Project Proposal

What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.

Kefan Chen, kchen61@illinois.edu, (captain)
Lijuan Geng, lijuang2@illinois.edu

What system have you chosen? Which subtopic(s) under the system?

System Extension -> MeTA Toolkit -> Extend the MeTA Toolkit to support text analysis

Briefly describe any datasets, algorithms or techniques you plan to use

EM algorithm to interpret pre-determined topics.
PLSA algorithm and its extensions, for example, adopting different priors.

If you are adding a function, how will you demonstrate that it works as expected? If you are improving a function, how will you show your implementation actually works better?

We are planning to add new functions/extensions to the existing Toolkit. We will use the following two ways to determine it works as expected.

- By empirical judgement, since lots of the NLP tasks require human input, so we will manually judge if the output makes sense
- By comparing the result to other Toolkit or framework that we can find, which implements a similar algorithm, or tries to achieve the same goal

How will your code communicate with or utilize the system? It is also fine to build your own systems, just please state your plan clearly

Our code will try to integrate closely and utilize as much as the current MeTA Toolkit can provide as possible. We do not plan to build our own systems, since everything provided by MeTA Toolkit would be super useful for our project and our goal is to make MeTA Toolkit support more use cases through our extensions.

Which programming language do you plan to use?

Python for the application layer.

Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.

- Spend more time to understand the current MeTA package setup, from a developer perspective instead of a pure user perspective. (4 hours)
- Determine the actual goals that we will want to achieve and how they should be integrated with the existing Toolkit. (4 hours)
- Implement the extensions based on earlier determined ideas. (22 hours)
- Test and validate our results against other available tools or procedures and further improvements. (6 hours)
- Extra documentation, refactor, examples and code clean up. (4 hours)

This track could contain multiple small milestones, goals and features that we can implement and achieve. We will determine the exact and appropriate scope as the investigation and project goes on depending on our progress.

Since we are a team of 2 people and we would like each of us to get involved as much as possible, so not everything is parallelizable. For example, each of us will need to understand the current package, investigate the integration points and perform needed tests and validations.