COMPSCI-546
Learning to Rank project report
Jaideep Rao

- **Description of the system, design tradeoffs, questions you had and how you resolved them, etc. List the software libraries you used, and for what purpose.**
  - The project is structured across 4 different components. Trec_data.py is a program contains functions to load provided data into memory, extract_features.py contains feature extractors that are run against loaded data, ranking_test.py creates train/test folds out of the feature set and runs a random-forest classifier that outputs a mean AUC score and feature contribution for that run, and a bash script that runs the ranked jar to perform random forest classification and output ndcg@5 scores for each train/test fold
  - For this project I added 3 new feature extractors into the extract_features.py program. I did not use any additional libraries

- **Describe each feature you chose to implement. What was your motivation for selection of that feature**
  - *Title mentions*: This feature calculates the ratio of how many titles contain an entity within a document to how many titles occur within a given document. This includes titles of the article itself, as well as any other titles occurring within the article such as those for specific contents or images. I chose this feature because I believe mentions of an entity within any kind of title occurring within a document should be considered to be of significance because there is emphasis being placed on that entity enough for it to be a part of the title that is supposed to be representative of the content it encapsulates. So there is a good chance that an entity being mentioned in a title could mean this document or item within this document is relevant to this entity
  - *Caption mentions*: This feature calculates the ratio of how many image captions contain an entity within a document to how many image captions occur within that document. I chose this feature because image captions are direct textual representations of content captured in images present in a document. An entity being mentioned in an image could be important because it is referring to that entity in a very specific way in terms of the image containing some information related to the entity. Since image captions tend to be concise, if a caption mentions an entity there is a high

chance that the image is closely related to that entity and by extension the document may be as well

- *Anchor text mentions*: For the paragraphs that link to other pages, this feature calculates the ratio of how many anchor texts mention an entity to how many anchor texts exist within a document. I thought this might be a useful feature because anchor texts are a great way to summarize what the linked documents are centered around. An entity being mentioned in the text that leads to a page that is centered around that entity, there could be a good chance that it was linked to from this document in the context of this entity in some way, which could make this document more relevant to this entity

- **Explain how your chosen features behave when used on the TREC Entity dataset when using normalized discounted cumulative gain (NDCG@5). Show performance of all of the features in combination, as well as the impact of each individual feature**.
    - Using my added features as well as combining all the available features, these are the average ndcg@5 scores observed across all train/test folds:
    - Title_mentions feature:
        - Mean-AUC: 0.514
        - title-fraction 1.0
        - Avg ndcg@5 score: 0.24505
    - Caption_mentions feature:
        - Mean-AUC: 0.501
        - caption-fraction 1.0
        - Avg ndcg@5 score: 0.25952
    - Anchor_text_mentions feature:
        - Mean-AUC: 0.504
        - anchor-text-fraction 1.0
        - Avg ndcg@5 score: 0.19826
    - Combination of all features:
        - Mean-AUC: 0.465
        - anchor-text-fraction 0.1692109320603809
        - caption-fraction 0.08358552417818842
        - para-fraction 0.7155305239831082
        - title-fraction 0.031673019778322445
        - Avg ndcg@5 score: 0.24186

As we can see, that the avg ndcg@5 score is greatest with caption-mentions feature, suggesting that this feature could potentially be quite informative in a larger dataset. We see that anchor-text mentions outputs the lowest score. This might be because of the sparse nature of the feature values since anchor texts are fewer and far between. The poor performance obtained from the combination of all features could possibly be attributed to the model being fed potentially misrepresentative information about the data from the other features that may be sparse and inaccurate to some extent