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 training model loads the data extracted by `extract-data.py` and uses as classification features the identifier name and only the first line of the comment docstring. All other comment lines are filtered out as this significantly increases performance when evaluating the models.

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. All four models are able to correctly report the ground truth required by the file `ground-truth-unique.txt` as the first result with $> 90\%$ similarity, with the except of the Doc2Vec model which reports 71.63% similarity.

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 are quite high for all models. The word frequency model has the highest precision and recall (93.33% and 100.00% respectively), while the Doc2Vec model has the lowest precision (73.33%) and lowest recall (80.00%).

Engine	Avg Precision	Recall
Frequencies	93.33%	100.00%
TD-IDF	90.00%	90.00%
LSI	90.00%	90.00%
Doc2Vec	73.33%	80.00%

Table 2: Evaluation of search engines.

TBD Section 4: Visualisation of query results

The two-dimensional T-SNE plots (computed with perplexity = 2) 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. However, even considering the good performance for both models, it is hard to distinguish from the plots given distinct "regions" where results and their respective query are located.

```
Similarity: 90.45%
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: 57.74%
Python function: gather_memory_info
Description: Gather memory info.
File: tensorflow/tensorflow/tools/test/system_info_lib.py
Line: 70

Similarity: 57.74%
Python function: gather_platform_info
Description: Gather platform info.
File: tensorflow/tensorflow/tools/test/system_info_lib.py
Line: 146

Similarity: 55.47%
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: 55.47%
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
```

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

```
Similarity: 90.95%
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: 59.12%
Python function: gather_memory_info
Description: Gather memory info.
File: tensorflow/tensorflow/tools/test/system_info_lib.py
Line: 70

Similarity: 56.40%
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: 55.25%
Python function: gather_platform_info
Description: Gather platform info.
File: tensorflow/tensorflow/tools/test/system_info_lib.py
Line: 146

Similarity: 53.97%
Python function: info
File: tensorflow/tensorflow/python/platform/tf_logging.py
Line: 167
```

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

```
Similarity: 98.38%
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: 97.66%
Python function: device
Description: Uses gpu when requested and available.
File: tensorflow/tensorflow/python/framework/test_util.py
Line: 1581

Similarity: 97.66%
Python function: device
Description: Uses gpu when requested and available.
File: tensorflow/tensorflow/python/keras/testing_utils.py
Line: 925

Similarity: 96.79%
Python class: ParallelDevice
Description: A device which executes operations in parallel.
File: tensorflow/tensorflow/python/distribute/parallel_device/parallel_device.py
Line: 42

Similarity: 96.67%
Python method: get_var_on_device
File: tensorflow/tensorflow/python/distribute/packed_distributed_variable.py
Line: 90
```

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

```
Similarity: 71.63%
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: 66.71%
Python function: device
Description: Uses gpu when requested and available.
File: tensorflow/tensorflow/python/keras/testing_utils.py
Line: 925

Similarity: 65.23%
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: 64.33%
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: 64.29%
Python method: hosts
Description: A list of device names for CPU hosts. Returns: A list of devic...
File: tensorflow/tensorflow/python/tpu/tpu_embedding.py
Line: 1011
```

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

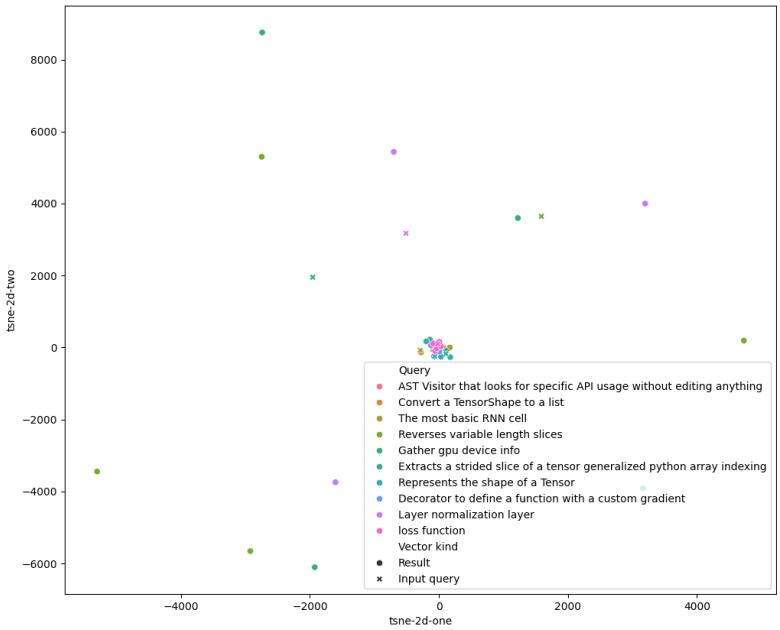

Figure 5: T-SNE plot for the LSI model over the queries and ground truths given in `ground-truth-unique.txt`.

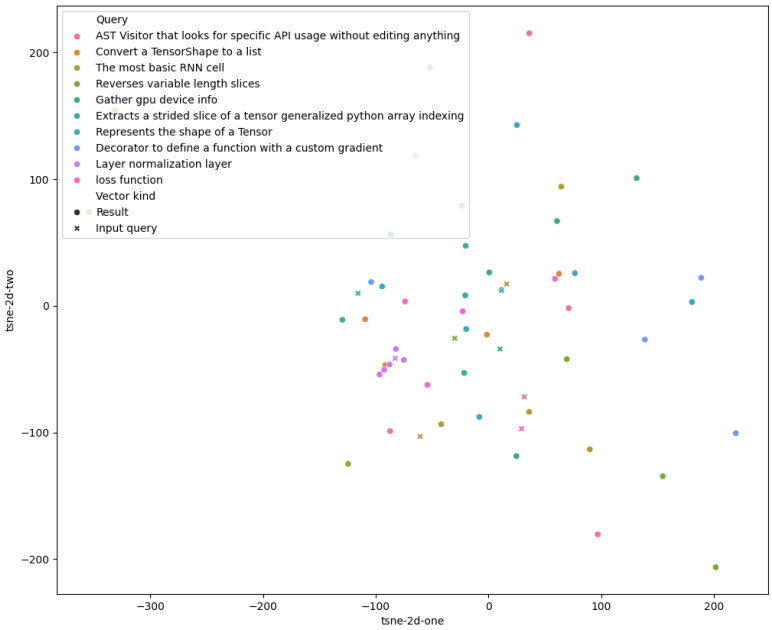

Figure 6: T-SNE plot for the Doc2Vec model over the queries and ground truths given in `ground-truth-unique.txt`.