Q&A System
Alex Merryman
IEMS 308

Executive Summary

Much of the worldwide data generated on a daily basis is text data. It is therefore very important to be able to properly mine and analyze this data, which requires specialized techniques. These data contain valuable, actionable business insights; the Q&A system can serve as a knowledge search engine, or a sophisticated encyclopedia. Furthermore, it can be used as a due diligence tool to collect or verify information about the business world.

The high-level approach taken by the system is as follows:

1. Upon executing the program, the user is prompted to type a question into the command line.
2. The system parses the question to determine what sort of answer the user is looking for.
3. Keywords are extracted from the question.
4. The corpus is searched for documents that contain any of the keywords.
5. The relative significance of each document-word pairing is evaluated using the Okapi formula.
6. Documents scoring highest are returned as candidates.
7. Sentences matching the desired answer type are extracted from each document and scored by TF-IDF.
8. The best candidate sentence is returned as the answer.

While the system is not perfect, it can yield relevant results to each question.

Next Steps

- Extract the specific answer from the highest-ranked sentence. For example, instead of returning “Ford's record results in the third quarter show the strength of our One Ford plan around the world," said Alan Mulally, Ford president and CEO.” for the question “Who is the CEO of Ford?” the Q&A system could extract “Alan Mulally” from the sentence and present that as the answer.
- Optimize performance – despite the lightning-fast performance of Solr, the overall Q&A process takes far too long to be reasonable alternative to simple Google searching.
- Make the system more flexible; interpret questions more flexibly (rather than strict rules), optimize Okapi parameters, etc.