Requirement Document

Intended Users:

The final consumers of our system output will be Question Answering systems which intend to operate within a particular domain and have a set of initial questions that define the domain (topic). Our system will be a module that can be used with any Question Answering system like Watson, to extract data from the web, generate a relevant corpus, and a Knowledge Graph that will be given to the QA system to use for answering questions later.

Functions:

1. **Read questions given by Watson API** - Our system reads and performs preprocessing (Question analysis, discard irrelevant information in questions etc.) on the questions given by the QA system.
2. **Compile and construct the corpus of relevant documents** - For the given questions, the system uses our internal APIs to extract relevant data from the web by doing an iterative source expansion and Machine Reading. The system constructs a knowledge graph. The graph contains the pruned and refined corpus.
3. **Return the knowledge graph and corpus to Watson** - After compiling the high precision and recall corpus and meeting the threshold recall, the system returns the knowledge graph and RDF records of the same to QA system API, that can be used for Information extraction by the QA system later.

User Interaction:`

The user provides a set of questions and the correct [gold-standard] answers for those questions for training our system. The system then uses our internal logic and comes up with a corpus and knowledge graph constructed from it, that ensures a right combination of precision and recall, to make sure that the correct answer has a high confidence score.

Non-Functional Requirements:

Our system should be able to run fast, as it has to explore and extract relevant data from the data sources and the size of the data sources available is large [the whole web]. Also, since there is a dependency of the QA system on ours for further processing and data extraction, time is a really important factor here. The performance of the MR system should not vary too much with different sets of initial questions given to our system i.e. it should be consistent for different sets of initial questions, irrespective of the domain .

Since the system depends on external online sources for the initial set of documents and data, it should be able to handle errors in cases when the external data sources have faulted.

Resource Requirements:

The system, as of now, relies on the documents available on the web that can be retrieved by the Bing search Engine or DBPedia for a given query. So, there is a clear dependency on the search engine for the quality of documents and the input data to the system.

As of now, our corpus size for training on 62 questions is ~60 MB (for worst case, when we retrieve the top 12 documents for every query from the search engine), so there is certainly a huge requirement of memory. We will try and minimize it as much as possible, by ensuring that we keep only the relevant content from the knowledge graph in memory and have a good recall at the same time.

Hardware requirement - We would need to keep the knowledge graph and its content somewhere, since we are training the system offline. So this may need some hard drives or we can store the data on cloud. Working with the system and the implementation will need around 5-6 machines.

Scalability:

We can scale our system to a large set of questions wisely since our current design allows us to split the questions into multiple subsets, each of which can be handled by different machines and we can perform the Machine Reading process in parallel and then combine the corpora to generate a final combined corpus. Although, this seems straight forward, it may not actually be easy to parallelize the process naively, as our current model needs to look at the whole set of questions and corpus to determine the importance (weight) of the seed (document, Named entity, question etc.) in the iterative expansion.

Scope:

We limit our work to only the MR system and any work towards improving the quality of data and the quality of the resources provided by the current data sources is beyond the scope of this project.

We will be working on generalizing the system so that it can work for any domain though the initial development of the system will be on the terrorism dataset. As of now, we do not have a bound on memory and computation time. We will restrict ourselves to the data sources - Bing and DBPedia. We have used only the metric - Binary Answer Recall, so far to measure the accuracy of our system.