# Project Proposal: Machine Learning with Large Datasets

Team: Nitish Kulkarni (10805), Prakruthi Prabhakar (10805)

October 2, 2017

## 1 Proposal

Question Answering systems solve the problem of extracting answers to a wide variety of questions pertaining to any domain. Research in this area has traditionally looked at the problem from two perspectives:

1. A reading comprehension task, where given a document and an associated question, the answer is extracted in the context of the document.
2. A collection of knowledge sources and documents are provided and the answers are extracted using the knowledge corpus.

The challenges of this problem include syntactic and semantic understanding of the posed question along with efficient parsing of documents and knowledge bases.

Some of the recent research focus has been on employing answer extraction using multiple large-scale sources of data and knowledge bases [5, 6].

In this project, drawing ideas from the aforementioned research, we aim to work on the problem:

> *Given a large collection of documents, knowledge sources and associated questions, extract relevant answers to the questions from the corpus.*

## 2 Methodology

We will begin by establishing simple baselines for answer extraction using machine comprehension of text:

- sliding window approach as well as distance based extension [8]
- logistic regression model using lexicalized features or dependency tree path features [7]

We will then explore scalable implementations of alternative NLP and deep learning techniques to enhance the model performance . We also aim to enhance the approach using *open-domain question answering* techniques by augmenting large-scale knowledge sources to extract relevant answers to the posed questions. [5, 6, 8]

## 3 Data

We are planning to use the following datasets:

- **SQuAD Dataset** The SimpleQuestions [7] dataset consists of a total of 108,442 questions written in natural language by human English-speaking annotators each paired with a corresponding fact, formatted as (subject, relationship, object), that provides the answer but also a complete explanation. Facts have been extracted from the Knowledge Base Freebase.

- **WikiMovies Dataset** The WikiMovies dataset [4] contains question-answer pairs in the movies domain.
- **CMU Question-Answer Dataset** This dataset [9] provides a corpus of Wikipedia articles along with manually-generated factoid questions and answers.
- **MCTest Dataset** [2] This dataset comprises of 660 annotated reading comprehension stories split into two categories, where each story set is a story and its associated questions.
- **Freebase Dataset** We will use Freebase knowledge base [1], which provides structured information in the form of (subject, predicate, object) triples. FB2M, which was used in (Bordes et al., 2014a), contains about 2M entities and 5k relationships. FB5M, is much larger with about 5M entities and more than 7.5k relationships.
- **Reverb Dataset** The Reverb dataset [3] contains about 2M entities and 600k relationships in an unstructured format in comparison to Freebase knowledge base.

# 4 Team

We are currently a team of two 10-805 students. We are looking for another team member.

# References


[1] Freebase data. `https://developers.google.com/freebase/data`.

[2] Mctest data. `https://github.com/mcobzarenco/mctest`.

[3] Reverb data. `http://reverb.cs.washington.edu/`.

[4] Wikimovies data. `https://research.fb.com/downloads/babi/`.

[5] Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*, 2015.

[6] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. *arXiv preprint arXiv:1606.03126*, 2016.

[7] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250, 2016.

[8] Matthew Richardson, Christopher JC Burges, and Erin Renshaw. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*, volume 3, page 4, 2013.

[9] Noah A Smith, Michael Heilman, and Rebecca Hwa. Question generation as a competitive undergraduate course project. In *Proceedings of the NSF Workshop on the Question Generation Shared Task and Evaluation Challenge*, 2008.