Project 2: Multi-source code search

Ardigò Susanna

15 December, 2020

https://github.com/SusyPinkBash/multi_source_code_search

Contents

1 Data Extraction

1.1 Goal and Input parameter

This part of the project consists of extracting names and comments of Python classes, methods and functions and save them in a csv file.
This file takes as argument the path of the directory of the project that we want to analyze. For this project we use the project `tensorflow`.

1.2 Description of the code

To efficiently parse the files in the directory, we created a class named `Visitor`, which extends the `NodeVisitor` class of the standard library ast (which stands for Abstract Syntax Tree). This class holds the path of the file. There is a global variable `data` used throughout the execution to store all the information extracted.
The function `start(directory_path)` *'walks'* the given directory using the function `walk` which generates a 3-tuple of directory path, directory names and file names. We open and read all the python files, checked with the extension of the file, we create a Visitor object and start to visit. The class we created has two different visit methods which differ in if the node visiting is a definition of a class or a function. The method `visit_FunctionDef(self, node: FunctionDef)` adds the node information to the array of data if the function or method is not a main or a test. Since this method is used both for functions and methods, we know if the node is a method by checking if the first argument is `self`. The method `visit_ClassDef(self, node: ClassDef)` calls a generic visit (of the ast library) and, as the previous method, adds the node information to the array of data if the class is not a main or a test.
After the parsing is complete we create a pandas dataframe, feeding it as data the data array, and export it in a csv extension.

1.3 Results

Table 1 show the number of Python files, classes, methods and functions found while parsing the Tensorflow directory. The results can be found in the file `res/data.csv`.

Type	**#**
Python files	2817
Classes	1904
Methods	7271
Functions	4881

Table 1: Count of data found in Tensorflow

2 Training of search engines

2.1 Goal and Input parameter

This part of the project consists of representing code entities using the four embeddings frequency, TF-IDF, LSI and Doc2Vec.
This file takes as argument a query which will be fed to the four search engines.

2.2 Description of the code

The function `start(query)` loads the csv into a pandas dataframe and then computes the results.
The first part of function `compute_results(query, dataframe)` creates the necessary data and normalizes the query that the second part needs to produce the results.
The first part of function `create_data(dataframe)` extracts the names and comments of the data extracted in the first part. to create a clean array of arrays of tokens and a dictionary with the frequencies of each token. In the second part we create the corpus by processing the tokens, we create a gensim dictionary and the bag of words. At the end of the creation, we save the corpus, dictionary and bag of words in external files to then load them in future runs. In the second part of function `compute_results(query, dataframe)` we create a dictionary that hold the results of the searches and a dictionary to save the embedding vectors.
The function `query_frequency(query, bow, dictionary)` creates a sparse matrix of the bag of words and returns an array with the similarity scores of each entity of the given csv file. This array is then filtered to extract only the top 5 scoring entities. Similarly, the function `query_tfidf(query, bow, dictionary)` creates a sparse matrix of the tfidf model of the bag of words and returns an array with the similarity scores which is then filtered. The function `query_lsi(query, bow, dictionary)` creates a lsi model based on the bag of words, a vector based on the model and the dictionary, the matrix of the similarities and the embedding vectors. The result of the matrix, as in the previous cases, is filtered to get only the top 5 scores. The function `query_doc2vec(query, bow, dictionary)` creates a doc2vec model which then feed the corpus to and train it. We create a vector infering it from the query, we create the similarity and take only the top 5 scores and the embedding vectors.
We save the trained models in external pickle files to load then load them in the next runs. This improves the running time of the function.
We create a dataframe with the information stored in the dictionary, we print the results and save them in a separate file.

2.3 Results

To show the results we run this part of the project with the query:

'AST Visitor that looks for specific API usage without editing anything' .

The correct document is `PastaAnalyzeVisitor` with path `../tensorflow/tensorflow/tools/compatibility/ast_edits.py`.
Figure 1 show the result of the given query.
As we can see in the image all search engine find the correct document as a first result. We can see that Frequency and TF-IDF have three results in common, if we don't consider the correct result, but at different positions.

	name	file	line	type	comment	search
1	PastaAnalyzeVisitor	../tensorflow/tensorflow/tools/compatibility/ast_edits.py	815	class	AST Visitor that looks for specific API usage without editing anything.	FREQ
2	CleanCopier	../tensorflow/tensorflow/python/autograph/pyct/ast_util.py	30	class	NodeTransformer-like visitor that copies an AST.	FREQ
3	CompatV1ImportReplacer	../tensorflow/tensorflow/tools/compatibility/tf_upgrade_v2.py	72	class	AST Visitor that replaces `import tensorflow.compat.v1 as tf`.	FREQ
4	FunctionVisitor	../tensorflow/tensorflow/python/autograph/pyct/static_analysis/type_inference.py	394	class	AST visitor that applies type inference to each function separately.	FREQ
5	track_usage	../tensorflow/tensorflow/python/platform/analytics.py	21	function	No usage tracking for external library.	FREQ
6	PastaAnalyzeVisitor	../tensorflow/tensorflow/tools/compatibility/ast_edits.py	815	class	AST Visitor that looks for specific API usage without editing anything.	TF-IDF
7	FunctionVisitor	../tensorflow/tensorflow/python/autograph/pyct/static_analysis/type_inference.py	394	class	AST visitor that applies type inference to each function separately.	TF-IDF
8	track_usage	../tensorflow/tensorflow/python/platform/analytics.py	21	function	No usage tracking for external library.	TF-IDF
9	CleanCopier	../tensorflow/tensorflow/python/autograph/pyct/ast_util.py	30	class	NodeTransformer-like visitor that copies an AST.	TF-IDF
10	PublicAPIVisitor	../tensorflow/tensorflow/tools/common/public_api.py	29	class	Visitor to use with `traverse` to visit exactly the public TF API.	TF-IDF
11	PastaAnalyzeVisitor	../tensorflow/tensorflow/tools/compatibility/ast_edits.py	815	class	AST Visitor that looks for specific API usage without editing anything.	LSI
12	matches	../tensorflow/tensorflow/python/autograph/pyct/ast_util.py	214	function	Basic pattern matcher for AST.	LSI
13	CountingVisitor	../tensorflow/tensorflow/python/autograph/pyct/cfg_test.py	28	class	*<null>*	LSI
14	ProfileOptionBuilder	../tensorflow/tensorflow/python/profiler/option_builder.py	27	class	Option Builder for Profiling API.	LSI
15	concentration	../tensorflow/tensorflow/python/ops/distributions/dirichlet.py	213	method	Concentration parameter; expected counts for that coordinate.	LSI
16	PastaAnalyzeVisitor	../tensorflow/tensorflow/tools/compatibility/ast_edits.py	815	class	AST Visitor that looks for specific API usage without editing anything.	Doc2Vec
17	APIChangeSpec	../tensorflow/tensorflow/tools/compatibility/ast_edits.py	192	class	This class defines the transformations that need to happen.	Doc2Vec
18	Profile	../tensorflow/tensorflow/python/profiler/profiler_v2.py	181	class	Context-manager profile API.	Doc2Vec
19	TFAPIChangeSpec	../tensorflow/tensorflow/tools/compatibility/tf_upgrade_v2_safety.py	26	class	List of maps that describe what changed in the API.	Doc2Vec
20	Profiler	../tensorflow/tensorflow/python/eager/profiler.py	176	class	Context-manager eager profiler api.	Doc2Vec

Figure 1: Results of the given query

3 Evaluation of search engines

3.1 Goal and Input parameter

This part of the project consists of measuring the precision and recall given 10 queries along with their ground truth.
This file takes as argument the path of the ground truth file.

3.2 Description of the code

The function `start(path_ground_truth)` loads the csv of the data into a pandas dataframe, parses the ground truth and then computes the precision and recall.
To efficiently parse the ground truth file, we created a class named `Truth` which holds the name, path and query. We read the ground truth file and create an array with all the entries of the ground truth and the queries.
To compute precision and recall we get the data of the results and the embedding vectors from the previous part. We create a dictionary to save the scores of the queries and a dictionary for the vectors. We then compute the precision and recall, by comparing our results and the ground truth.

3.3 Results

Table 2 show the statistics of precision and recall compared to the unique ground truth. We can see that the precision is high for all engines except for one. The engine with the highest precision is `TF-IDF`, with score 8.85. The second highest precision is `Frequencies`, with score 0.83. The third highest precision is `LSI`, with score 0.80. The least precise search engine is `Doc2Vec` with score 0.55, which is the only score lower than 0.8
The recall is higher than the precision. The `TF-IDF` engine has a recall equal to 1, which means that for each query the search engine has found the correct result in the top 5. The second highest recall is of `Frequencies` with score 0.9. It is followed by `LSI` with a score of 0.9. At last `Doc2Vec` has a recall of 0.6.
We can say that almost all search engine have similar scores for precision and recalls. The only exception is `Doc2Vec`.

Engine	**Precision**	**Recall**
Frequencies	0.83	0.90
TD-IDF	0.85	1.00
LSI	0.80	0.80
Doc2Vec	0.55	0.60

Table 2: Statistics of the search engines

4 Visualisation of query results

4.1 Goal and Input parameter

This part of the project consists of visualizing the embedding vectors of the queries and the top 5 answers in a 2D plot. This file takes as argument the ground truth file.

4.2 Description of the code

The first part of the execution is the same as the previous file. After the results are calculated, we use the embedding vectors, that we retrieved in the explanation above but we did not use. For vector we apply TSNE to produce 2D vectors composed of queries and the top 4 results. The plot is straight-forward: we create a dataframe with the information of x and y coordinates and print them of different hues. We use the library `seaborn` to create the charts and we then save them on disk.

4.3 Results

Figures 2 and 3 shows the plots of the visualization of the queries. At first we notice that the LSI scatterplot tends to have better query clusters. The optimal solution is to have defined clusters for each query. This does not happen in any of the two images. There are queries that create clusters but this does not happen for all queries.

4.3.1 LSI

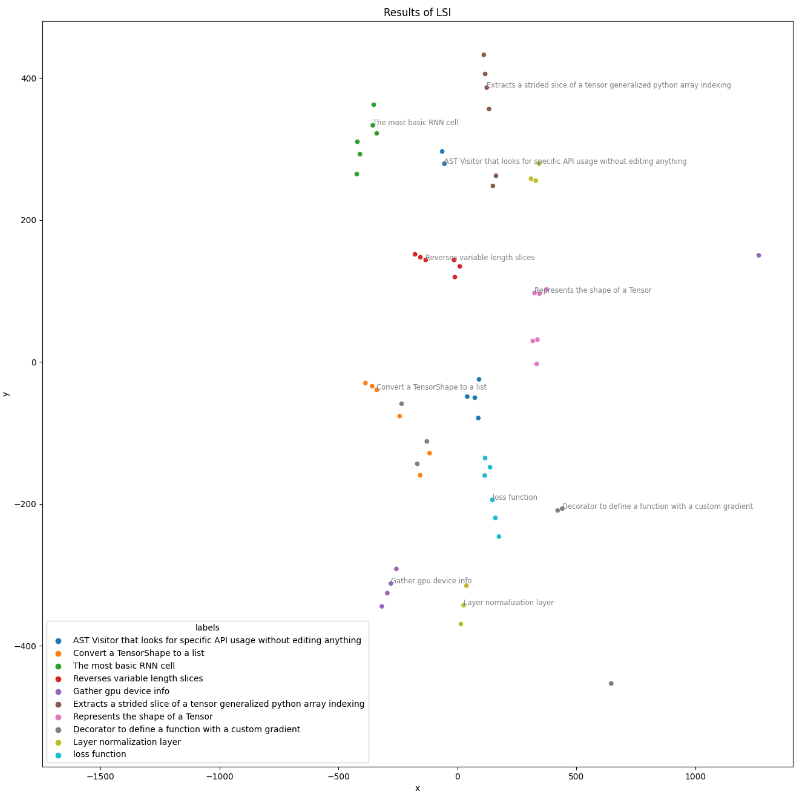

Figure 2: Visualization of the queries using LSI

Analyzing the plot of LSI, shown in figure 2, we can see that some results of the queries tend to stay close, but not completely. There are some well defined clusters composed of the same queries.

Some examples are *loss function*, *The most basic RNN cell* and *Reverses variable length slices*. The queries of *Convert a TensorShape to a list* and *Decorator to define a function with a custom gradient* that tend to have some results that are somewhat close.

4.3.2 Doc2Vec

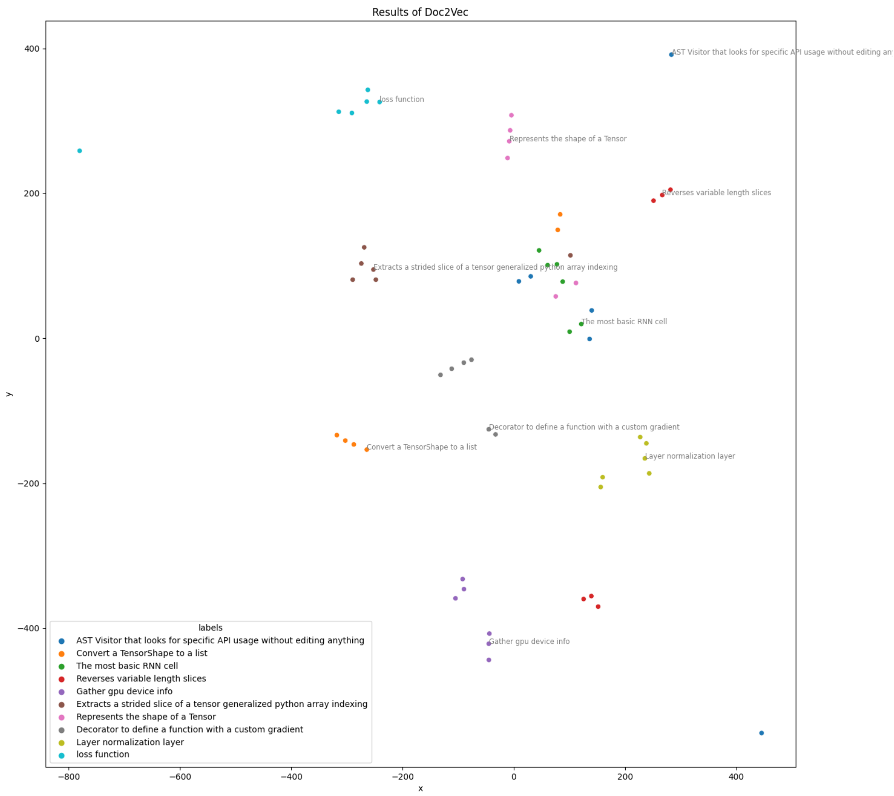

Figure 3: Visualization of the queries using Doc2Vec

Analyzing the plot of Doc2Vec, shown in figure 3, we can see that there are results of the queries tend to stay closer than the previous plot. In this image there are some cluster that are well defined and are composed only of one query. Some queries are divided into two clusters close to each other. The most well defined cluster is *'loss function'*, which was also well defined in the previous plot. There is a cluster which is composed of five different queries.

A Python code

A.1 Data Extraction

```
from sys import argv, exit
from ast import *
from os import walk
import pandas as pd


class Visitor(NodeVisitor):
    def __init__(self, file_path, node):
        super().__init__()
        self.file_path = clean_file_path(file_path)
        self.visit(parse(node))

    def visit_ClassDef(self, node: ClassDef):
        self.generic_visit(node)
        if is_valid_entity(node.name):
            self.append_data(node, "class")

    def visit_FunctionDef(self, node: FunctionDef):
        if is_valid_entity(node.name):
            self.append_data(node, "method" if is_method(node) else "function")

    def append_data(self, node, def_type):
        comment = get_docstring(node)
        comment = comment.split('\n')[0] if comment is not None else ""
        data.append((node.name, self.file_path, node.lineno, def_type, comment))


def clean_file_path(path):
    directories = path.split('/')
    return '../' + '/'.join(directories[directories.index('tensorflow'):])


def is_valid_entity(name):
    return name[0] != '_' and name != "main" and "test" not in name.lower()


def is_method(function):
    return function.args and len(function.args.args) > 0 and 'self' in function.args.args[0].
        arg


def start(directory_path):
    if directory_path[-1] == '/':
        directory_path = directory_path[: -1]
    counter = 0
    for path, _, files in walk(directory_path):
        for file_name in files:
            if file_name.endswith('.py'):
                counter += 1
                file_path = path + '/' + file_name
                with open(file_path) as file:
                    Visitor(file_path, file.read())
```

```python

    dataframe = pd.DataFrame(data=data, columns=["name", "file", "line", "type", "comment"])
    dataframe.to_csv('res/data.csv', index=False, encoding='utf-8')
    print("files\t    " + str(counter))
    print(dataframe["type"].value_counts())


if len(argv) < 2:
    print("Please give as input the path of the directory to analyze")
    exit(1)
data = []
start(argv[1])
```

A.2 Training of search engines

```python
from datetime import datetime
import string
from os import path
import pandas as pd
import pickle as pkl
from re import finditer
from sys import argv, exit
from collections import defaultdict
from gensim.corpora import Dictionary
from gensim.models.doc2vec import TaggedDocument
from gensim.utils import simple_preprocess
from gensim.models import TfidfModel, LsiModel, Doc2Vec
from gensim.similarities import MatrixSimilarity, SparseMatrixSimilarity


def start(query):
    dataframe = pd.read_csv("res/data.csv").fillna(value="")
    results_dictionary, _ = compute_results(query, dataframe)
    results = pd.DataFrame(data=create_result_dataframe(results_dictionary, dataframe),
                           columns=['name', "file", "line", "type", "comment", "search"])
    pd.options.display.max_colwidth = 200
    print_results(results)
    results.to_latex('res/search_data.tex', index=False, encoding='utf-8')
    results.to_csv('res/search_data.csv', index=False, encoding='utf-8')


def compute_results(query, dataframe):
    processed_corpus, frequencies, bag_of_words = get_data(dataframe)
    query_to_execute = normalize_query(query)
    results = {
        "FREQ": query_frequency(query_to_execute, bag_of_words, frequencies),
        "TF-IDF": query_tfidf(query_to_execute, bag_of_words, frequencies)
    }
    vectors = dict()
    results["LSI"], vectors["LSI"] = query_lsi(query_to_execute, bag_of_words, frequencies)
    results["Doc2Vec"], vectors["Doc2Vec"] = query_doc2vec(query_to_execute, processed_corpus)
    return results, vectors


```

```python
def get_data(df):
    return load_data_files() if exists_data_files() else create_data(df)


def create_data(df):
    tokens = [filter_stopwords(normalize_tokens(handle_camel_case(split_underscore(
        [row["name"]] + split_space(row["comment"]))))) for _, row in df.iterrows()]

    frequency = defaultdict(int)
    for token in tokens:
        for word in token:
            frequency[word] += 1

    corpus = [[token for token in text if frequency[token] > 1] for text in tokens]
    dictionary = Dictionary(corpus)
    bow = [dictionary.doc2bow(text) for text in corpus]

    save_data(corpus, 'corpus')
    save_data(dictionary, 'dictionary')
    save_data(bow, 'bow')
    return corpus, dictionary, bow


def exists_data_files():
    return exists_file('corpus') and exists_file('dictionary') and exists_file('bow')


def exists_file(name):
    return path.exists('res/pickle/' + name + '.pkl')


def load_data_files():
    return load_file('corpus'), load_file('dictionary'), load_file('bow')


def save_data(data, name):
    pkl.dump(data, open('res/pickle/' + name + '.pkl', "wb"), protocol=pkl.HIGHEST_PROTOCOL)


def load_file(name):
    return pkl.load(open('res/pickle/' + name + '.pkl', "rb"))


def split_space(text):
    return text.translate(str.maketrans('', '', string.punctuation)).split(' ') if text != ""
        else []


def split_underscore(tokens):
    return [word for token in tokens for word in token.split('_')]


def handle_camel_case(tokens):
    words = []
    for token in tokens:
        matches = finditer('.+?(?:(?<=[a-z])(?=[A-Z])|(?<=[A-Z])(?=[A-Z][a-z])|$)', token)
        words += [m.group(0) for m in matches]
```

```
    return words


def normalize_tokens(tokens):
    return [token.lower() for token in tokens]


def filter_stopwords(tokens):
    for token in tokens:
        if token in ['test', 'tests', 'main']:
            return []
    return tokens


def normalize_query(query):
    return query.strip().lower().split()


def query_frequency(query, bow, dictionary):
    return filter_results(create_top_5_result_tuples(get_freq_model(bow, dictionary)[
        dictionary.doc2bow(query)]))


def get_freq_model(bow, dictionary):
    return load_file('model_freq') if exists_file('model_freq') else create_freq_model(bow,
        dictionary)


def create_freq_model(bow, dictionary):
    model = SparseMatrixSimilarity(bow, num_features=len(dictionary.token2id))
    save_data(model, 'model_freq')
    return model


def query_tfidf(query, bow, dictionary):
    model = get_tfidf_model(bow)
    matrix = get_tfidf_matrix(model, bow, dictionary)
    return filter_results(create_top_5_result_tuples(matrix[model[dictionary.doc2bow(query)
        ]]))


def get_tfidf_model(bow):
    return load_file('model_tfidf') if exists_file('model_tfidf') else create_tfidf_model(bow
        )


def create_tfidf_model(bow):
    model = TfidfModel(bow)
    save_data(model, 'model_tfidf')
    return model


def get_tfidf_matrix(model, bow, dictionary):
    return load_file('matrix_tfidf') if exists_file('matrix_tfidf') else create_tfidf_matrix(
        model, bow, dictionary)

```

```
def create_tfidf_matrix(model, bow, dictionary):
    matrix = SparseMatrixSimilarity(model[bow], num_features=len(dictionary.token2id))
    save_data(matrix, 'matrix_tfidf')
    return model


def query_lsi(query, bow, dictionary):
    model = get_lsi_model(bow, dictionary)
    matrix = get_lsi_matrix(model, bow)
    vector = model[dictionary.doc2bow(query)]
    result = abs(matrix[vector])
    embedding = [[value for _, value in vector]] + [[value for _, value in model[bow][i]] for
         i, value in
                                                sorted(enumerate(result), key=lambda x: x
                                                    [1], reverse=True)[:5]]
    return filter_results(create_top_5_result_tuples(result)), embedding


def get_lsi_model(bow, dictionary):
    return load_file('model_lsi') if exists_file('model_lsi') else create_lsi_model(bow,
        dictionary)


def create_lsi_model(bow, dictionary):
    model = LsiModel(bow, id2word=dictionary, num_topics=300)
    save_data(model, 'model_lsi')
    return model


def get_lsi_matrix(model, bow):
    return load_file('matrix_lsi') if exists_file('matrix_lsi') else create_lsi_matrix(model,
         bow)


def create_lsi_matrix(model, bow):
    matrix = MatrixSimilarity(model[bow])
    save_data(matrix, 'matrix_lsi')
    return matrix


def create_top_5_result_tuples(arrg):
    return sorted(enumerate(arrg), key=lambda x: x[1], reverse=True)[:5]


def filter_results(tuples):
    return [i for i, v in tuples]


def query_doc2vec(query, corpus):
    model = get_doc2vec_model(get_doc2vec_corpus(corpus))
    vector = model.infer_vector(query)
    similar = model.docvecs.most_similar([vector], topn=5)
    return [index for (index, _) in similar], \
           [list(vector)] + [list(model.infer_vector(corpus[index])) for index, _ in similar]


def get_doc2vec_corpus(corpus):
```

```
    return [TaggedDocument(simple_preprocess(' '.join(element)), [index])
            for index, element in enumerate(corpus)]


def get_doc2vec_model(corpus):
    return load_file('model_doc2vec') if exists_file('model_doc2vec') else
        create_doc2vec_model(corpus)


def create_doc2vec_model(corpus):
    model = Doc2Vec(vector_size=300, min_count=2, epochs=77)
    model.build_vocab(corpus)
    model.train(corpus, total_examples=model.corpus_count, epochs=model.epochs)
    save_data(model, 'model_doc2vec')
    return model


def create_result_dataframe(queries_dictionary, df):
    for key, values in queries_dictionary.items():
        for index in values:
            row = df.iloc[index]
            yield [row["name"], row["file"], row["line"], row["type"], row["comment"], key]


def print_results(df):
    grouped = df.groupby(['search'])
    for key, item in grouped:
        print(grouped.get_group(key), "\n\n")


if len(argv) < 2:
    print("Please give as input the query")
    exit(1)

print("NOPE")
start(argv[1])
```

A.3 Evaluation of search engines and Visualisation of query results

```
import itertools
from datetime import datetime

import string
import pandas as pd
from os import path
import pickle as pkl
import seaborn as sns
from re import finditer
from sys import argv, exit
import matplotlib.pyplot as plt
from sklearn.manifold import TSNE
from collections import defaultdict
from gensim.corpora import Dictionary
from gensim.models.doc2vec import TaggedDocument
from gensim.utils import simple_preprocess
```

```
from gensim.models import TfidfModel, LsiModel, Doc2Vec
from gensim.similarities import MatrixSimilarity, SparseMatrixSimilarity

##################
def get_results(query, dataframe):
    results_dictionary, vectors = compute_results(query, dataframe)
    return pd.DataFrame(data=create_result_dataframe(results_dictionary, dataframe),
                        columns=['name', "file", "line", "type", "comment", "search"]),
                            vectors


def compute_results(query, dataframe):
    processed_corpus, frequencies, bag_of_words = get_data(dataframe)
    query_to_execute = normalize_query(query)
    results = {
        "FREQ": query_frequency(query_to_execute, bag_of_words, frequencies),
        "TF-IDF": query_tfidf(query_to_execute, bag_of_words, frequencies)
    }
    vectors = dict()
    results["LSI"], vectors["LSI"] = query_lsi(query_to_execute, bag_of_words, frequencies)
    results["Doc2Vec"], vectors["Doc2Vec"] = query_doc2vec(query_to_execute, processed_corpus
        )
    return results, vectors


def get_data(df):
    return load_data_files() if exists_data_files() else create_data(df)


def create_data(df):
    tokens = [filter_stopwords(normalize_tokens(handle_camel_case(split_underscore(
        [row["name"]] + split_space(row["comment"]))))) for _, row in df.iterrows()]

    frequency = defaultdict(int)
    for token in tokens:
        for word in token:
            frequency[word] += 1

    corpus = [[token for token in text if frequency[token] > 1] for text in tokens]
    dictionary = Dictionary(corpus)
    bow = [dictionary.doc2bow(text) for text in corpus]

    save_data(corpus, 'corpus')
    save_data(dictionary, 'dictionary')
    save_data(bow, 'bow')
    return corpus, dictionary, bow


def exists_data_files():
    return exists_file('corpus') and exists_file('dictionary') and exists_file('bow')


def exists_file(name):
    return path.exists('res/pickle/' + name + '.pkl')


def load_data_files():
```

```python
    return load_file('corpus'), load_file('dictionary'), load_file('bow')


def save_data(data, name):
    pkl.dump(data, open('res/pickle/' + name + '.pkl', "wb"), protocol=pkl.HIGHEST_PROTOCOL)


def load_file(name):
    return pkl.load(open('res/pickle/' + name + '.pkl', "rb"))


def split_space(text):
    return text.translate(str.maketrans('', '', string.punctuation)).split(' ') if text != "" else []


def split_underscore(tokens):
    return [word for token in tokens for word in token.split('_')]


def handle_camel_case(tokens):
    words = []
    for token in tokens:
        matches = finditer('.+?(?:(?<=[a-z])(?=[A-Z])|(?<=[A-Z])(?=[A-Z][a-z])|$)', token)
        words += [m.group(0) for m in matches]
    return words


def normalize_tokens(tokens):
    return [token.lower() for token in tokens]


def filter_stopwords(tokens):
    for token in tokens:
        if token in ['test', 'tests', 'main']:
            return []
    return tokens


def normalize_query(query):
    return query.strip().lower().split()


def query_frequency(query, bow, dictionary):
    return filter_results(get_freq_model(bow, dictionary)[dictionary.doc2bow(query)])


def get_freq_model(bow, dictionary):
    return load_file('model_freq') if exists_file('model_freq') else create_freq_model(bow, dictionary)


def create_freq_model(bow, dictionary):
    model = SparseMatrixSimilarity(bow, num_features=len(dictionary.token2id))
    save_data(model, 'model_freq')
    return model

```

```

def query_tfidf(query, bow, dictionary):
    model = get_tfidf_model(bow)
    matrix = get_tfidf_matrix(model, bow, dictionary)
    return filter_results(matrix[model[dictionary.doc2bow(query)]])


def get_tfidf_model(bow):
    return load_file('model_tfidf') if exists_file('model_tfidf') else create_tfidf_model(bow
        )


def create_tfidf_model(bow):
    model = TfidfModel(bow)
    save_data(model, 'model_tfidf')
    return model


def get_tfidf_matrix(model, bow, dictionary):
    return load_file('matrix_tfidf') if exists_file('matrix_tfidf') else create_tfidf_matrix(
        model, bow, dictionary)


def create_tfidf_matrix(model, bow, dictionary):
    matrix = SparseMatrixSimilarity(model[bow], num_features=len(dictionary.token2id))
    save_data(matrix, 'matrix_tfidf')
    return model


def query_lsi(query, bow, dictionary):
    model = get_lsi_model(bow, dictionary)
    matrix = get_lsi_matrix(model, bow)
    vector = model[dictionary.doc2bow(query)]
    result = abs(matrix[vector])
    embedding = [[value for _, value in vector]] + [[value for _, value in model[bow][i]] for
         i, value in
                                                      sorted(enumerate(result), key=lambda x: x
                                                          [1], reverse=True)[:5]]
    return filter_results(result), embedding


def get_lsi_model(bow, dictionary):
    return load_file('model_lsi') if exists_file('model_lsi') else create_lsi_model(bow,
        dictionary)


def create_lsi_model(bow, dictionary):
    model = LsiModel(bow, id2word=dictionary, num_topics=300)
    save_data(model, 'model_lsi')
    return model


def get_lsi_matrix(model, bow):
    return load_file('matrix_lsi') if exists_file('matrix_lsi') else create_lsi_matrix(model,
         bow)


```

```
def create_lsi_matrix(model, bow):
    matrix = MatrixSimilarity(model[bow])
    save_data(matrix, 'matrix_lsi')
    return matrix


def filter_results(arrg):
    return [i for i, v in sorted(enumerate(arrg), key=lambda x: x[1], reverse=True)[:5]]


def query_doc2vec(query, corpus):
    model = get_doc2vec_model(get_doc2vec_corpus(corpus))
    vector = model.infer_vector(query)
    similar = model.docvecs.most_similar([vector], topn=5)
    return [index for (index, _) in similar], \
           [list(vector)] + [list(model.infer_vector(corpus[index])) for index, _ in similar]


def get_doc2vec_corpus(corpus):
    return [TaggedDocument(simple_preprocess(' '.join(element)), [index])
            for index, element in enumerate(corpus)]


def get_doc2vec_model(corpus):
    return load_file('model_doc2vec') if exists_file('model_doc2vec') else
        create_doc2vec_model(corpus)


def create_doc2vec_model(corpus):
    model = Doc2Vec(vector_size=300, min_count=2, epochs=77)
    model.build_vocab(corpus)
    model.train(corpus, total_examples=model.corpus_count, epochs=model.epochs)
    save_data(model, 'model_doc2vec')
    return model


def create_result_dataframe(queries_dictionary, df):
    for key, values in queries_dictionary.items():
        for index in values:
            row = df.iloc[index]
            yield [row["name"], row["file"], row["line"], row["type"], row["comment"], key]


#####################################

class Truth:
    def __init__(self, query, name, path):
        self.name = name
        self.path = path
        self.query = query.lower()


class Stat:
    def __init__(self, precisions, recalls):
        self.precisions = precisions
        self.recalls = recalls

```

```python

def start(path_ground_truth):
    dataframe = pd.read_csv("res/data.csv").fillna(value="")
    ground_truth, labels = parse_ground_truth(path_ground_truth)
    scores, vectors = compute_precision_recall(ground_truth, dataframe)
    plot_vectors(compute_tsne(vectors), labels)
    print_scores(scores)


def parse_ground_truth(path_ground_truth):
    classes, labels = [], []
    for entry in open(path_ground_truth, "r").read().split("\n\n"):
        data = entry.split("\n")
        query = data[0]
        classes.append(Truth(query, data[1], data[2]))
        labels += [query]*6
    return classes, labels


def compute_precision_recall(ground_truth, dataframe):
    scores = {"FREQ": [], "TF-IDF": [], "LSI": [], "Doc2Vec": []}
    vectors = {"LSI": [], "Doc2Vec": []}
    for entry in ground_truth:
        results, vectors_i = get_results(entry.query, dataframe)
        vectors["LSI"] += vectors_i["LSI"]
        vectors["Doc2Vec"] += vectors_i["Doc2Vec"]
        for query_type in ["FREQ", "TF-IDF", "LSI", "Doc2Vec"]:
            precision = compute_precision(entry, query_type, results)
            scores[query_type].append(Stat(precision, compute_recall(precision)))
    return scores, vectors


def compute_precision(truth, search_type, dataframe):
    counter = 0
    for _, row in dataframe[dataframe['search'] == search_type].iterrows():
        counter += 1
        if row["name"] == truth.name and row["file"] == truth.path:
            return 1 / counter
    return 0


def compute_recall(precision):
    return 1 if precision > 0 else 0


def compute_tsne(dictionary):
    results = {}
    for key, values in dictionary.items():
        tsne = TSNE(n_components=2, verbose=1, perplexity=2, n_iter=3000)
        results[key] = tsne.fit_transform(values)
    return results


def plot_vectors(dictionary, labels):
    for key, values in dictionary.items():
        dataframe = pd.DataFrame()
        dataframe['x'] = values[:, 0]
```

```
        dataframe['y'] = values[:, 1]
        dataframe['labels'] = labels
        plt.figure(figsize=(16, 16))
        plt.title("Results of " + key)

        plot = sns.scatterplot(
            x="x",
            y="y",
            hue='labels',
            data=dataframe,
            legend="full",
            alpha=1.0
        )

        for label in range(0, len(labels), 6):
            plot.text(
                dataframe['x'][label],
                dataframe['y'][label],
                dataframe['labels'][label],
                horizontalalignment='left', size='small', color='gray'
            )

        plot.get_figure().savefig("res/plot_" + key.lower())


def print_scores(scores):
    print("##### PRINT #####")
    for key, values in scores.items():
        print(key)
        precision, recall = compute_mean(values)
        print("\tprecision:\t" + precision + "\n\trecall:\t\t" + recall)


def compute_mean(stats):
    precision, recall, counter = 0, 0, 0
    for stat in stats:
        precision += stat.precisions
        recall += stat.recalls
        counter += 1
    return str(precision / counter), str(recall / counter)


if len(argv) < 1:
    print("Please give as input ground truth file")
    exit(1)


start(argv[1])
```

B Bash Code

```
#!/bin/bash

rm -rf res/pickle
mkdir res/pickle
python3 src/extract_data.py $1
python3 src/search_data.py $2
python3 src/prec_recall.py res/ground-truth-unique.txt
```