement new algorithms/techniques that will provide better results than your baseline system. The improvement strategy should include the following mandatory NLP features:
    - Two (2) Lexical Features: word tokens, lemmas, stems, etc.

      Please note that this requires some type of word tokenization algorithm to be implemented or used from third-party tools.
    - Two (2) Syntactic Features: POS, syntactic phrase types, syntactic headwords, syntactic patterns, dependency parse patterns, etc.

      Please note that this requires some type of POS tagging, Syntactic parsing, or dependency parsing algorithm to be implemented or used from third-party tools.
    - Two (2) Semantic Features: thematic relations, semantic relations, etc.

      Please note that this requires some type of semantic parsing or role labeling algorithm to be implemented or used from third-party tools.
    - Your system improvements should be based on manual rules, machine learning (or similar methods), or some combination.

D. Project Point Distribution

1. Max points available: 100 points
2. Division of points:
    a. Group information: 3 points
    b. Project Short Description: 7 points
    c. Project implementation and demo: 80 points
    d. Project Report: 10 points