NLP Project

The NLP project will span five weeks, with a deliverable each week. The goal will be to implement and test an NLP system to accomplish a stated task. You will be working in pairs, unless you request to work independently, and as such will be submitting a weekly work report explaining how the tasks were divided up throughout the week. Below is a breakdown of the weekly deliverables:

Week 5 - Project proposal:

A one page proposal detailing your desired project. It will include:

- A description of the system you intend to build
- The purpose of the system (what you intend to do with it)
- The type of source material you intend to gather
- A "vision statement" describing what a successful implementation of your system would look like, including an explanation of input format and and example of correct output

Week 6 - Background and Prior Work:

A 2-3 page paper describing how similar systems have been implemented in the past. This should include:

- Techniques used for similar tasks in the past (Logistic Regression, State Vector Machines, Naive Bayes, etc) with a brief explanation of why they are used (i.e. what purpose do they serve?)
- Available tools and frameworks that are used for the task (NLTK, CoreNLP, etc)
  - Be specific about the modules in the frameworks and what they would do
- Examples of previous systems (at least 3) that accomplished similar tasks and how they were built
  - Give a brief summary of what the goal of the system was
  - Include the tools and techniques used
- Summarize what your system will do (what is the task you will be accomplishing and what information are you expecting to get from your system), what tools and techniques you plan on using, and how what your system will do is different from the examples you described.

Week 7 - Training Data Collection and Processing:

A 1-2 page paper describing the data you will be using for your project. It will include:

- How the data was/is being captured
  - Where is it coming from?
  - Who are the authors of the data?
  - How did you collect the data?
  - What is the raw format of the data?
- How will the data be formatted to work with your system?
  - How will it be read in?
  - Any normalization or pruning?
  - What annotations will you be adding to the training data? (sentiment, POS tags, meaning, etc)
  - What will a trained model look like?
    - What information will it contain?
    - What features of the data will be used for decision-making?

Week 8 - Prototype and Cross-Validation Testing:

A basic implementation of your system. The prototype should have been run through at least 5-fold cross-validation (10-fold would be better, if possible) to test the resulting model before submission. The prototype will consist of:

- The system, compiled as a self-contained executable.
  - Any dependencies, libraries, etc, must be included in the executable
  - The user (me) should not have to build, download, or configure anything.
  - You may assume I have the necessary language SDKs installed for interpreted languages (Java 18, Python 3, etc),
  - The model generated by your training data, saved as a separate binary file that can be loaded by your system
  - In other words, to use your system, I won't have to wait for the model to be built.
- The set of training data that was used to build the model.
- A set of test data
- Clear instructions on how to use the system. This will include:
  - How the user will feed the test data into the system
  - How the output will be reported to the user
  - How to interpret the output
- A report detailing the output of the cross-validation tests, including the Precision, Recall, and F-measure of all folds and any actions taken as a result of these tests. Was any overfitting identified? Were any features under-represented? How close is the current performance to your vision statement's desired performance?

Week 9 - Draft Project Submission:

A draft of the final paper for your project. It will include:

- A one paragraph summary of your project
- The three previous papers you submitted (background, data, and performance), formatted as sections of the paper
- An explanation of how your system was implemented.
  - Tools and techniques used
    - Include algorithms and even pseudocode detailing how your system functions
    - Discuss why these tools and techniques were chosen
- A discussion on the final performance of the system
  - Include the results of running your test data
  - Discuss how this performance compares to your vision statement's desired performance. How close is it? What, if anything, is missing?
- Discuss what the next step would be in implementing the system
  - What improvements could be made? How could they be implemented?
- Discuss lessons learned. In a more free-form section, explain what you learned from the project. What did you expect at the start of the project? Were there any surprises? What did you learn from the process?

Week 10 - Final Project Submission:

The completed project. It will include:

- The system as an executable, as described in week 8
- All the training and test data used.
- Full, formal, user documentation of your system
- The final version of your paper.
- A set of PowerPoint slides that will be used for your final presentation

Week 11 - Final Presentation:

The final exam time will be spent with each group giving a short (8-10 minute) presentation and demonstration of their project. The presentation will be targeted to non-technical people. You will need to describe the techniques you used without relying on the audience understanding the algorithms, jargon, and math involved. Furthermore, the demonstration should be something a non-technical person should be able to execute on their own (no computer expertise required, though you may assume they can type commands into a terminal window, with good documentation and examples).