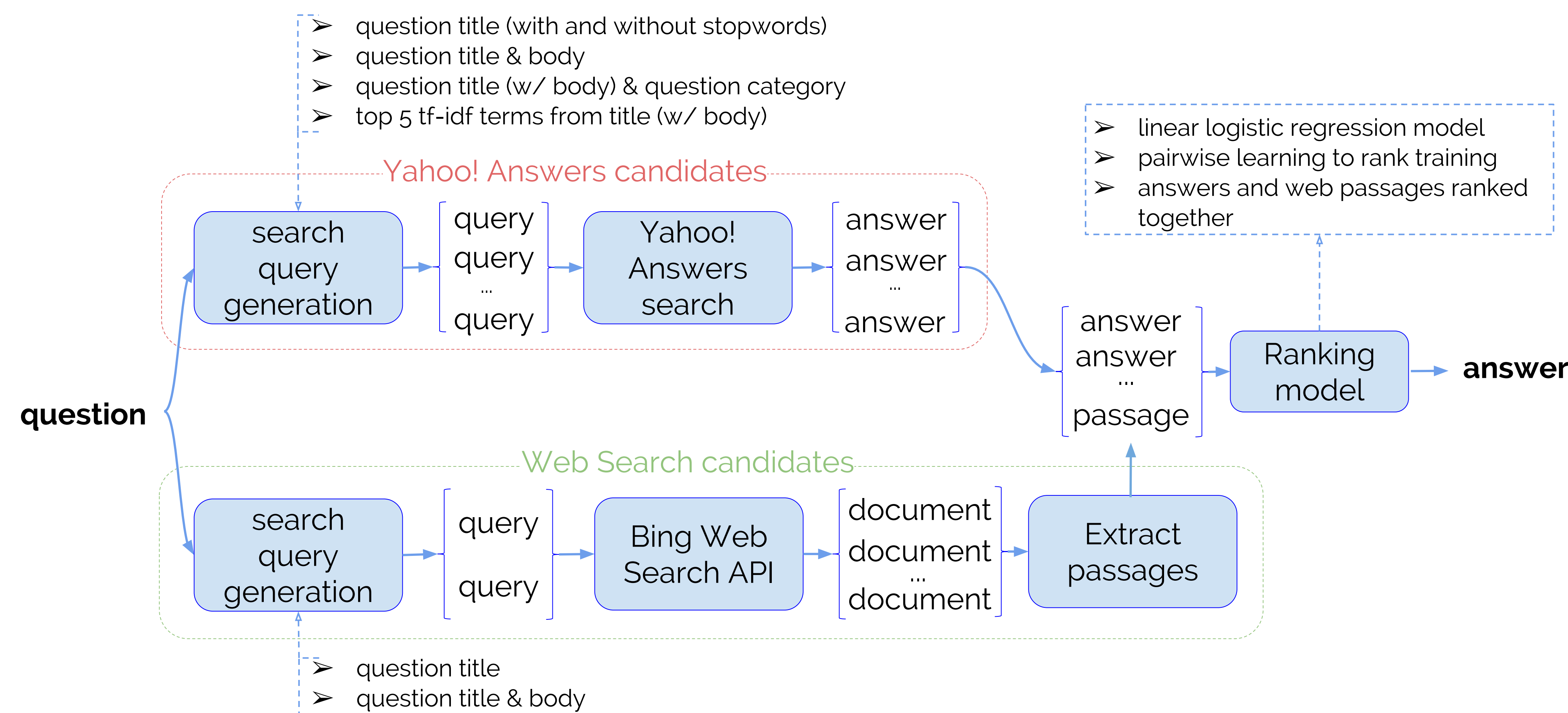




System overview

Our question answering system extracts candidate answers from passages of documents retrieved by a web search engine and answers to similar questions previously posted on Yahoo! Answers. All candidates are ranked using a single linear logistic regression model based on a set of term match, retrieval score, PMI and neural network score features and the candidate with the highest rank is returned as the answer.



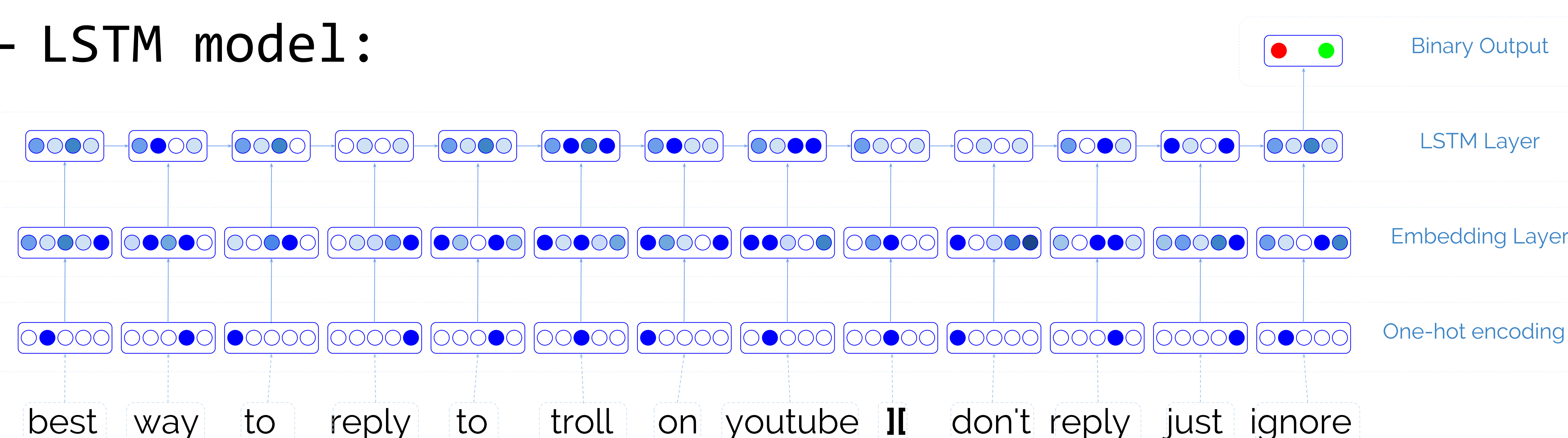
Candidate ranking features

- o answer statistics (length in chars, tokens, sentences, etc)
- o BM25 scores (using title or title + body as query)
- o term matches features (number of matches, maximum spans of matched terms, etc.)
- o question and retrieved question categories match
- o bag of word pairs from question and candidate answer
- o statistics of NPMI scores between question and answer terms
- o score computed by an LSTM neural network model

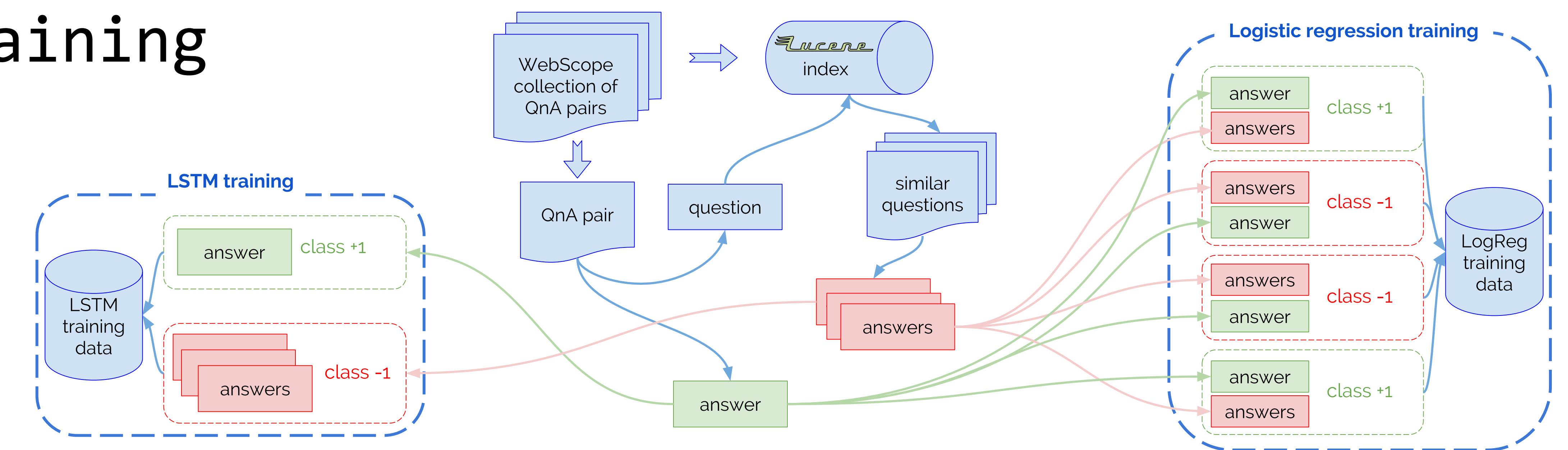
- Normalized PMI scores:

Term	NPMI score	Term	NPMI score	Term	NPMI score
visit	0.376	weight	0.655	diabetes	0.728
beach	0.373	calories	0.556	insulin	0.608
tourist	0.370	exercise	0.517	glucose	0.555
snorkeling	0.352	diet	0.506	niddm	0.551
museum	0.349	pounds	0.495	hyperglycemia	0.547

- LSTM model:

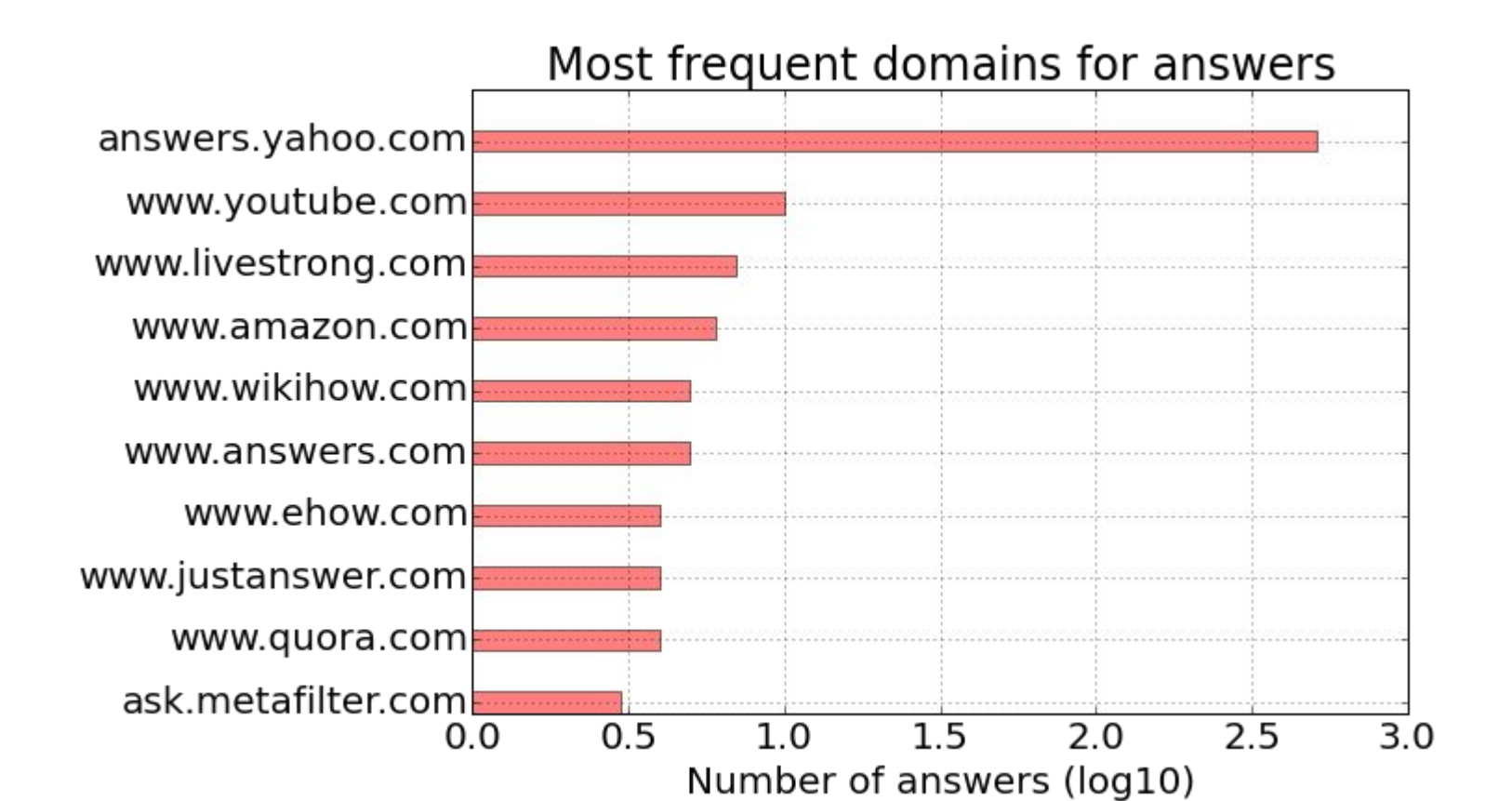
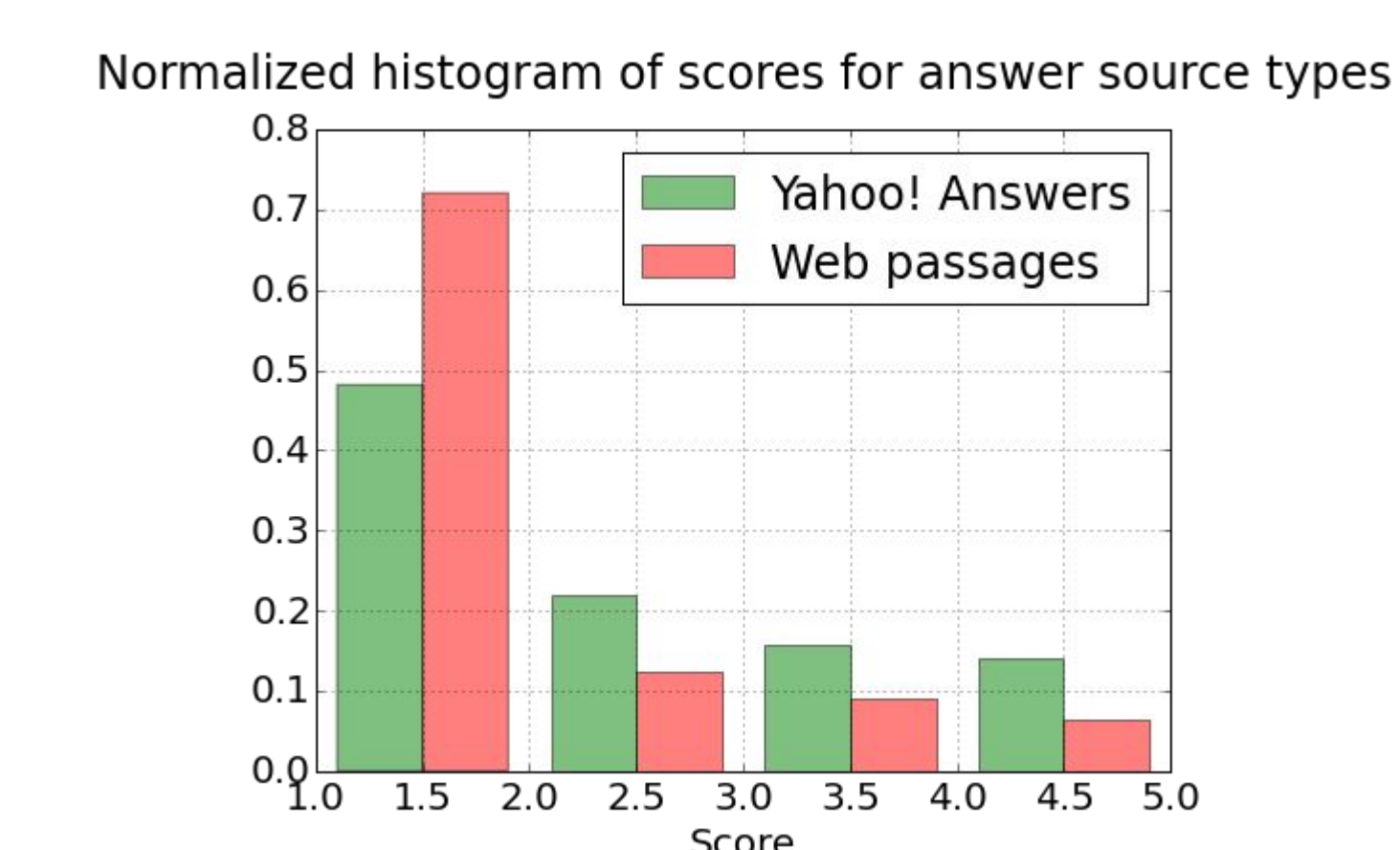
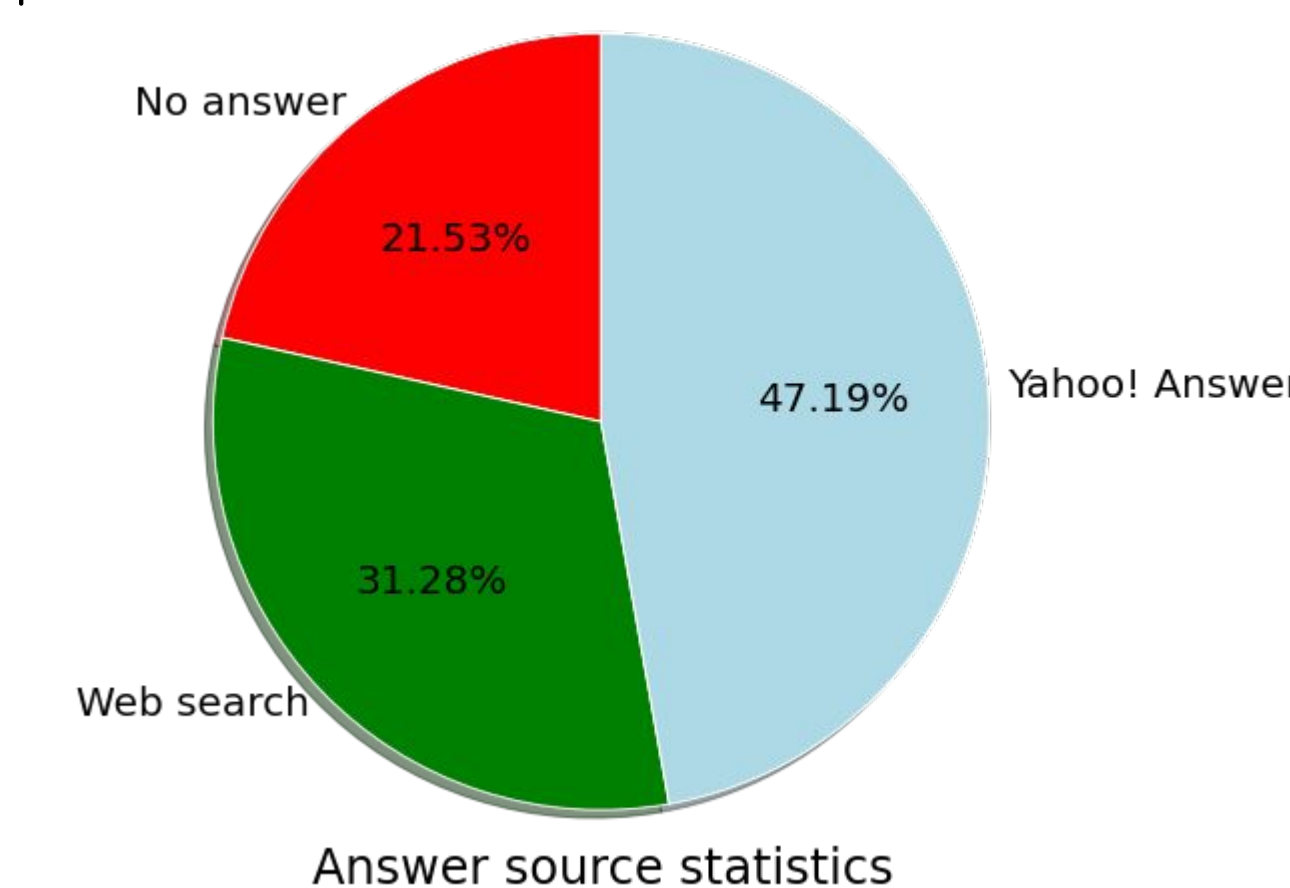


Training

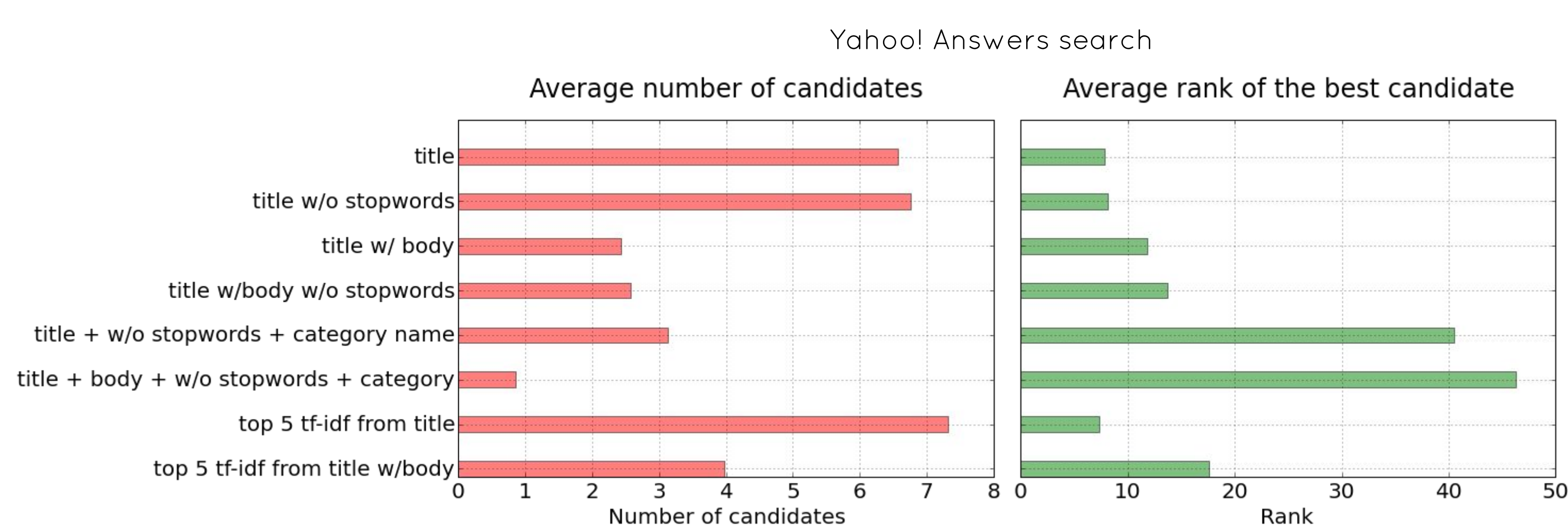


System analysis

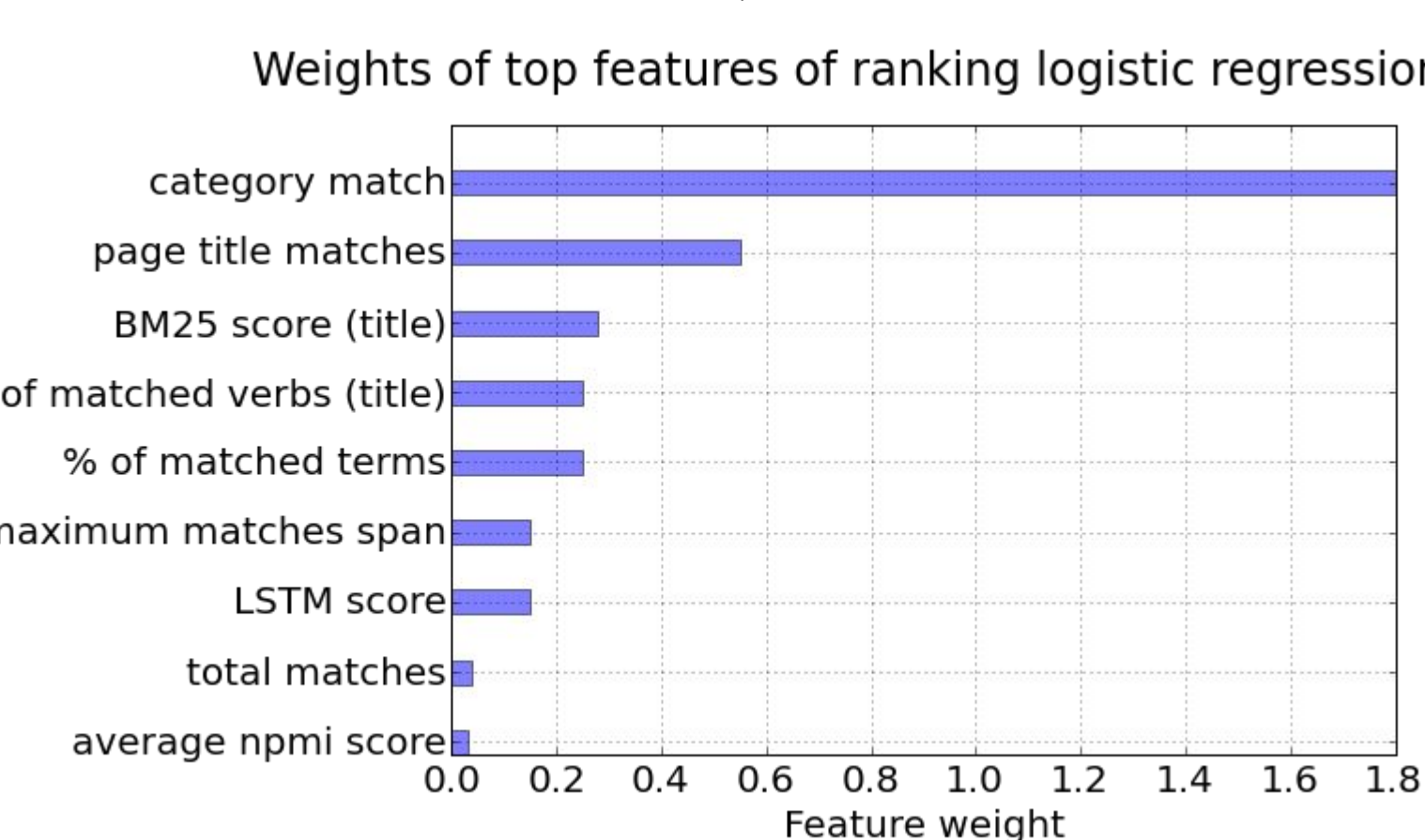
- ✓ Most of the responses were extracted from answers to previous Yahoo! Answers questions
- ✓ Average score of candidates from Yahoo! Answers is higher than of extracted web passages
- ✓ Many other answers extracted from web search results came from CQA websites



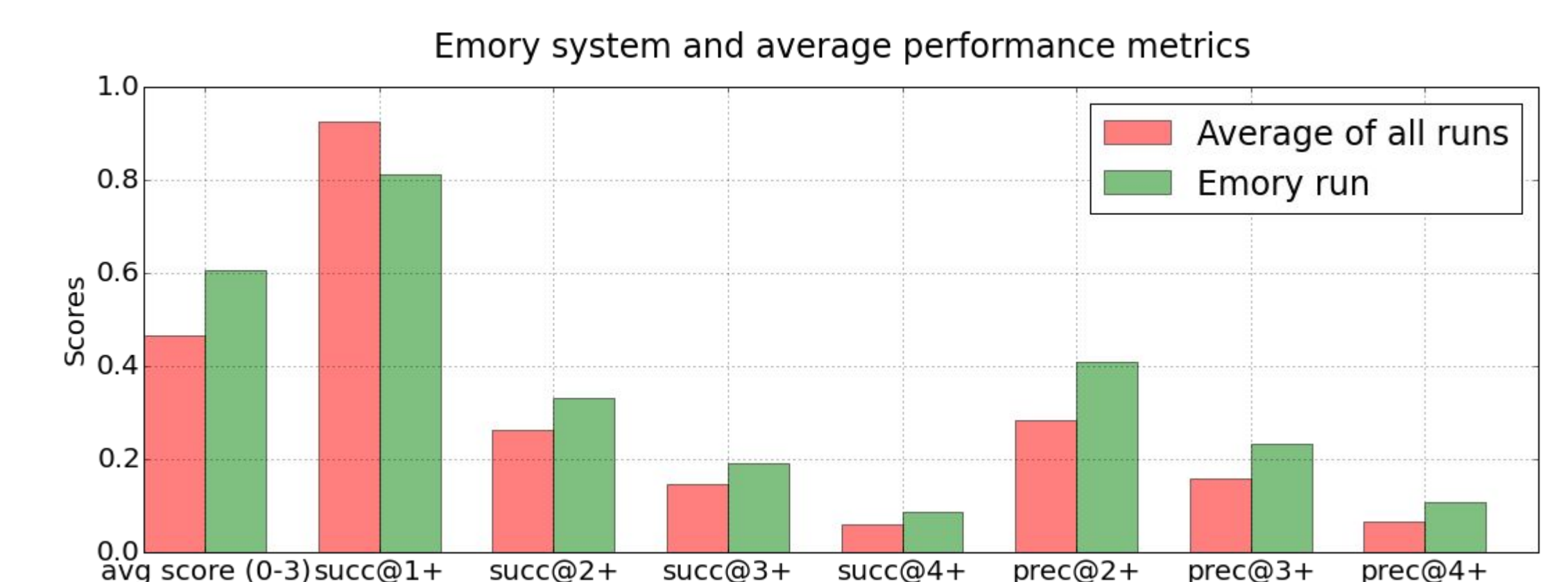
- ✓ Question title and top terms by tf-idf are more effective for generating search queries from questions



- ✓ Questions categories, terms match and BM25 score are among the top most useful features



- ✓ The system has higher than average non-response rate, but other metrics are better than average



Conclusions

- o answers to previously posted similar questions can be effectively reused to answer new questions
- o query generation strategies for CQA and web search are important as LiveQA questions are long
- o term matches-based features are the most useful in our model. It will most likely benefit from "better" text similarity features: translation models, n-grams, tree kernels, etc.
- o there is a huge room for improvement: only ~20% of questions were answered fully or partially