MyProject

1

Generated by Doxygen 1.8.11

Contents

1 Namespace Index **1**

1.1 Namespace List 1

2 File Index **3**

2.1 File List 3

3 Namespace Documentation **5**

3.1 create_dict Namespace Reference 5

3.1.1 Variable Documentation 5

3.1.1.1 delimiters 5

3.1.1.2 dict_file 5

3.1.1.3 document_dictionary 5

3.1.1.4 document_word_count 5

3.1.1.5 end 6

3.1.1.6 file 6

3.1.1.7 folder_path 6

3.1.1.8 key 6

3.2 dict_functions Namespace Reference 6

3.2.1 Function Documentation 6

3.2.1.1 create_dictionary(file_path, document_word_count) 6

3.2.1.2 format_text(text_line) 7

3.2.1.3 remove_delimiters(text_line) 7

3.2.1.4 text_formatting(my_dictionary) 7

3.2.2 Variable Documentation 7

3.2.2.1 delimiters . . . 7
3.3 diff Namespace Reference . . . 7
3.3.1 Function Documentation . . . 8
3.3.1.1 calculate_distance(document_one_dictionary, document_two_dictionary) . . . 8
3.3.1.2 ensure_dir(file_path) . . . 8
3.3.1.3 inner_product(document_one_dictionary, document_two_dictionary) . . . 9
3.3.2 Variable Documentation . . . 9
3.3.2.1 delimiters . . . 9
3.3.2.2 dict_file . . . 9
3.3.2.3 document_dictionary . . . 9
3.3.2.4 document_one_dictionary . . . 9
3.3.2.5 document_two_dictionary . . . 9
3.3.2.6 document_word_count . . . 9
3.3.2.7 end . . . 9
3.3.2.8 file . . . 9
3.3.2.9 file_index . . . 10
3.3.2.10 file_index_one . . . 10
3.3.2.11 file_index_two . . . 10
3.3.2.12 folder_path . . . 10
3.3.2.13 idf_key . . . 10
3.3.2.14 output_folder_path . . . 10
3.3.2.15 result . . . 10
3.3.2.16 result_dictionary . . . 10
3.3.2.17 result_file . . . 10
3.3.2.18 result_file_name . . . 10
3.3.2.19 sorted_result . . . 11
3.3.2.20 start . . . 11
3.3.2.21 temp . . . 11
3.3.2.22 temp_dictionary . . . 11
3.3.2.23 text_line . . . 11

3.3.2.24 text_line_divided . . . 11
3.3.2.25 tf_key . . . 11
3.4 result Namespace Reference . . . 11
3.4.1 Function Documentation . . . 12
3.4.1.1 calculate_nearest(file_name) . . . 12
3.4.2 Variable Documentation . . . 12
3.4.2.1 cur_home_dir . . . 12
3.4.2.2 end . . . 12
3.4.2.3 file . . . 12
3.4.2.4 file_name . . . 12
3.4.2.5 folder_path . . . 12
3.4.2.6 key . . . 12
3.4.2.7 res_out_file . . . 12
3.4.2.8 result_dict . . . 12

4 File Documentation **13**

4.1 create_dict.py File Reference . . . 13
4.2 dict_functions.py File Reference . . . 13
4.3 diff.py File Reference . . . 14
4.4 result.py File Reference . . . 14

Index **17**

Chapter 1

Namespace Index

1.1 Namespace List

Here is a list of all namespaces with brief descriptions:

create_dict 5
dict_functions 6
diff 7
result 11

Chapter 2

File Index

2.1 File List

Here is a list of all files with brief descriptions:

create_dict.py 13
dict_functions.py 13
diff.py 14
result.py 14

Chapter 3

Namespace Documentation

3.1 create_dict Namespace Reference

Variables

- string delimiters = "\'\"/<?>,.:*-+\\=‘~!@#^&()_ ;{}[ ]|"
- folder_path = raw_input("Please input the path to the folder\t")
- dictionary document_dictionary = {}
- dictionary document_word_count = {}
- dict_file = open('dict.txt','w')
- key
- end
- file

3.1.1 Variable Documentation

3.1.1.1 string create_dict.delimiters = "\'\"/<?>,.:*-+\\=‘~!@#^&()_ ;{}[ ]|"

Definition at line 18 of file create_dict.py.

3.1.1.2 create_dict.dict_file = open('dict.txt','w')

Definition at line 34 of file create_dict.py.

3.1.1.3 dictionary create_dict.document_dictionary = {}

Definition at line 25 of file create_dict.py.

3.1.1.4 create_dict.document_word_count = {}

Definition at line 26 of file create_dict.py.

3.1.1.5 create_dict.end

Definition at line 36 of file create_dict.py.

3.1.1.6 create_dict.file

Definition at line 36 of file create_dict.py.

3.1.1.7 create_dict.folder_path = raw_input("Please input the path to the folder\t")

Definition at line 24 of file create_dict.py.

3.1.1.8 create_dict.key

Definition at line 36 of file create_dict.py.

3.2 dict_functions Namespace Reference

Functions

- def remove_delimiters (text_line)
- def format_text (text_line)
- def create_dictionary (file_path, document_word_count)
- def text_formatting (my_dictionary)

Variables

- string delimiters = "\'\"/<?>,.:*-+\\=‘~!@#^&()_ ;{}[ ]|"

3.2.1 Function Documentation

3.2.1.1 def dict_functions.create_dictionary (*file_path, document_word_count*)

```
Function name : create_dictionary
Input arguments :
    1. string file_path : The path to the file to be read.
    2. dictionary document_word_count : The dicitonary responsible for training
Purpose : To create the word frequency table/ vector
Return Value : The dictionary containing the word frequenct vector
```

Definition at line 48 of file dict_functions.py.

3.2.1.2 def dict_functions.format_text (*text_line*)

```
Function name : format_text
Input arguments :
    1. string text_line : Piece of text
Purpose : To format the text by removing delimiters and trailing whitespaces
Return Value : The formatted string.
```

Definition at line 35 of file dict_functions.py.

3.2.1.3 def dict_functions.remove_delimiters (*text_line*)

```
Function name : remove_delimiters
Input arguments :
    1. string text_line : Piece of text
Purpose : To remove unnecessary characters
Return Value : The string after removing characters.
```

Definition at line 22 of file dict_functions.py.

3.2.1.4 def dict_functions.text_formatting (*my_dictionary*)

```
Function name : inner_product
Input arguments :
    1. dictionary my_dictionary : The dictionary to be formatted
Purpose : To remove the words which have very less importance to the meaning.
Return Value : The modified dictionary.
```

Definition at line 83 of file dict_functions.py.

3.2.2 Variable Documentation

3.2.2.1 string dict_functions.delimiters = "\'\"/<?>,.:*-+\=‘~!@#^&()_ ;{}[ ]|"

Definition at line 17 of file dict_functions.py.

3.3 diff Namespace Reference

Functions

- def inner_product (document_one_dictionary, document_two_dictionary)
- def calculate_distance (document_one_dictionary, document_two_dictionary)
- def ensure_dir (file_path)

Variables

- string delimiters = "\'\"/<?>,.:∗-+\\=‘∼!@#^&()_ ;{}[ ]|"
- dictionary document_word_count = {}
- start = time.clock()
- folder_path = raw_input("Please input the path to the folder\t")
- dictionary document_dictionary = {}
- output_folder_path = raw_input("Please input the path to the output folder\t")
- dictionary result_dictionary = {}
- dictionary temp = {}
- dict_file = open('dict.txt','r')
- text_line = text_line.rstrip()
- text_line_divided = text_line.split(' ')
- dictionary temp_dictionary = {}
- idf_key = math.log(len(document_dictionary) / (document_word_count[key] ∗ 1.0))
- tf_key = document_dictionary[filename][key]
- file_index = os.listdir(folder_path)
- file_index_one = file_index[index_one]
- file_index_two = file_index[index_two]
- document_one_dictionary = document_dictionary[file_index_one]
- document_two_dictionary = document_dictionary[file_index_two]
- result = calculate_distance(document_one_dictionary, document_two_dictionary)
- end
- sorted_result = sorted(result_dictionary.items(), key=operator.itemgetter(1))
- result_file_name = raw_input("Please enter the name of file where you wish to save the result.\t")
- result_file = open(result_file_name ,'w')
- file

3.3.1 Function Documentation

3.3.1.1 def diff.calculate_distance (*document_one_dictionary, document_two_dictionary*)

```
Function name : inner_product
Input arguments :
    1. dictionary document_one_dictionary: The dictionary corresponding to document one.
    2. dictionary document_two_dictionary: The dictionary corresponding to document two.
Purpose : To find the distance between two documents
Return Value : The cosine distance of two vectors
```

Definition at line 42 of file diff.py.

3.3.1.2 def diff.ensure_dir (*file_path*)

```
Function name : inner_product
Input arguments :
    1. string file_path : The path of the folder.
Purpose : To create a directory if it does not exists.
```

Definition at line 58 of file diff.py.

3.3.1.3 def diff.inner_product (*document_one_dictionary, document_two_dictionary*)

```
Function name : inner_product
Input arguments :
    1. dictionary document_one_dictionary : The dictionary corresponding to document one.
    2. dictionary document_two_dictionary : The dictionary corresponding to document two.
Purpose : To find the inner product of two vectors
Return Value : The inner product of two vectors
```

Definition at line 26 of file diff.py.

3.3.2 Variable Documentation

3.3.2.1 string diff.delimiters = "\'\"/<?>,.:*-+\\=‘∼!@#^&()_ ;{}[ ]|"

Definition at line 18 of file diff.py.

3.3.2.2 diff.dict_file = open('dict.txt','r')

Definition at line 83 of file diff.py.

3.3.2.3 dictionary diff.document_dictionary = {}

Definition at line 72 of file diff.py.

3.3.2.4 diff.document_one_dictionary = document_dictionary[file_index_one]

Definition at line 111 of file diff.py.

3.3.2.5 diff.document_two_dictionary = document_dictionary[file_index_two]

Definition at line 113 of file diff.py.

3.3.2.6 dictionary diff.document_word_count = {}

Definition at line 22 of file diff.py.

3.3.2.7 diff.end

Definition at line 117 of file diff.py.

3.3.2.8 diff.file

Definition at line 130 of file diff.py.

3.3.2.9 diff.file_index = os.listdir(folder_path)

Definition at line 103 of file diff.py.

3.3.2.10 diff.file_index_one = file_index[index_one]

Definition at line 108 of file diff.py.

3.3.2.11 diff.file_index_two = file_index[index_two]

Definition at line 109 of file diff.py.

3.3.2.12 diff.folder_path = raw_input("Please input the path to the folder\t")

Definition at line 71 of file diff.py.

3.3.2.13 diff.idf_key = math.log(len(document_dictionary) / (document_word_count[key] * 1.0))

Definition at line 96 of file diff.py.

3.3.2.14 diff.output_folder_path = raw_input("Please input the path to the output folder\t")

Definition at line 73 of file diff.py.

3.3.2.15 diff.result = calculate_distance(document_one_dictionary, document_two_dictionary)

Definition at line 115 of file diff.py.

3.3.2.16 dictionary diff.result_dictionary = {}

Definition at line 77 of file diff.py.

3.3.2.17 diff.result_file = open(result_file_name ,'w')

Definition at line 128 of file diff.py.

3.3.2.18 diff.result_file_name = raw_input("Please enter the name of file where you wish to save the result.\t")

Definition at line 127 of file diff.py.

3.3.2.19 diff.sorted_result = sorted(result_dictionary.items(), key=operator.itemgetter(1))

Definition at line 123 of file diff.py.

3.3.2.20 diff.start = time.clock()

Definition at line 69 of file diff.py.

3.3.2.21 diff.temp = {}

Definition at line 78 of file diff.py.

3.3.2.22 dictionary diff.temp_dictionary = {}

Definition at line 92 of file diff.py.

3.3.2.23 diff.text_line = text_line.rstrip()

Definition at line 85 of file diff.py.

3.3.2.24 diff.text_line_divided = text_line.split(' ')

Definition at line 86 of file diff.py.

3.3.2.25 diff.tf_key = document_dictionary[filename][key]

Definition at line 97 of file diff.py.

3.4 result Namespace Reference

Functions

- def calculate_nearest (file_name)

Variables

- folder_path = raw_input("Please input the path to the folder\t")
- cur_home_dir = os.getcwd()
- file_name = raw_input("Please input the name of result file.\t")
- result_dict = calculate_nearest(file_name)
- res_out_file = open('eval_res-nearest', 'w')
- key
- end
- file

3.4.1 Function Documentation

3.4.1.1 def result.calculate_nearest (*file_name*)

```
Function name : calculate_nearest
Input arguments :
    1. string file_name : Name of file that contains result
Purpose : To find the nearest file of each and every file
Return Value : The dictionary containing the nearest file index of each file
```

Definition at line 16 of file result.py.

3.4.2 Variable Documentation

3.4.2.1 result.cur_home_dir = os.getcwd()

Definition at line 49 of file result.py.

3.4.2.2 result.end

Definition at line 59 of file result.py.

3.4.2.3 result.file

Definition at line 59 of file result.py.

3.4.2.4 result.file_name = raw_input("Please input the name of result file.\t")

Definition at line 52 of file result.py.

3.4.2.5 result.folder_path = raw_input("Please input the path to the folder\t")

Definition at line 48 of file result.py.

3.4.2.6 result.key

Definition at line 59 of file result.py.

3.4.2.7 result.res_out_file = open('eval_res-nearest', 'w')

Definition at line 57 of file result.py.

3.4.2.8 result.result_dict = calculate_nearest(file_name)

Definition at line 53 of file result.py.

Chapter 4

File Documentation

4.1 create_dict.py File Reference

Namespaces

- create_dict

Variables

- string create_dict.delimiters = "\'\"/<?>,.:*-+\\=‘~!@#^&()_ ;{}[ ]|"
- create_dict.folder_path = raw_input("Please input the path to the folder\t")
- dictionary create_dict.document_dictionary = {}
- dictionary create_dict.document_word_count = {}
- create_dict.dict_file = open('dict.txt','w')
- create_dict.key
- create_dict.end
- create_dict.file

4.2 dict_functions.py File Reference

Namespaces

- dict_functions

Functions

- def dict_functions.remove_delimiters (text_line)
- def dict_functions.format_text (text_line)
- def dict_functions.create_dictionary (file_path, document_word_count)
- def dict_functions.text_formatting (my_dictionary)

Variables

- string dict_functions.delimiters = "\'\"/<?>,.:*-+\\=‘~!@#^&()_ ;{}[ ]|"

4.3 diff.py File Reference

Namespaces

- diff

Functions

- def diff.inner_product (document_one_dictionary, document_two_dictionary)
- def diff.calculate_distance (document_one_dictionary, document_two_dictionary)
- def diff.ensure_dir (file_path)

Variables

- string diff.delimiters = "\'\"/<?>,.:*-+\\=‘~!@#^&()_ ;{}[ ]|"
- dictionary diff.document_word_count = {}
- diff.start = time.clock()
- diff.folder_path = raw_input("Please input the path to the folder\t")
- dictionary diff.document_dictionary = {}
- diff.output_folder_path = raw_input("Please input the path to the output folder\t")
- dictionary diff.result_dictionary = {}
- dictionary diff.temp = {}
- diff.dict_file = open('dict.txt','r')
- diff.text_line = text_line.rstrip()
- diff.text_line_divided = text_line.split(' ')
- dictionary diff.temp_dictionary = {}
- diff.idf_key = math.log(len(document_dictionary) / (document_word_count[key] * 1.0))
- diff.tf_key = document_dictionary[filename][key]
- diff.file_index = os.listdir(folder_path)
- diff.file_index_one = file_index[index_one]
- diff.file_index_two = file_index[index_two]
- diff.document_one_dictionary = document_dictionary[file_index_one]
- diff.document_two_dictionary = document_dictionary[file_index_two]
- diff.result = calculate_distance(document_one_dictionary, document_two_dictionary)
- diff.end
- diff.sorted_result = sorted(result_dictionary.items(), key=operator.itemgetter(1))
- diff.result_file_name = raw_input("Please enter the name of file where you wish to save the result.\t")
- diff.result_file = open(result_file_name ,'w')
- diff.file

4.4 result.py File Reference

Namespaces

- result

Functions

- def result.calculate_nearest (file_name)

Variables

- result.folder_path = raw_input("Please input the path to the folder\t")
- result.cur_home_dir = os.getcwd()
- result.file_name = raw_input("Please input the name of result file.\t")
- result.result_dict = calculate_nearest(file_name)
- result.res_out_file = open('eval_res-nearest', 'w')
- result.key
- result.end
- result.file

Index

calculate_distance
    diff, 8
calculate_nearest
    result, 12
create_dict, 5
    delimiters, 5
    dict_file, 5
    document_dictionary, 5
    document_word_count, 5
    end, 5
    file, 6
    folder_path, 6
    key, 6
create_dict.py, 13
create_dictionary
    dict_functions, 6
cur_home_dir
    result, 12

delimiters
    create_dict, 5
    dict_functions, 7
    diff, 9
dict_file
    create_dict, 5
    diff, 9
dict_functions, 6
    create_dictionary, 6
    delimiters, 7
    format_text, 6
    remove_delimiters, 7
    text_formatting, 7
dict_functions.py, 13
diff, 7
    calculate_distance, 8
    delimiters, 9
    dict_file, 9
    document_dictionary, 9
    document_one_dictionary, 9
    document_two_dictionary, 9
    document_word_count, 9
    end, 9
    ensure_dir, 8
    file, 9
    file_index, 9
    file_index_one, 10
    file_index_two, 10
    folder_path, 10
    idf_key, 10
    inner_product, 8
    output_folder_path, 10
    result, 10
    result_dictionary, 10
    result_file, 10
    result_file_name, 10
    sorted_result, 10
    start, 11
    temp, 11
    temp_dictionary, 11
    text_line, 11
    text_line_divided, 11
    tf_key, 11
diff.py, 14
document_dictionary
    create_dict, 5
    diff, 9
document_one_dictionary
    diff, 9
document_two_dictionary
    diff, 9
document_word_count
    create_dict, 5
    diff, 9

end
    create_dict, 5
    diff, 9
    result, 12
ensure_dir
    diff, 8

file
    create_dict, 6
    diff, 9
    result, 12
file_index
    diff, 9
file_index_one
    diff, 10
file_index_two
    diff, 10
file_name
    result, 12
folder_path
    create_dict, 6
    diff, 10
    result, 12
format_text
    dict_functions, 6

idf_key
    diff, 10
inner_product
    diff, 8

key
    create_dict, 6
    result, 12

output_folder_path
    diff, 10

remove_delimiters
    dict_functions, 7
res_out_file
    result, 12
result, 11
    calculate_nearest, 12
    cur_home_dir, 12
    diff, 10
    end, 12
    file, 12
    file_name, 12
    folder_path, 12
    key, 12
    res_out_file, 12
    result_dict, 12
result.py, 14
result_dict
    result, 12
result_dictionary
    diff, 10
result_file
    diff, 10
result_file_name
    diff, 10

sorted_result
    diff, 10
start
    diff, 11

temp
    diff, 11
temp_dictionary
    diff, 11
text_formatting
    dict_functions, 7
text_line
    diff, 11
text_line_divided
    diff, 11
tf_key
    diff, 11