**Retrieval of Relevant Answers to Stack Overflow Queries by Ranking the Answers**

Akhil Kumar Baipaneni | Gayathri Chandrasekaran | Madhuri Yadav

# **Introduction:**

Stack Overflow is a question and answer site for professional and enthusiast programmers. It provides a platform for its users to post questions and answers. It also allows its users to upvote or downvote these questions and answers. Stack overflow has a wide range of questions and challenges faced by developers and their solutions in the computer engineering field. As per the Stack Overflow survey 2019, around 97% of developers visited the Stack Overflow website last year. And about 50 million people visit Stack Overflow to learn, share, and build their careers every month.

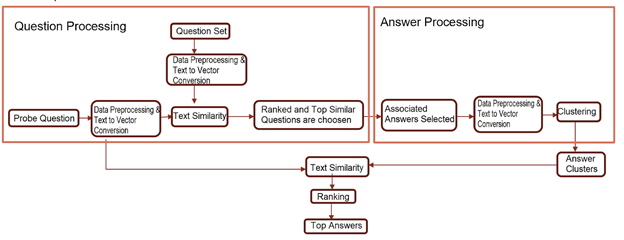
Although this forum offers numerous advantages for developers, on the flip side there are few drawbacks associated with it. Sometimes Stack overflow’s search engine does not retrieve any results even if a similar question exists in the repository. And as stack overflow allows duplicate questions and each question has a different set of answers many users find it difficult to get relevant answers to their queries. As the main source of revenue for stack overflow is through advertising these are the advantages to drive more traffic. However, the problem arises when users use more sophisticated search engines like google and these search engines redirect the user to other sites. Hence to overcome these gaps this project aims at improving Stack Overflow search using semantic similarity between questions & answers and retrieving the most relevant answers using a ranking methodology.

# **Dataset & Scope:**

The dataset "Python Questions from Stack Overflow" is used for this research which was published in Kaggle. The dataset contains a set of Python questions posted on the stack overflow website. A set of python questions is selected as Python is one of the fastest-growing programming languages and widely used by most of the developers. The dataset has three files Answers.csv, Questions.csv, and Tags.csv of which Answers.csv & Questions.csv are used in the research. These files contain questions and answers posted from August 2, 2008, to October 19, 2016.

Questions.csv contains 607,282 Questions whereas Answers.csv contains 987,122 answers in total. Each question may have more than one answer associated with it. Questions and answers are associated with each other using the "id" field. At the outset, this project focuses on WH-questions to limit the number of observations (i.e) 22,722 questions.

# **Experimental Setup:**

**Figure 1. Experimental setup and Data Flow description**

As shown in Figure 1 the project is divided into three parts:

1. Question Processing
2. Answer Processing
3. Q&A Relevance Mapping

## **Question Processing:**

In this phase, the probe question is compared with the stack overflow question corpus to find the most similar questions in the corpus. Initially, basic text preprocessing is performed on question title part and search question which includes lowercasing the tokens, converting accented characters, and removing stop words from tokens. However, stemming and lemmatization is not performed as the data is related to technical questions with different domain-specific meanings. Also, numeric data is preserved as removing numbers from questions changes the contextual meaning. On the preprocessed data various embedding techniques like word2vec, Bert, etc are performed to convert text to numbers. Embedded data is then compared using cosine similarity to get top similar questions for the search question. Each question is scored against distance value i.e. questions with less distance value are more similar to each other and vice versa. The questions with similar values are studied manually for various embedding techniques. It is observed that word2vec and Bert perform better than other embedding techniques for a given dataset. Hence, to improve the quality of the question similarity outcome results of both the embeddings are combined and questions with a similarity score greater than 0.90 are selected. Question processing helps gather answers spread across various duplicate questions for further analysis.

## **Answer Processing:**

In this phase, associated answers for selected questions are filtered and grouped to form various answer clusters. Similar to the question processing, basic text preprocessing was performed on the answer dataset. Apart from that, HTML tags related to code, image, URL were removed to extract the text information from the answer corpus. This processed data is converted to numbers using embedding models like Bert, tf-idf, word2vec, and fed to the hierarchical clustering model. Like question processing, data clusters are evaluated manually for each embedding technique. It is observed that Bert clusters perform better compared to other cluster data.

## **Question and Answer Mapping:**

Each answer cluster is compared to search questions using cosine similarity to check relevance between questions and answers. The similarity score is aggregated at a cluster level to rank the clusters. However, most of the answer part contains code chunks, leaving less information in the natural text to get actual relevance. To overcome this issue, the score features available in answer.csv. is used. The score value is the user annotated score for each answer to the question. Both similarity and user annotated scores are combined to get an overall score for each cluster. Clusters with maximum scores are ranked as top answer output.

Refer to the [Appendix](#_heading=h.3rdcrjn) for manual evaluated samples and results for each embedded model used for this project.

# **Challenges:**

Stack Overflow Q&A system is domain-specific, every stage requires additional preprocessing to preserve domain details associated with it. To overcome the issue various data preprocessing steps were revisited to have maximum details for the next stage in the process.

The answer set is a mixture of text, codes, images, and links. The answer body is dominated with code chunks than natural text making it difficult to conclude only with the similarity score of the model. Different strategies were used to get the most relevant answers and to overcome the issue.

# **Results:**

The word relevance of word2vec combined with the Bert sequence model to capture the semantic importance of the word in numeric form works great for the dataset in retrieving relevant questions for search questions. Grouping the most similar answers from answer corpus helps with relevance matching and eliminates redundant answers from the result set. Overall, the Bert model produces good results for retrieval of matched answers for given questions.

# **Future Scope:**

* Extend research for other question pairs apart from WH-questions.
* Build models that represent the entire Stack Overflow corpus having various programming domain-specific results.
* Build an interactive Q&A search engine and related API for wider use.

# **References:**

1. Stack Overflow Developer Survey Results from 2019,

Retrieved from:<https://insights.stackoverflow.com/survey/2019#overview>

1. Kaggle dataset - Python Questions from Stack Overflow

Retrieved from:<https://www.kaggle.com/stackoverflow/pythonquestions>

1. Google Cloud Big Query StackOverflow Public dataset

Retrieved from: [https://cloud.google.com/stack-overflow-q-a](https://cloud.google.com/blog/products/gcp/google-bigquery-public-datasets-now-include-stack-overflow-q-a)

1. Training a Ranking Function for Open-Domain Question Answering Research Paper

Retrieved from:<https://www.aclweb.org/anthology/N18-4017.pdf>

1. Comparison of different Word Embeddings on Text Similarity — A use case in NLP

Retrieved from: [https://medium.com/@Intellica.AI/](https://medium.com/@Intellica.AI/comparison-of-different-word-embeddings-on-text-similarity-a-use-case-in-nlp-e83e08469c1c)

1. Word embedding

Retrieved from: <https://medium.com/data-science-group-iitr/word-embedding-2d05d270b285>

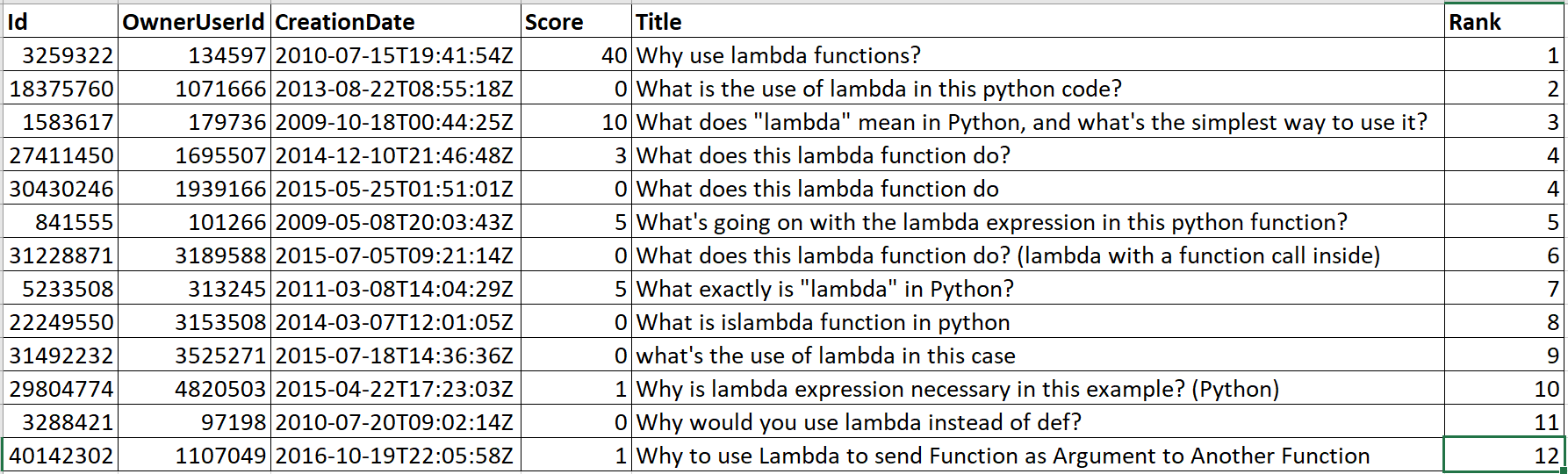
# **Appendix**

The appendix shows the various techniques used in the project and their result.

## **Model Evaluation Results:**

Search Question: What is the use of lambda function?

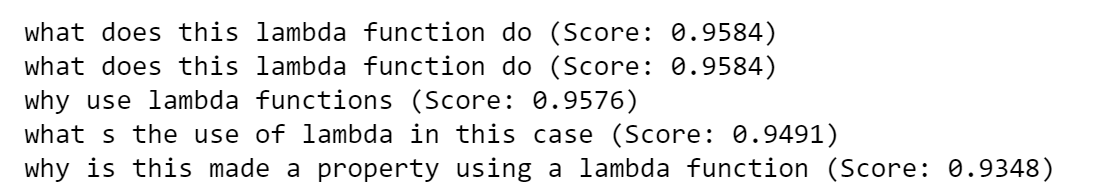
## **Question Processing:**



**Figure 2. Results of Manual Ranking from Question Corpus**

For a given question questions.csv is explored to get similar questions and each question is ranked manually to compare and check the performance of model result as shown in Figure 2.

### **Bert Embedding with Cosine Similarity Results:**



**Figure 3. Top 5 Similar results from Bert Embedding with Cosine Similarity (Duplicate questions treated as separate entities)**

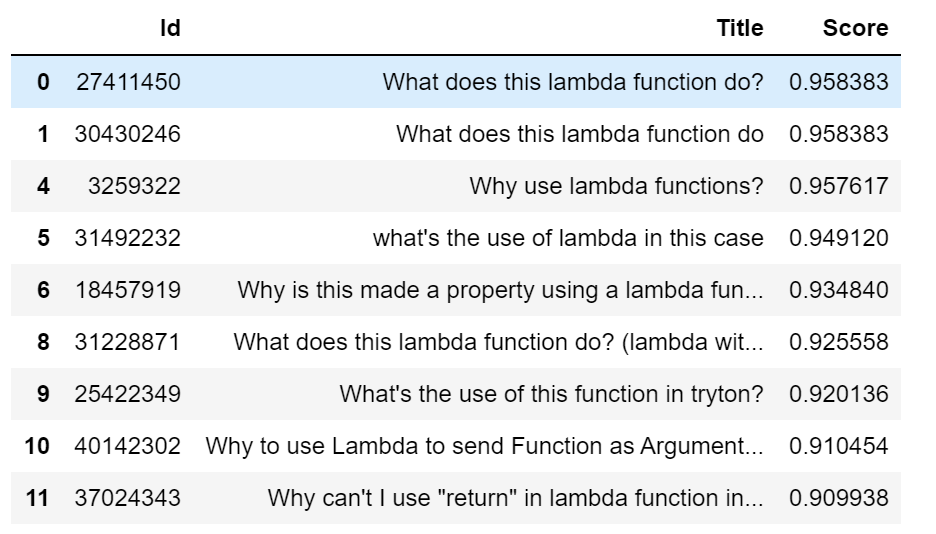
As shown in Figure 3 in the result obtained by Bert model, 4 out of 5 questions match with our manual ranking.

### **Word2vec Embedding with Cosine Similarity Results:**

**Figure 4. Top 5 Similar results from Word2vec Embedding with Cosine Similarity**

As shown in Figure 4 in the result obtained by the word2vec model, 3 out of 5 questions match with our manual ranking.

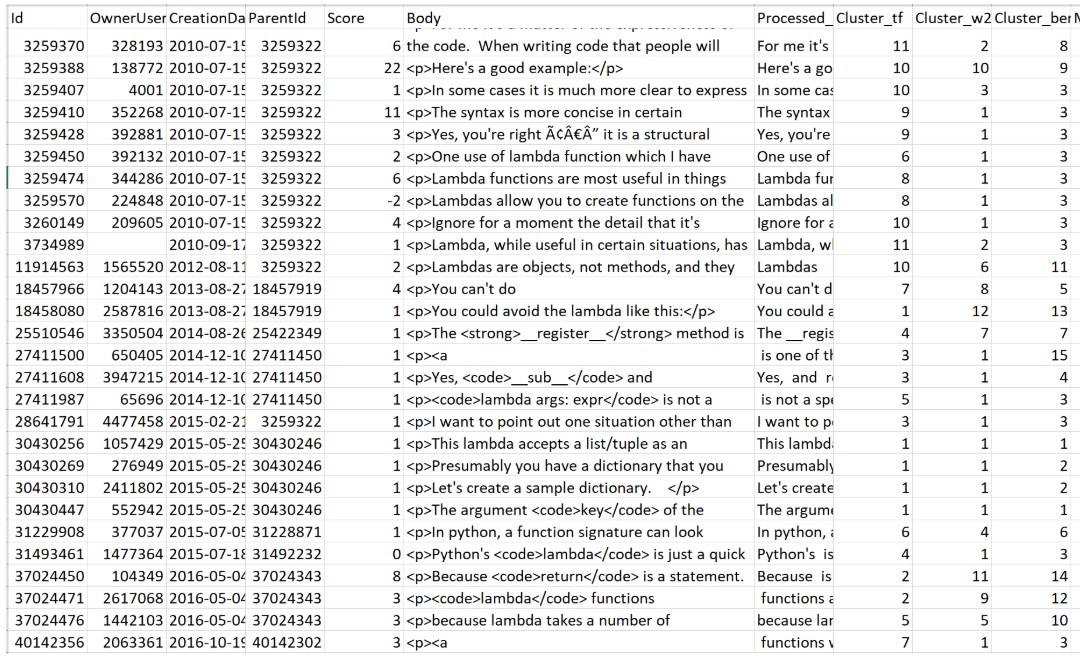
### **Bert+Word2vec Embedding with Cosine Similarity Results:**



**Figure 5. Top Similar Question Set Retrieved for the given search question**

As shown in Figure 5 in the result obtained by the Bert+Word2vec model, 4 out of 5 questions match with our manual ranking.

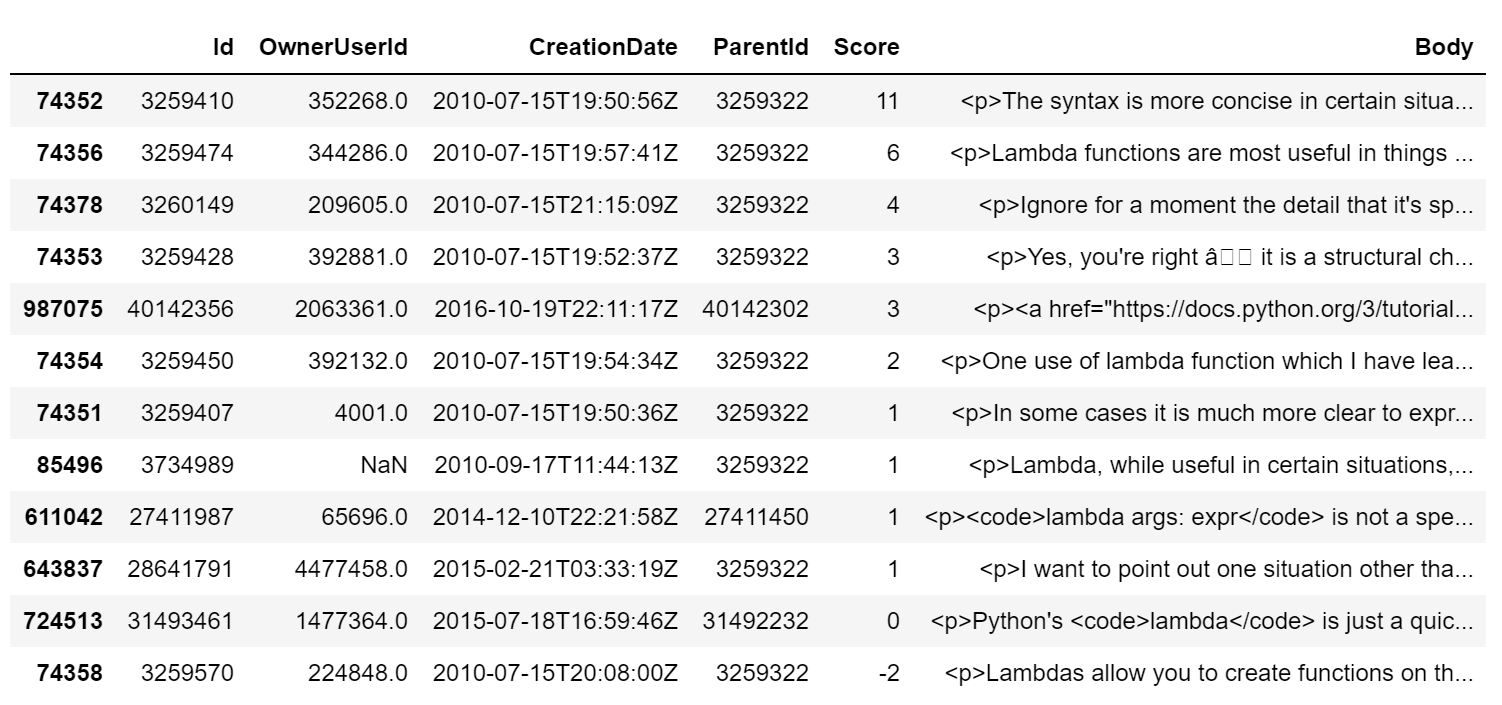
## **Answer Processing:**



**Figure 6. Answer Set Retrieved for given search question along with Hierarchical Cluster values**

Figure 6 shows hierarchical cluster results on answers set over various embedding models. After manual evaluation of each of the results, it is observed that Bert embedded results give good results with hierarchical clustering.

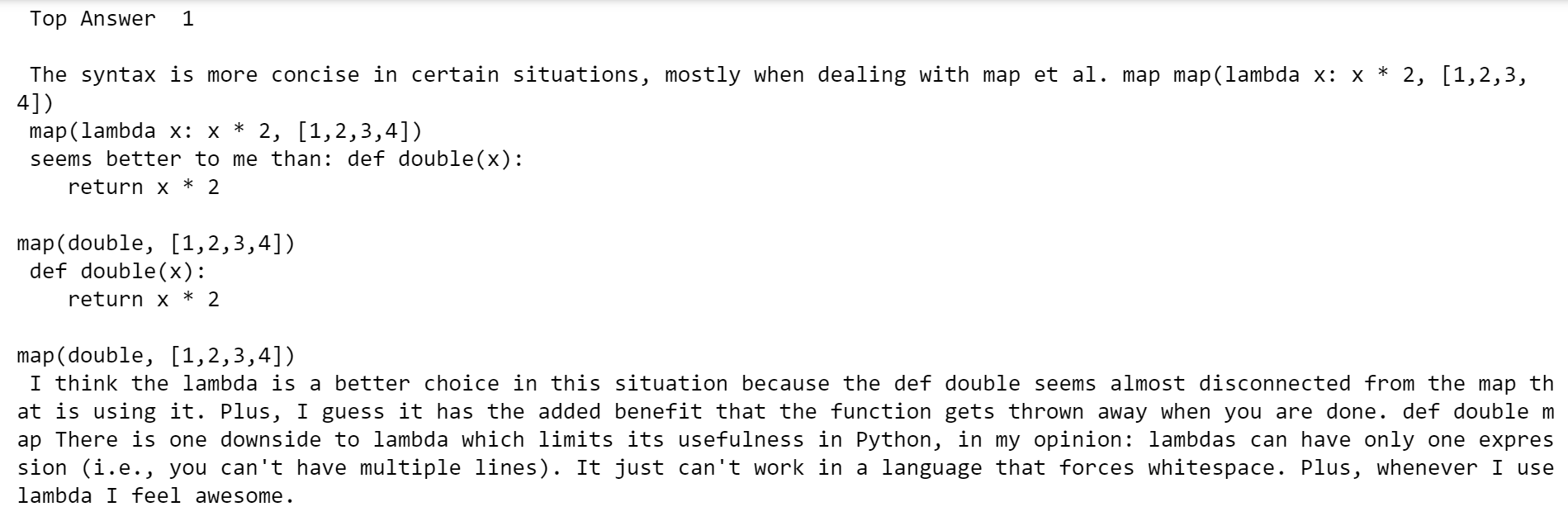
## **Question and Answer Relevance Mapping:**



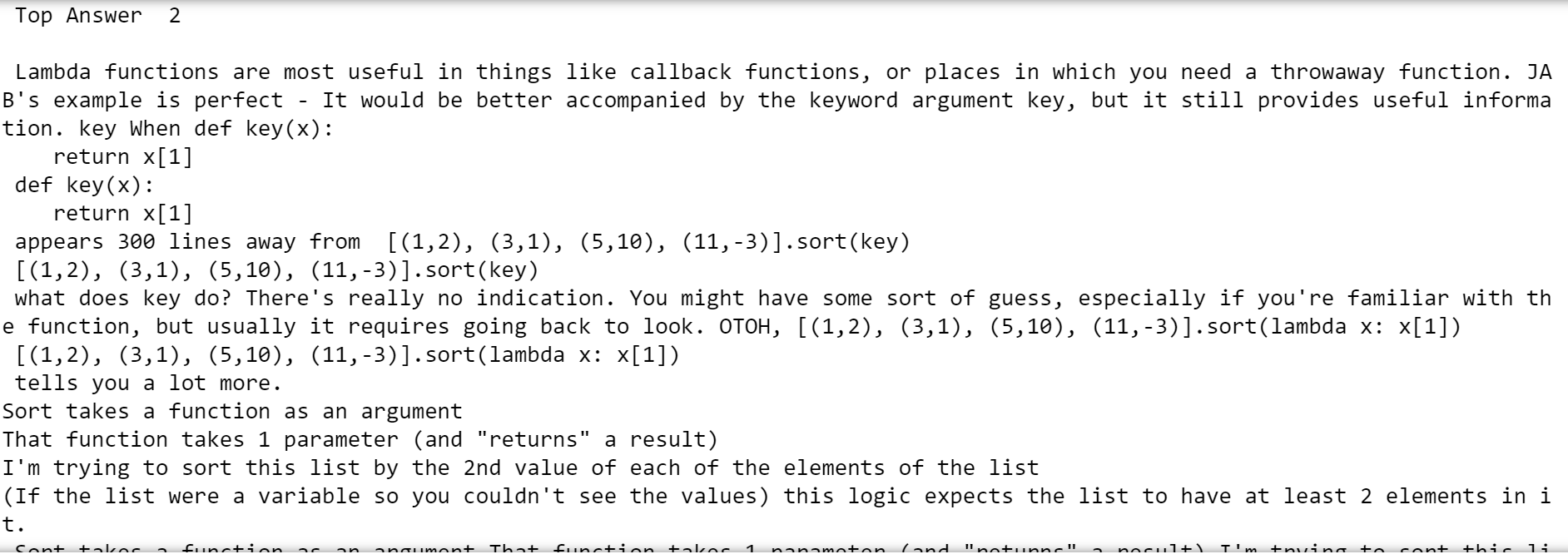
**Figure 7. Top Ranked Cluster Results (Bert Cluster 3)**

Figure 7 shows the results retrieved after question and answer relevance mapping.

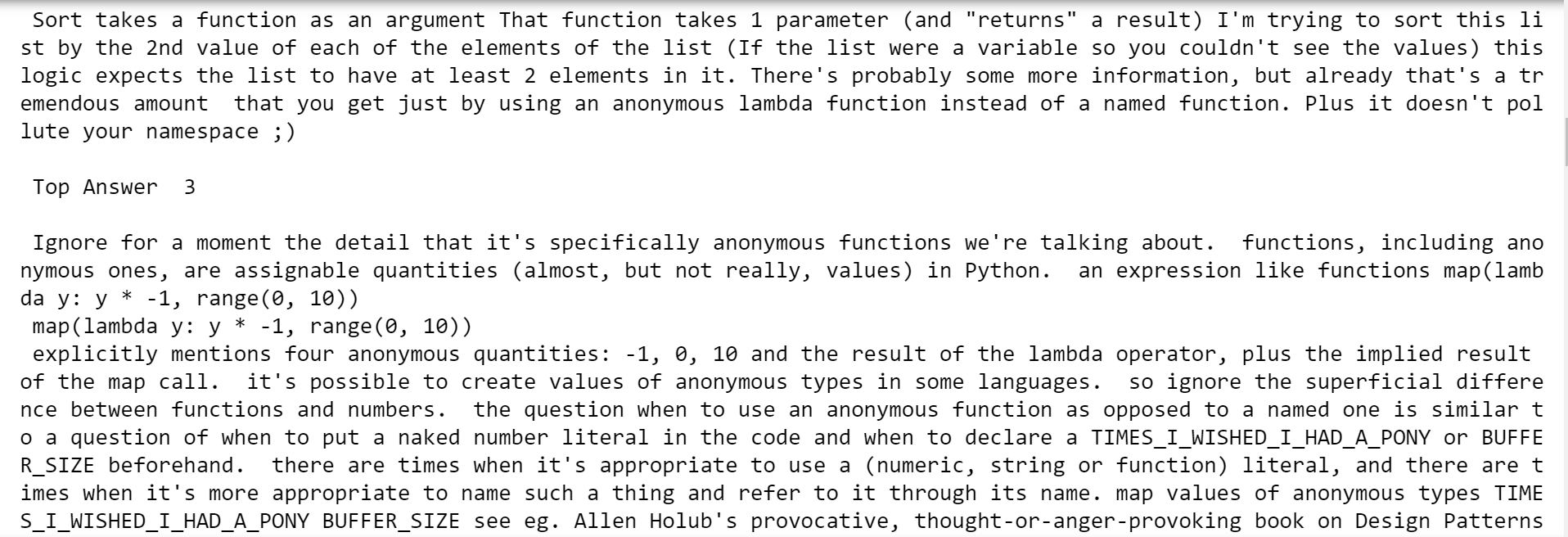
Figure 8, Figure 9, & Figure 10 below are the detailed views of the first 3 top answers, respectively.



**Figure 8. First Top Answer**



**Figure 9. Second Top Answer**



**Figure 10. Third Top Answer**