

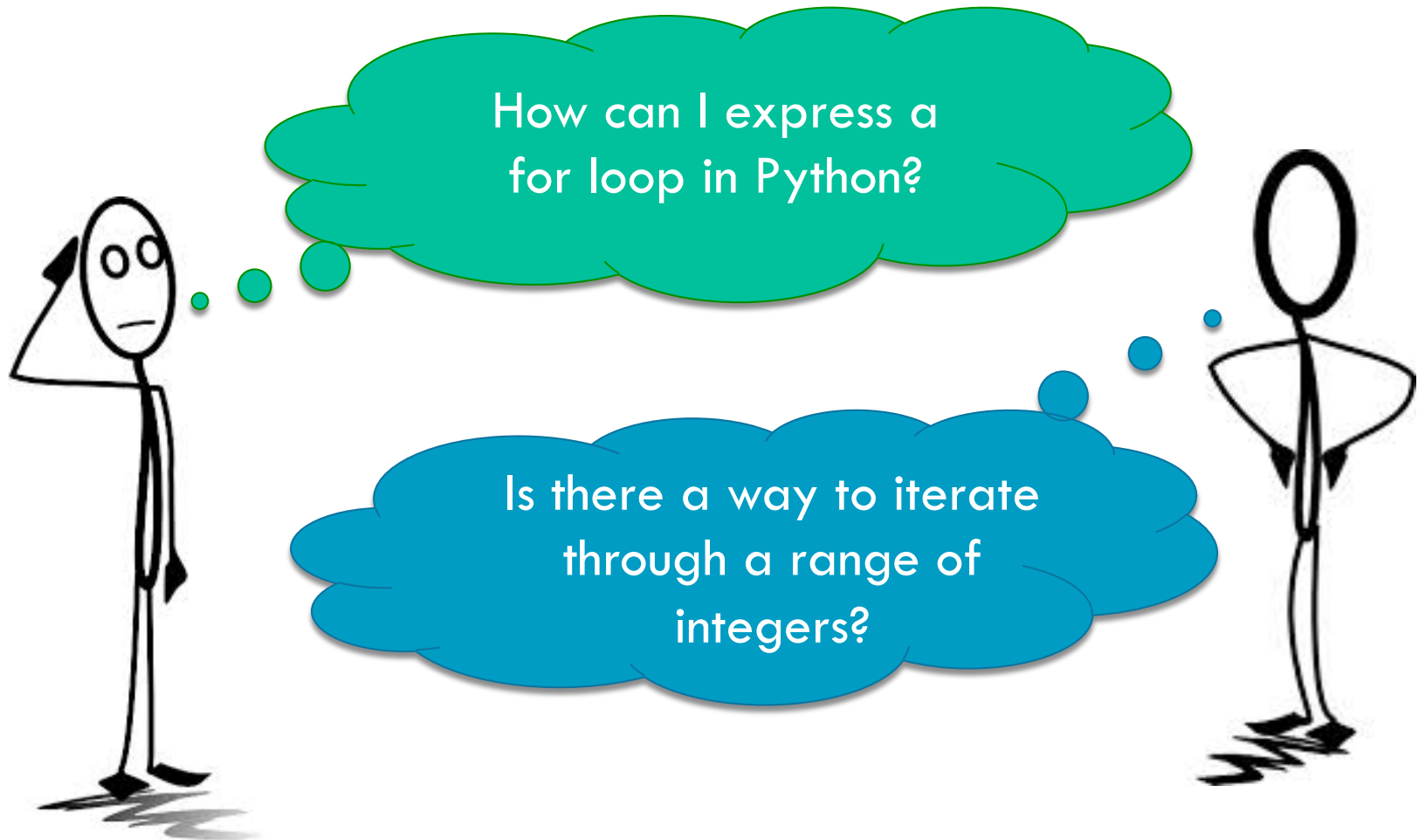
FINDING SIMILAR QUESTIONS FROM COMMUNITY-BASED Q&A SERVICES

22 April, 2014

Stephanie Valentine, Jung-in Koh, Shiqiang Guo

How to find similar questions?

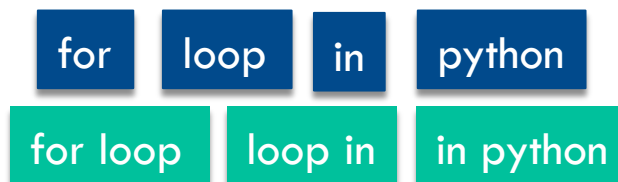
2



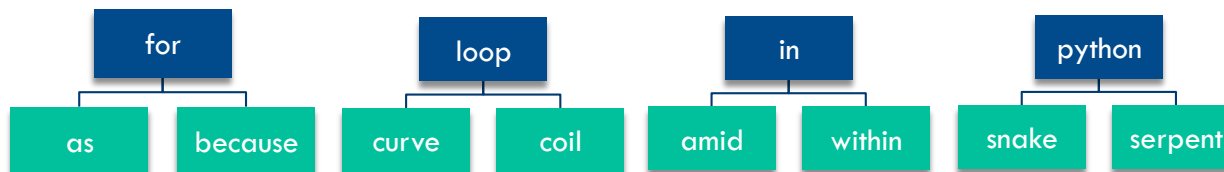
How to Build the Vector?

3

Add N-Grams (phrases of length N) [5]



Add Synonyms of Terms [6]



Weight Nouns & Verbs More Heavily [2]



TFIDF Weighting (Baseline)



Data Source and Metrics

4

- Gathered questions (title & body) from...



all posts with tag “Python”

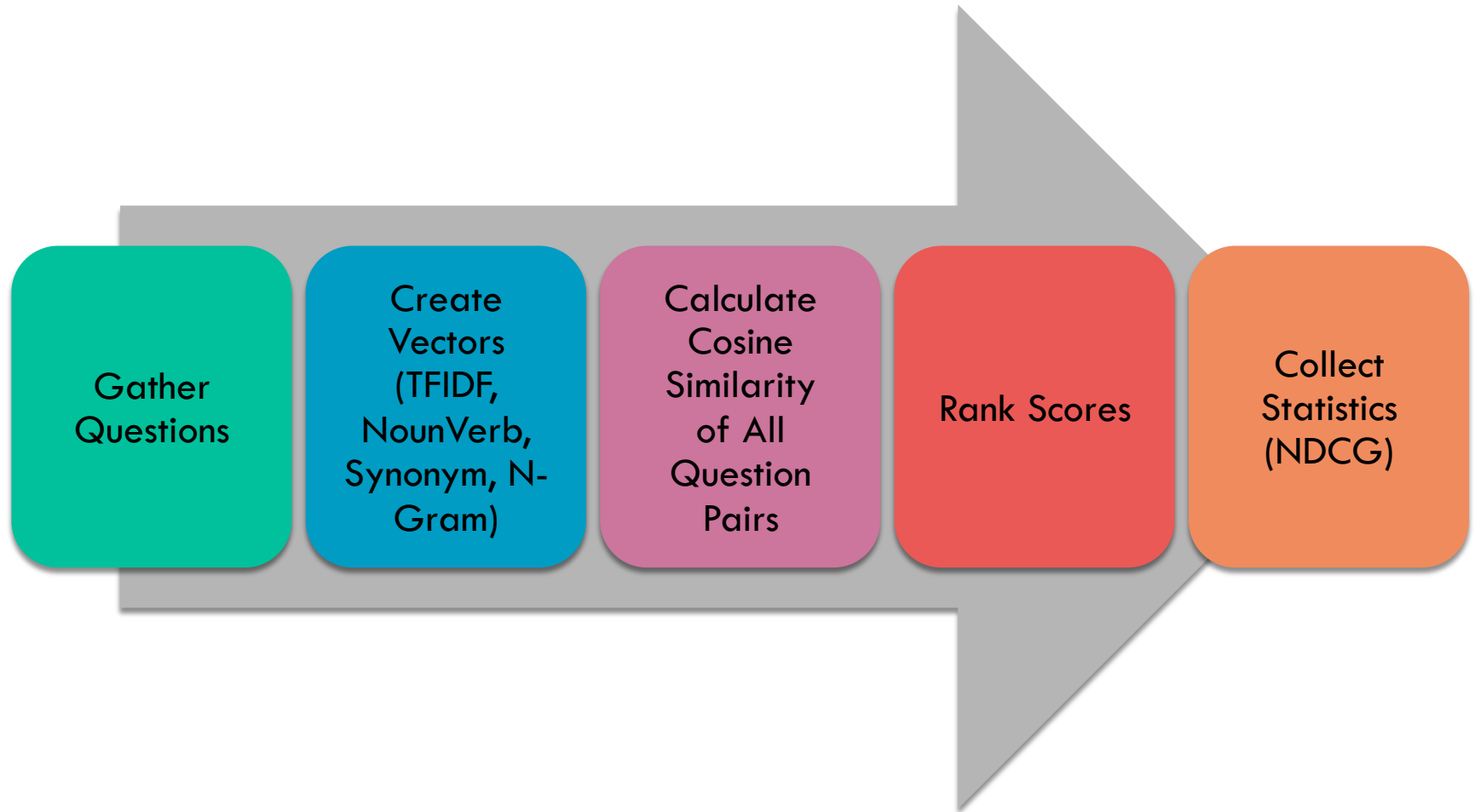


all posts

- Similarity determined with ‘links’
 - ▣ answer-er posts link to other similar question
- Relevance Scores:
 - $\left\{ \begin{array}{l} 1: \text{if link exists between doc1 and doc2 (known similarity)} \\ 0: \text{if no link exists (unknown similarity)} \end{array} \right.$

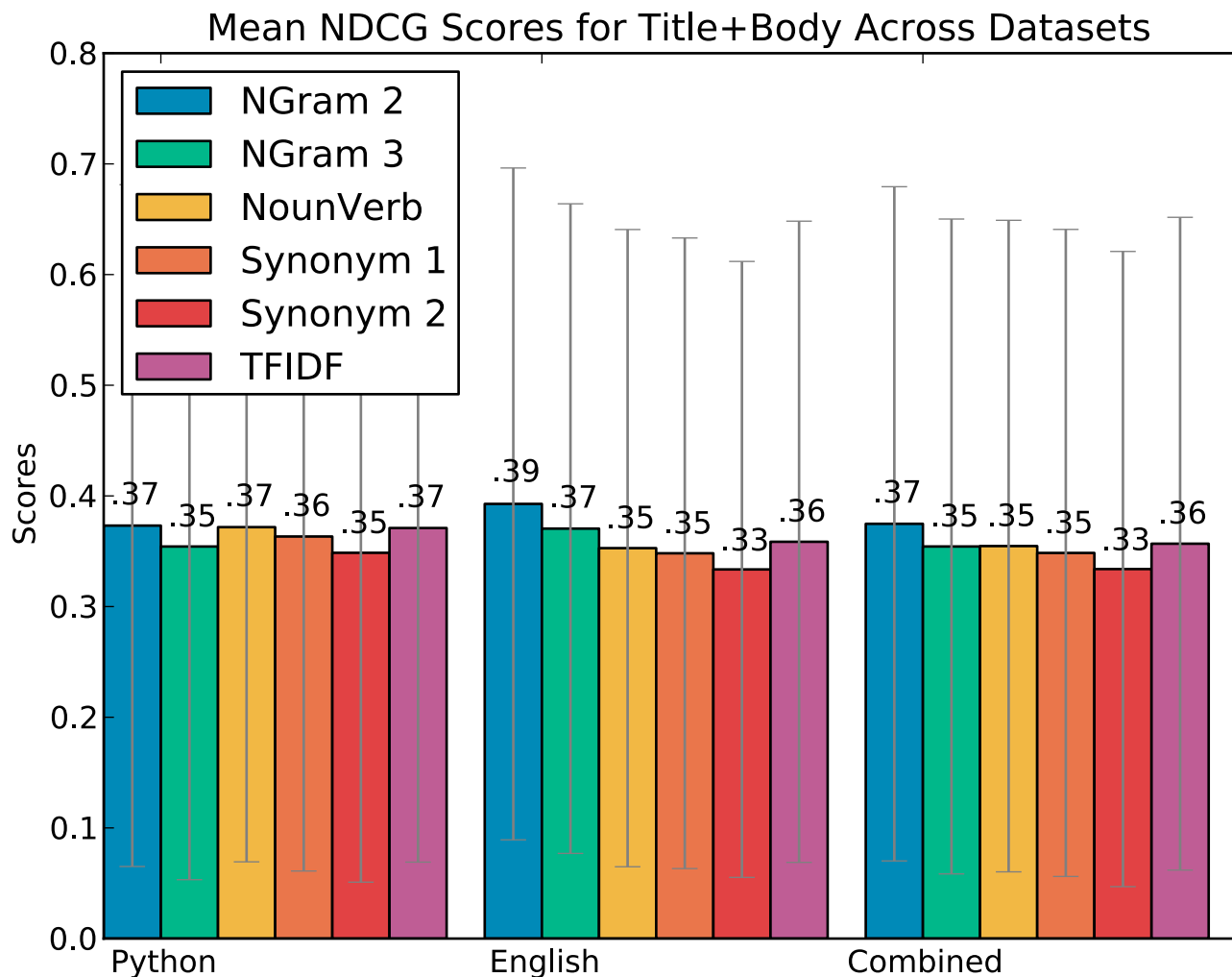
20-Second Summary of Experiments

5



Results and Evaluation

6



References

- [1] Jeon, Jiwoon, W. Bruce Croft, and Joon Ho Lee. "Finding similar questions in large question and answer archives." In Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 84-90. ACM, 2005.
- [2] Wang, Kai, Zhaoyan Ming, and Tat-Seng Chua. "A syntactic tree matching approach to finding similar questions in community-based qa services." In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 187-194. ACM, 2009.
- [3] Cao, X., Cong, G., Cui, B., Jensen, C.S., Zhang, C.: The use of categorization information in language models for question retrieval. In: CIKM 2009, pp. 265–274. ACM, New York (2009)
- [4] Jijkoun, V., de Rijke, M.: Retrieving answers from frequently asked questions pages on the web. In: CIKM 2005, pp. 76–83. ACM, New York (2005)
- [5] Barzilay, Regina, and Lillian Lee. "Learning to paraphrase: an unsupervised approach using multiple-sequence alignment." *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 2003.
- [6] Liu, Mingrong, Yicen Liu, and Qing Yang. "Searching semantically similar questions from a large community-based question archive." *Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on*. IEEE, 2009.