Knowledge Management and Analysis
Project 01: Code Search

Claudio Maggioni

Section 1 - Data Extraction

The data extraction (implemented in the script `extract-data.py`) process scans through the files in the TensorFlow project to extract Python docstrings and symbol names for functions, classes and methods. A summary of the number of features extracted can be found in table 1. The collected figures show that the number of classes is more than half the number of files, while the number of functions is about twice the number of files. Additionally, the data shows that a class has slightly more than 2 methods in it on average.

Type	Number
Python files	2817
Classes	1882
Functions	4565
Methods	5817

Table 1: Count of created classes and properties.

Section 2: Training of search engines

The training and model execution of the search engines is implemented in the Python script `search-data.py`. The script is able to search a given natural language query among the extracted TensorFlow corpus using four techniques. These are namely: Word Frequency Similarity, Term-Frequency Inverse Document-Frequency (TF-IDF) Similarity, Latent Semantic Indexing (LSI), and Doc2Vec.

An example output of results generated from the query "Gather gpu device info" for the word frequency, TF-IDF, LSI and Doc2Vec models are shown in figures 1, 2, 3 and 4 respectively. Both the word frequency and TF-IDF model identify the correct result (according to the provided ground truth for this query) as the first recommendation to output. Both the LSI and Doc2Vec models fail to report the correct function in all 5 results.

Section 3: Evaluation of search engines

The evaluation over the given ground truth to compute precision, recall, and the T-SNE plots is performed by the script `prec-recall.py`. The calculated

average precision and recall values are reported in table 2.

Precision and recall is quite low for all models, less so for the word frequency and the TF-IDF models. The word frequency model has the highest precision and recall (27% and 40% respectively), while the LSI model has the lowest precision (4%) and Doc2Vec has the lowest recall (10%).

Engine	Avg Precision	Recall
Frequencies	27.00%	40.00%
TD-IDF	20.00%	20.00%
LSI	4.00%	20.00%
Doc2Vec	10.00%	10.00%

Table 2: Evaluation of search engines.

TBD Section 4: Visualisation of query results

The two-dimensional T-SNE plots (computed with perplexity = 1) for the LSI and Doc2Vec models are respectively in figures 5 and 6.

The T-SNE plot for the LSI model shows evidently the presence of outliers in the search result. The Doc2Vec plot shows fewer outliers and more distinct clusters for the results of each query and the query vector itself.

```
Similarity: 87.29%
Python function: gather_gpu_devices
Description: Gather gpu device info. Returns: A list of test_log_pb2.GPUInf...
File: tensorflow/tensorflow/tools/test/gpu_info_lib.py
Line: 167

Similarity: 60.63%
Python function: compute_capability_from_device_desc
Description: Returns the GpuInfo given a DeviceAttributes proto. Args: devi...
File: tensorflow/tensorflow/python/framework/gpu_util.py
Line: 35

Similarity: 60.30%
Python function: gpu_device_name
Description: Returns the name of a GPU device if available or the empty str...
File: tensorflow/tensorflow/python/framework/test_util.py
Line: 129

Similarity: 58.83%
Python function: gather_available_device_info
Description: Gather list of devices available to TensorFlow. Returns: A lis...
File: tensorflow/tensorflow/tools/test/system_info_lib.py
Line: 126

Similarity: 57.74%
Python function: gather_memory_info
Description: Gather memory info.
File: tensorflow/tensorflow/tools/test/system_info_lib.py
Line: 70
```

Figure 1: Search result output for the query "Gather gpu device info" using the word frequency similarity model.

```
Similarity: 86.62%
Python function: gather_gpu_devices
Description: Gather gpu device info. Returns: A list of test_log_pb2.GPUInf...
File: tensorflow/tensorflow/tools/test/gpu_info_lib.py
Line: 167

Similarity: 66.14%
Python function: gather_memory_info
Description: Gather memory info.
File: tensorflow/tensorflow/tools/test/system_info_lib.py
Line: 70

Similarity: 62.52%
Python function: gather_available_device_info
Description: Gather list of devices available to TensorFlow. Returns: A lis...
File: tensorflow/tensorflow/tools/test/system_info_lib.py
Line: 126

Similarity: 57.98%
Python function: gather
File: tensorflow/tensorflow/compiler/tf2xla/python/xla.py
Line: 452

Similarity: 57.98%
Python function: gather_v2
File: tensorflow/tensorflow/python/ops/array_ops.py
Line: 4736
```

Figure 2: Search result output for the query "Gather gpu device info" using the TF-IDF model.

```
Similarity: 92.11%
Python function: device
Description: Uses gpu when requested and available.
File: tensorflow/tensorflow/python/framework/test_util.py
Line: 1581

Similarity: 92.11%
Python function: device
Description: Uses gpu when requested and available.
File: tensorflow/tensorflow/python/keras/testing_utils.py
Line: 925

Similarity: 89.04%
Python function: compute_capability_from_device_desc
Description: Returns the GpuInfo given a DeviceAttributes proto. Args: devi...
File: tensorflow/tensorflow/python/framework/gpu_util.py
Line: 35

Similarity: 85.96%
Python class: CUDADeviceProperties
File: tensorflow/tensorflow/tools/test/gpu_info_lib.py
Line: 51

Similarity: 85.93%
Python function: gpu_device_name
Description: Returns the name of a GPU device if available or the empty str...
File: tensorflow/tensorflow/python/framework/test_util.py
Line: 129
```

Figure 3: Search result output for the query "Gather gpu device info" using the LSI model.

```
Similarity: 81.85%
Python method: benchmark_gather_nd_op
File: tensorflow/tensorflow/python/kernel_tests/gather_nd_op_test.py
Line: 389

Similarity: 81.83%
Python function: gather_hostname
File: tensorflow/tensorflow/tools/test/system_info_lib.py
Line: 66

Similarity: 81.07%
Python method: benchmarkNontrivialGatherAxis1XLA
File: tensorflow/tensorflow/compiler/tests/gather_test.py
Line: 210

Similarity: 80.53%
Python method: benchmarkNontrivialGatherAxis4
File: tensorflow/tensorflow/compiler/tests/gather_test.py
Line: 213

Similarity: 80.45%
Python method: benchmarkNontrivialGatherAxis4XLA
File: tensorflow/tensorflow/compiler/tests/gather_test.py
Line: 216
```

Figure 4: Search result output for the query "Gather gpu device info" using the Doc2Vec model.

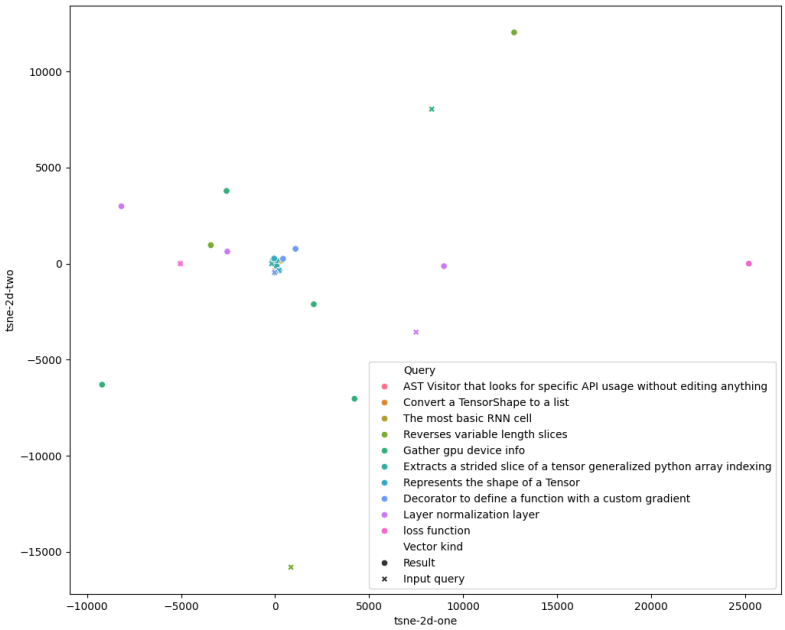

Figure 5: T-SNE plot for the LSI model over the queries and ground truths given in `ground-truth-unique.txt`.

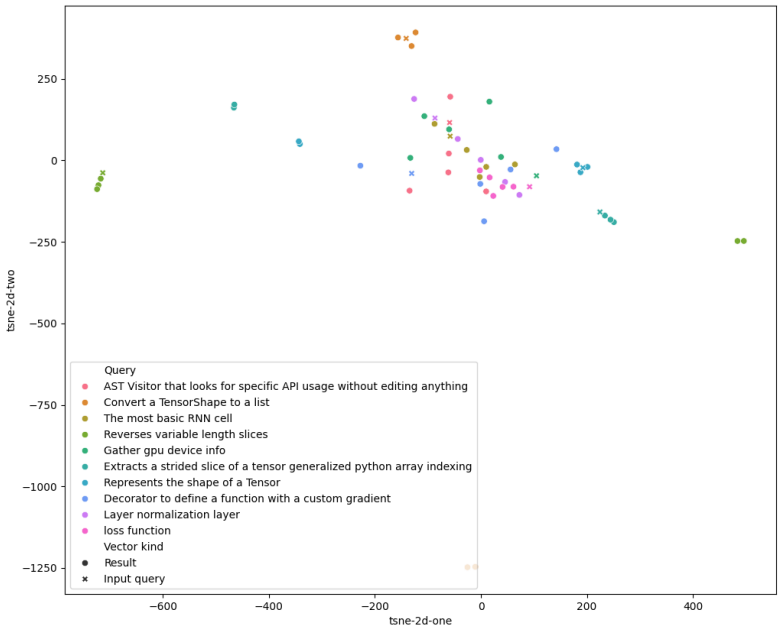

Figure 6: T-SNE plot for the Doc2Vec model over the queries and ground truths given in `ground-truth-unique.txt`.