CS 6320.002: Natural Language Processing
Fall 2019

Project Milestone 2 – 10 Points
Due 8:30am 11 Nov. 2019

Deliverables: A PDF writeup. You can submit just one per team. Please put the names of ALL members of your team on your writeup.

0 Planning Ahead

After Milestone 2, you have the following milestones:

- Milestone 3, due 1.5 weeks after Milestone 2. Implement a baseline from an existing paper and evaluate it on your test data.
- Milestone 4, due 2.5 weeks after Milestone 3. Implement your improvements over the baseline and evaluate. Complete your final project paper.
- Presentation, also 2.5 weeks after Milestone 3.

Each milestone is worth 8% of your final grade; the proposal and presentation are worth 4% each.

You don't have to (and shouldn't!) wait for the milestone due dates to work on each stage of your project! Keep working steadily so that you aren't rushing at the end of the semester. There are no more homeworks in this class, so set aside the time you would have spent each week doing homework to work on your project instead.

1 Instructions

Answer the following questions about the data you are using for the project. Each question can be answered in 1-2 sentences.

Note that every project must have a test set, but not every project necessarily has training or validation sets (ie. if you are using an unsupervised approach or don't have hyperparameters to tune). If you are not using a training or validation set, skip the last question.

- What does the data look like? What is the input – document? sentence? word? And what is the gold standard output – class label? word embedding? real number?

- Who collected the data? Was it you, or are you using someone else's dataset? If the latter, give the citation for the dataset.

- Where did the inputs come from? For example, the documents in the New York Times annotated corpus for summarization comes from archived NYT articles.

- Where did the gold standard labels or annotations come from? For example, the NYT annotated corpus's gold standard summaries were written by the humans of the NYT Indexing Service for archival purposes.

- How many gold standard labels or annotations are there per document? If there are multiple labels or annotations per document, what is the interannotator agreement?

- How large is the dataset? How many input/output pairs?

- What is the train/validate/test split? How many input/output pairs are in each set, and is there a standard split for this dataset? For example, the New York Times dataset has a standard split of 90%/5%/5% based on the dates that the articles were published.