CSCI 404 Final Project Information

Start on May 10, 2021. Due on June 9 (Wed), 2021 11:59pm on Canvas.

Project Topics

Your first task is to pick a project topic. Pick an NLP application that you are passionate about or something you find interesting. One important aspect of designing your project is to identify one or several well-defined dataset that you can use for your project. If that data needs considerable pre-processing to suit your task, or if you intend to collect the needed data yourself, keep in mind that this is only one part of the expected project work, but can often take considerable time. We still expect a solid implementation and discussion of results, so pace your project accordingly.

If you are still looking for some idea, below is a list of suggested project ideas. You are also welcome to drop by my office hours to talk about this.

Many of the items below represent a topic rather than a specific project. Thus, many projects or variations of a single project are possible for each topic. The general idea is that the project will allow you to get a hands-on experience and allow you to learn/investigate a topic in far more detail than what would be possible from class discussions alone. Please note that the criteria for success in the project is not limited to an effective system (accurate, efficient etc.). For the purposes of this course, I care more about why you considered the investigated problem to be important, how you addressed the main issues/challenges, what you learnt etc. More detail on the project assessment can be found in the latter part of this document.

The final project can be a group project. Indeed, by working as a group, you can attempt something larger and more interesting. However, the amount of work should be appropriately scaled to the size of the group, and you should include a brief statement on the responsibilities of different members of the team. Team members will normally get the same grade, but I reserve the right to differentiate in egregious cases. In general we would like group size of 2; if you are considering a bigger group, you need to talk to me. Solo projects are, of course, allowed.

You are free (and, where appropriate, encouraged) to make use of existing code and systems as part of your project, but you should make sure their use is properly acknowledged, and make clear what additional value your project is adding.

Project Ideas

Here is a sample of ideas and projects that have been used in similar NLP courses and in the previous quarters. You should feel free to come up with your own. Use these as a helpful idea generator.

- **Text Summarization**: Summarization is the task of producing a shorter version of one or several documents that preserves most of the input's meaning. This could potentially help people with reading disabilities.
    - Summarize Restaurant Reviews: take a list of reviews about a restaurant, and generate a single English summary for that restaurant. Use Yelp.com or some other website for the data.
    - Similar projects could be summarizing movie reviews, Amazon reviews, etc.
    - Others summmarization tasks could be scientific journals summarization, movie/game transcripts summarization, etc.
- **Text Classification**
    - Text Complexity Classificaition: a sample project report:
        - classifying texts based on their difficulty levels. This task's outcomes have the potential of enhancing education immensely. Texts of proper level of complexity can then be recommended to readers according to their reading ability, ultimately resulting in a more personalized educational experience. Generating or transforming text into simpler levels of complexity encourages more widespread knowledge, approachable from different fields and backgrounds.
    - Toxic Comments Classification. Similar tasks: Cyberbullying detection in social media, abusive/offennsive language detection in social media.
    - Fake News Challenge github
    - Spam / Click-bait detection
    - Topic identification: Multi-label classification of printed media articles to topics. Dataset: MultiLabel Classification - Reuters News Dataset
    - Author Idenfication
- **Song Generator**: Use a corpus of actual song lyrics and automatically generate new songs, perhaps given an initial sentence from the user. Make it rhyme. Similar ideas could be poem generator, genre-based story generator, TV show script generator, news article generator, President inauguration speech generation, celebrity graduation generation, user-based tweet generator etc.
    - Obama Speeches? For instance, you can create a bot which writes some new speeches in Obama's style
    - Narendra Modi bot saying "doston"? Start by scrapping off his Hindi speeches from his personal website
    - Example Dataset: English Transcript of Modi speeches
- **Information Extraction**: Choose a type of text format that typically contains useful information, but in written language. For instance, classifieds on craigslist. Build a system that extracts relevant information for products being offered for sale, such as the price, make, model, etc. Your system will read the sentences and extract the key pieces of information automatically. Similar projects could be information extraction from housing advertisement, customer reviews, clinical dictations, literatures, or any unstructured text of your interest.
    - Related task: Keyword/Concept identification :
        - Identify keywords from millions of questions. Dataset: StackOverflow question samples by Facebook

  - Good use cases of Information Extraction/Keyword extraction
    - Keyword Highlighting Improves Comprehension for People with Dyslexia
    - Evaluating the Benefit of Highlighting Key Words in Captions for People who are Deaf or Hard of Hearing

- **Text Simplification**: Text Simplification is the task of reducing the complexity of the vocabulary and sentence structure of text while retaining its original meaning, with the goal of improving readability and understanding. Simplification has a variety of important societal applications, for example increasing accessibility for those with cognitive disabilities such as aphasia, dyslexia, and autism, or for non-native speakers and children with reading difficulties.
  - A System for the Simplification of Numerical Expressions at Different Levels of Understandability in this workshop
  - Other related tasks:
    - Quality Assessment for Text Simplification (QATS)
    - Wiki sentence simplification: a sample project from Stanford Deep NLP class
  - You may also combine the text simplification and summarization tasks. That is to simplify the generated summary.
  - Some resources
- **NLP and AI fairness**:
  - NLP Bias Against People with Disabilities (a medium article)
    - There exist undesirable biases against people with disabilities exist in NLP tasks and models, specifically toxicity prediction, sentiment analysis, and word embeddings.
- **Tweet clustering and topic modeling**: apply clustering algorithms to tweets and do topic extraction from each cluster. Similar projects could be do clustering and topic modeling to the text in differnt domains.
  - For example, find out how covid-19 pandemic has been affecting people with disabilities. You would need to scrape tweets (with a certain time window, and with possible hashtags such as `#accessibility`, `#disabled`, `#DisabiityDivide`, `#a11y`, `#pwd`, etc.
- **Sentiment Analysis**: Choose a domain of interest and apply sentiment analysis to it. This is a special case of text classification.
  - Twitter sentiment analysis
  - sentiment analysis on tweets related to a certain topic. For example, find out how covid-19 pandemic has been affecting people with disabilities. You would need to scrape tweets (with a certain time window, and with possible hashtags such as `#accessibility`, `#disabled`, `#DisabiityDivide`, `#a11y`, `#pwd`, etc.
- **Complex Word Identification**: extend your homeworkd code to produce an **innovative** idea (you cannot just use a basic machine learning method any more) for this task. You must build the system on the dataset of this shared task.
- **Sentence to Sentence semantic similarity**: Can you identify question pairs that have the same intent or meaning? Dataset: Quora question pairs with similar questions marked.
  - similar tasks: measure similarity between sentences and their paraphrases. measure similarity between sentences and their simplified versions. Plagiarism detection. Datasets, PAWS.
- **Chatbot**: Build a system that can have a conversation with you. The user types messages, and your system replies based on the user's text. Many approaches here ... you could use a large twitter corpus and do language similarity.
  - References on more advanced implementations: Chat-bot architecure based on Neural Machine Translation. Dataset: Reddit Dataset. mlm/blog.
- **Automatic Image Caption Generation**: Automatic image captioning is the task where, given a photograph, the system must generate a caption that describes the contents of the image. (Warning: *This project will require knowledge such as image encoding, encoder-decoder architecure/deep neural network, which is out of the scope of our NLP class*.)

Unlabeled Data for Clustering, Language Models, etc.

- Web data
- BootCat

A list of **shared tasks** (including the dataset) is provided below.

- Sebastian Ruder’s curated collection
- Kaggle NLP tasks

SemEval Shared Tasks

- SemEval 2020
- SemEval 2019
- SemEval 2018
- SemEval 2017
- SemEval 2016
- you may work on a shared task in even earlier years.

NOTE: the honor concept applies. For the projects with code, you can adopt their topics, but not take their actual implementation ideas off-the-shelf and simply reproduce their system. You must make it your own.

Important Dates:

- 5/14 (**Friday noon 12:00pm**): Project Proposal Due.

  - Must run by me before you finalize your project topic.
- 5/24 Project Progress Report Due.
- 6/1-6/4 Final Project Presentations
- 6/9 Project report/code due

Grading:

- Project Proposal (5%)
- Progress Report (5%)
- Final Presentation + demo (optional) (8%)
- Final Report + code (15%)

Project Proposal

One- or two- page document that addresses the following issues:

1. What is the problem or task you propose to solve?
2. What is interesting about this problem from an NLP perspective?
3. What technical method or approach will you use?
4. On what data will you run your system?
5. How will you evaluate the performance of your system?
6. What NLP-related difficulties and challenges do you anticipate?

Note:

1. Must get approval by me before you finalize your project topic.
2. Must cite and briefly describe at least two pieces of relevant existing work.

Project Progress Report

Report the progress/achievement of your project so far (fit on 1-2 pages). On my judgement 25-30% of the work should get done.

Key information. Your PDF should have the following information in :

1. Title: The title of your project (you can change this later).
2. Team member names: List the names and @wwu.edu email addresses of all of your team members.
3. About the progress of the project:
    - Have you set up the data? How have you set up the data? links to the data and description of the data.
    - What baseline model have you set up? Or are you planning to set up a baseline? Describe the baseline in enough detail.
    - What's the core technique are you using? Did you find out any other techniques that can be potentitally used? Describe the baseline in enough detail.
    - Will your model be implemented from scratch? Or which part of your model will be implemented from scratch? Describe the baseline in enough detail.
    - Is your model based on an existing codebase? If yes, give the link(s) of the codebase.
    - Plan of work for the next two weeks. Describe the baseline in enough detail.
    - Justify why do you think you've gotten 25-30% done. Describe the baseline in enough detail.
    - Roles: if this is a teamwork, who did which part. Describe the baseline in enough detail.

In the report, you need to demonstrate and communicate well that you have obtained a better understanding of the problem/tasks/methods/metrics/research context.

Final Project Presentation: (10 minutes including questions)

- Problem description, motivation, societal/ethical implications of the task
- Proposed solution
- Related work knowledge (the task, the method)
- Description of implementation: data set (training data, test data, development set if any), feature set, toolkit, experimental setup
- Discussions of your experimental results
- Critical analysis of your work (potential improvement and future work)

The above is the basic points you have to cover in your presentation. Please feel free to cover more, and do some demo if helps.

All the teams are required to participate in and observe all the presentations. Participation in peer-grading will be a part of the assessment of your final presentations. This is the peer-grading sheet. Please make a copy of your own and submit it on Canvas by the end of each presentation day.

Final Submission

Your final report should be between 4-5 pages (including references) using the **following template**:

- Style sheets (Latex, Word) are available here: http://acl2020.org/downloads/acl2020-templates.zip
- And the Overleaf template is also available here: https://www.overleaf.com/latex/templates/acl-2020-proceedings-template/zsrkcwjptpcd

You can use either latex or word document templates.

There are **three things** you need to turn in:

- Final Project Writeup: A PDF file of your final report. Submitted to Canvas.
- Final Project Supplementary Materials: A zip file of your supplementary materials. You are required to include all the code for your project in the supplementary materials. Submitted to here.
- Final Project workload split-up and/or any additional information. Submit it as a comment along with the report on Canvas.
- Your team only needs to submit each of these things once, but make sure that the submission tags all members of the team.
- ***A quick reminder: Have you submitted your final presentation slides to this folder and submited your peer assessment sheet on Canvas? Please do so if not yet.***

Report:

The following is a suggested structure for the report:

- Title, Authors
- Abstract: It should not be more than 300 words.
- Introduction
  - Problem description
  - What are the societal or ethical implications of your application? What categories of end-users this application may benefit?
  - Your overall approach to the problem.
- Related Work: Relevant literature for your project.
- Approach: This section details your approach to the problem. You should be specific. This section is optional if you would like to include it in the next section.
- Experiments: In this section, you describe:
  - The dataset(s) you used (links to the data) and data pre-processing.
  - How you ran your experiments
  - The evaluation metric(s) you used
  - Others should be abel to replicate your experiment through reading your description/explanation here.
- Results and Discussion:
  - It's very important to show both quantitative evaluation (show numbers, figures, tables etc. relating to your evaluation metric(s)) and qualitative evaluation (show example results, etc.).
- Conclusion and Future Work (if you had more time, what would you do next?)
- References: Include references to all literature that informed your project work. This is absolutely necessary.

You can of course have other sections.

Please make sure you address these issues

- You need to adhere to the format described above
- You need to provide data description
- You need to describe the toolkit and data pre-processing you did (if any)
- You need to describe the feature set (if any)
- You need to briefly describe the theory behind algorithm you used
- You need to present at least 1 table of results
- Your need to discuss the results and provide critical analysis of your work (Such as: Have you thought about ways to improve the model? How do you think the model accuracy? etc.)

Your final paper should be of sufficient length. Two pages is not sufficient.

Supplementary Material

Examples of things to include in your supplementary material:

- Source code (required)
- Cool videos, interactive visualizations, demos, etc. (optional)

Examples of things not to include in your supplementary material:

- The source code for an entire submodule (e.g. nltk/SpaCy submodules, CoreNLP)
- Any code that is larger than 1MB
- Model checkpoints
- A computer virus

Honor Code

You may use any existing code, libraries, etc. and consult and any papers, books, online references, etc. for your project. However, you must cite your sources in your writeup and clearly indicate which parts of the project are your contribution and which parts were implemented by others. Under no circumstances may you look at another CSCI 404 group's code or incorporate their code into your project.

If you are doing a similar project for another class, you must make this clear and write down the exact portion of the project that is being counted for CSCI 404.