AIT626 – Introduction to Natural Language Processing The Term Project Guide

This is a self-defined **teamwork** project designed for the students who are required to prepare an NLP system that solves a well-defined NLP problem. The technologies are NOT limited to those studied during the class. Students are encouraged to apply other technologies beyond the course content. Every project needs up to **3 students** to team up. Each student in a team will be evaluated by other members in the same team. Every student in one team may have different credit based on their own overall contributions and performances.

General Instructions

You are free to pick your own topic. You must have your topic approved by the instructor. Turn in a paragraph long description of your proposed topics via the Blackboard before you start working on your proposal.

Some project ideas are as follows:

- Extend your implementation of a solution to one of the class assignments. You must have some set of specific functionality extensions in mind and be ready to articulate them, why they are important, what they will add to your existing solution, etc.
- You can browse the web for publicly available NLP corpora, e.g., https://github.com/niderhoff/nlp-datasets, and identify a data set that would serve as a testbed for you. Example data sets in the above link include document classification corpora, spam detection corpora, sentiment analysis corpora, etc. Linguistic Data Consortium is also a great resource for NLP corpora.
- Any other NLP topic that you find interesting.

For your topic, you must know what a typical baseline solution looks like and how well it performs. You must also know the state of the art and its performance level. You must run and evaluate your system on your data, then present us your solution and its performance, along with a comparison to the baseline and the state of the art.

Your system must run on the SW/HW platform as described in your README file and final report. It must be clearly documented and implemented in such a way that the instructor and/or GTA can easily understand and run your system.

Objectives

- Prepare an NLP system that solves a well-defined NLP problem
- Select a suitable dataset for analysis
- Apply tools and methods of NLP and/or data analytics technologies
- Develop skills in applying NLP and/or data analytics techniques to particular tasks
- Design and Develop a preliminary NLP system based on the requirements
- Learn how to write scientific/technical documentations

- Learn some project management techniques

Deliverables

The NLP Project is a course-long teamwork effort consisting of several separate deliverables. You are strongly encouraged to work ahead on these deliverables. You may submit them early, if you would like, but they will not be assessed until after their respective due dates.

This project **(30 points)** consists of 3 deliverables:

- Check Point 1 **(3 point) – Project Topic and Project Proposal**
- Check Point 2 **(7 points)** – **Progress Report**
- Check Point 3 **(4 points)** – **Project Presentation** (in-class presentation)
  (9 points) – **Final project Report** (10-15 pages, single space, 11 or 12 font size)
  (7 points) – **A Working System** (one zipped file)

Due dates: check the latest class schedule in the blackboard.

Submission

Please submit all your files with the file names "**AIT626_Team#_NameInitials_*.docx/pptx/zip**" (e.g., * could be ***proposal.docx***, ***progress.docx***, ***PPT.pptx***, ***final.docx***, ***sys.zip***). For example, **AIT626_Team3_NJ_KB_proposal.docx**.

Please submit to Blackboard under your team group.

Note that your submission may not be graded if you do not follow the required file naming convention.

Grading

The whole project will be graded based upon scope, complexity, quality and deliverable.

Specific Requirements / Templates for Deliverables

Read the sections below for the specific requirements of each deliverable. **Use the following structures as the templates for your reports and submissions.**

1. check Point 1: Project Topic (*submitted in **one 1-page Word file***)

- Free to pick your own topic but need to get approved by the instructor.
- Turn one or two paragraph description of your proposed topic before you start to work on your proposal.

2. Check Point 1: Project Proposal (*submitted in **one Word file***)

- **Cover** (One separate page)
  - The Proposal Title

- Author(s)' Name(s) and/or Team #
- Course Project Professor's Name
- Course Name, Number, and Section #
- University Name
- Date

- **Introduction**
  - Briefly explain the project and its scope
    - Describe the problem that you are trying to solve, and explain why it should be interesting (to the audience), why it is challenging (technically)
- **Related Work**
  - Discuss some of the previous attempts (by others in the field) to the problem you are trying to solve.
    - It should highlight the best solutions in the field, why they work, where they fall short, and how these approaches relate to your proposed solution to the problem.
    - Briefly describe what's **difference** of your proposed solution(s) from other's work.
- **Objectives**
  - Briefly describe/list the objectives of the overall project you proposed
- **Selected Dataset**
  - Briefly describe the overall selected dataset
  - Briefly describe the features/columns you would like to select for your data analysis
- **Proposed system**
  - Briefly describe and draw a conceptual system architecture/high-level framework you proposed
  - *Briefly describe potential NLP and/data analytics approaches you will explore for your solution to the problem.
  - Explain how your solutions are **different** from what has already been done?
  - May include data visualization and other data analysis methods (e.g., descriptive statistics, inferential statistics, machine learning, etc.).
- **Proposed development platforms**
  - Briefly describe the software and hardware development platforms
- **References**
- *Other contents could be added if necessary*
  - *show a piece of data as an appendix*

3. check Point 2: Project Progress Report (*submitted in **one Word file***)

- **Baseline solution**
  - Implement and evaluate a baseline solution to your problem on your data.
  - Describe your baseline solution and present its results.
  - Present an analysis and discussion of the errors of the baseline solution.
  - What errors should be addressed by your (next iteration of the) system?
- **Proposed solutions and present preliminary results**

- Given the errors of the baseline solution, implement your proposed solutions and present preliminary results.
- Analyze the preliminary results of your solution and comment on whether or not it addressed the shortcomings of the baseline solution.
  - If it did not, what modifications should you make to the solution in order to improve the results?

- **Extend your write-up for Proposal with the preliminary results**
  - Analyze the results
  - Explain for how your proposed solution succeeded (or not) to address the previously identified errors.
  - What errors remain open and what should be done about those?

4. **Check Point 3: Presentation** (*submitted in one **PPT file**)*

   - **Make Power Point slides**
   - **In-Class Presentation**
     - Presentation of the final report
     - Demonstrations of the working system

5. **Check point 3: Final Technical Report,(submitted in one Word file)Deliverable 3: Working system** (*submitted in python file)*
6. *Final Technical Report:*

   - **Cover** (One separate page)
7. **Abstract** (200- 300 words) and **Keywords**
8. **Introduction**
9. **Related Work**
10. **Objectives**
11. **The Dataset(s)**
12. **The System**
    - **Architecture/framework**
    - **NLP and/or Data Analytics Approaches**
    - **SW/HW Development platforms**
13. **Experimental Results and Analysis**
    - Explore and present analysis of the dataset using relevant approaches
    - Prepare relevant analysis and visualizations for selected data items
    - Analyze and interpret the results
    - Discuss errors and open questions
    - Explain for how your proposed solution succeeded (or not)
14. **Conclusions**
    - Draw conclusions for overall project
    - Lessons learned from this project
    - Future work
15. **References**
    - Provide appropriate citations and references
    - Be sure to include a citation and link(s) for the dataset(s)
    - see http://infoguides.gmu.edu/citingdata
16. *Other contents could be added if necessary*

6. A Working System (*submitted in one zipped file)*

- **Introductory Comments (README)**
  Write a **README.txt** file. It should
  - clearly introduce to problem, point by point outline of solution, and examples of actual program input and output;
  - concisely describe how to set up and run your system, dataset link(s), and any other info you need to tell others to re-compile and re-run your system.
- **Detailed Comments (in source code)**
  - The detailed comments should be included throughout body of code, where authors are clearly identified. The comments should be clear and easy to follow and understand.
- **Project Functionality**
  - The project should implement **all** required functionality, runs on **all** evaluation data and achieves a score **greater than** the task baseline(s).
- **Code References/Acknowledgements**
  - If you code comes from others, please pay attribute to them.
- **Submissions**
  - All the source code, datasets, README, and related files including running output HTML files should be submitted in **one zipped file: AIT626_Team#_Initials_sys.zip**.
  - If the original dataset is too big, please provide the link(s) for downloading in the README file.
  - If the zipped file is too big to submit into Blackboard, please contact Instructor.
- NOTE: the system will be re-compiled and re-run on instructor's computer to double check for grading.