

# MyProject

1

Generated by Doxygen 1.8.11



# Contents

<b>1</b>	<b>Namespace Index</b>	<b>1</b>
1.1	Namespace List . . . . .	1
<b>2</b>	<b>File Index</b>	<b>3</b>
2.1	File List . . . . .	3
<b>3</b>	<b>Namespace Documentation</b>	<b>5</b>
3.1	create_dict Namespace Reference . . . . .	5
3.1.1	Function Documentation . . . . .	5
3.1.1.1	create_dictionary(file_path) . . . . .	5
3.1.1.2	format_text(text_line) . . . . .	6
3.1.1.3	remove_delimiters(text_line) . . . . .	6
3.1.1.4	text_formatting(my_dictionary) . . . . .	6
3.1.2	Variable Documentation . . . . .	6
3.1.2.1	delimiters . . . . .	6
3.1.2.2	dict_file . . . . .	6
3.1.2.3	document_dictionary . . . . .	6
3.1.2.4	document_word_count . . . . .	6
3.1.2.5	end . . . . .	7
3.1.2.6	file . . . . .	7
3.1.2.7	folder_path . . . . .	7
3.1.2.8	key . . . . .	7
3.2	diff Namespace Reference . . . . .	7
3.2.1	Function Documentation . . . . .	8

3.2.1.1	<code>calculate_distance(document_one_dictionary, document_two_dictionary)</code>	8
3.2.1.2	<code>create_dictionary(file_path)</code>	8
3.2.1.3	<code>ensure_dir(file_path)</code>	8
3.2.1.4	<code>format_text(text_line)</code>	8
3.2.1.5	<code>inner_product(document_one_dictionary, document_two_dictionary)</code>	8
3.2.1.6	<code>remove_delimiters(text_line)</code>	9
3.2.1.7	<code>text_formatting(my_dictionary)</code>	9
3.2.2	Variable Documentation	9
3.2.2.1	<code>delimiters</code>	9
3.2.2.2	<code>dict_file</code>	9
3.2.2.3	<code>document_dictionary</code>	9
3.2.2.4	<code>document_one_dictionary</code>	9
3.2.2.5	<code>document_two_dictionary</code>	9
3.2.2.6	<code>document_word_count</code>	9
3.2.2.7	<code>end</code>	10
3.2.2.8	<code>file</code>	10
3.2.2.9	<code>file_index</code>	10
3.2.2.10	<code>file_index_one</code>	10
3.2.2.11	<code>file_index_two</code>	10
3.2.2.12	<code>folder_path</code>	10
3.2.2.13	<code>idf_key</code>	10
3.2.2.14	<code>output_folder_path</code>	10
3.2.2.15	<code>result</code>	10
3.2.2.16	<code>result_dictionary</code>	10
3.2.2.17	<code>result_file</code>	11
3.2.2.18	<code>result_file_name</code>	11
3.2.2.19	<code>sorted_result</code>	11
3.2.2.20	<code>start</code>	11
3.2.2.21	<code>temp_dictionary</code>	11
3.2.2.22	<code>text_line</code>	11

---

3.2.2.23	<a href="#">text_line_divided</a>	11
3.2.2.24	<a href="#">tf_key</a>	11
3.3	<a href="#">result Namespace Reference</a>	11
3.3.1	<a href="#">Function Documentation</a>	12
3.3.1.1	<a href="#">calculate_nearest(file_name)</a>	12
3.3.2	<a href="#">Variable Documentation</a>	12
3.3.2.1	<a href="#">cur_home_dir</a>	12
3.3.2.2	<a href="#">end</a>	12
3.3.2.3	<a href="#">file</a>	12
3.3.2.4	<a href="#">file_name</a>	12
3.3.2.5	<a href="#">folder_path</a>	12
3.3.2.6	<a href="#">key</a>	13
3.3.2.7	<a href="#">res_out_file</a>	13
3.3.2.8	<a href="#">result_dict</a>	13
4	<a href="#">File Documentation</a>	15
4.1	<a href="#">create_dict.py File Reference</a>	15
4.2	<a href="#">diff.py File Reference</a>	15
4.3	<a href="#">result.py File Reference</a>	16
	<a href="#">Index</a>	17



# Chapter 1

## Namespace Index

### 1.1 Namespace List

Here is a list of all namespaces with brief descriptions:

<a href="#">create_dict</a>	5
<a href="#">diff</a>	7
<a href="#">result</a>	11





## Chapter 2

# File Index

### 2.1 File List

Here is a list of all files with brief descriptions:

<a href="#">create_dict.py</a>	15
<a href="#">diff.py</a>	15
<a href="#">result.py</a>	16



## Chapter 3

# Namespace Documentation

### 3.1 create\_dict Namespace Reference

#### Functions

- def `remove_delimiters` (text\_line)
- def `format_text` (text\_line)
- def `create_dictionary` (file\_path)
- def `text_formatting` (my\_dictionary)

#### Variables

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

#### 3.1.1 Function Documentation

##### 3.1.1.1 def create\_dict.create\_dictionary ( file\_path )

Function name : `create_dictionary`

Input arguments :

1. string `file_path` : The path to the file to be read.

Purpose : To create the word frequency table/ vector

Return Value : The dictionary containing the word frequenct vector

Definition at line 50 of file `create_dict.py`.

**3.1.1.2 def create\_dict.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 37 of file create\_dict.py.

**3.1.1.3 def create\_dict.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 24 of file create\_dict.py.

**3.1.1.4 def create\_dict.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 79 of file create\_dict.py.

**3.1.2 Variable Documentation****3.1.2.1 string create\_dict.delimiters = "\\\"/ < ? > , : \* - + \\ = ' ~ ! @ # ^ & ( ) \_ ; { [ ] | "**

Definition at line 17 of file create\_dict.py.

**3.1.2.2 create\_dict.dict\_file = open('dict.txt','w')**

Definition at line 107 of file create\_dict.py.

**3.1.2.3 dictionary create\_dict.document\_dictionary = {}**

Definition at line 101 of file create\_dict.py.

**3.1.2.4 dictionary create\_dict.document\_word\_count = {}**

Definition at line 21 of file create\_dict.py.

## 3.1.2.5 create\_dict.end

Definition at line 109 of file create\_dict.py.

## 3.1.2.6 create\_dict.file

Definition at line 109 of file create\_dict.py.

## 3.1.2.7 create\_dict.folder\_path = raw\_input("Please input the path to the folder\t")

Definition at line 100 of file create\_dict.py.

## 3.1.2.8 create\_dict.key

Definition at line 109 of file create\_dict.py.

## 3.2 diff Namespace Reference

### Functions

- def [remove\\_delimiters](#) (text\_line)
- def [format\\_text](#) (text\_line)
- def [create\\_dictionary](#) (file\_path)
- def [inner\\_product](#) (document\_one\_dictionary, document\_two\_dictionary)
- def [calculate\\_distance](#) (document\_one\_dictionary, document\_two\_dictionary)
- def [text\\_formatting](#) (my\_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](#) = {}
- [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](#)

### 3.2.1 Function Documentation

#### 3.2.1.1 `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 94 of file `diff.py`.

#### 3.2.1.2 `def diff.create_dictionary ( file_path )`

Function name : `create_dictionary`

Input arguments :

1. string `file_path` : The path to the file to be read.

Purpose : To create the word frequency table/ vector

Return Value : The dictionary containing the word frequenct vector

Definition at line 50 of file `diff.py`.

#### 3.2.1.3 `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 126 of file `diff.py`.

#### 3.2.1.4 `def diff.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 37 of file `diff.py`.

#### 3.2.1.5 `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 78 of file `diff.py`.

### 3.2.1.6 def diff.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 24 of file diff.py.

### 3.2.1.7 def diff.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 109 of file diff.py.

## 3.2.2 Variable Documentation

### 3.2.2.1 string diff.delimiters = "\"'/?>.,:~!@#^&()\_ ;{}|\""

Definition at line 17 of file diff.py.

### 3.2.2.2 diff.dict\_file = open('dict.txt','r')

Definition at line 150 of file diff.py.

### 3.2.2.3 dictionary diff.document\_dictionary = {}

Definition at line 140 of file diff.py.

### 3.2.2.4 diff.document\_one\_dictionary = document\_dictionary[file\_index\_one]

Definition at line 178 of file diff.py.

### 3.2.2.5 diff.document\_two\_dictionary = document\_dictionary[file\_index\_two]

Definition at line 180 of file diff.py.

### 3.2.2.6 dictionary diff.document\_word\_count = {}

Definition at line 21 of file diff.py.

#### 3.2.2.7 `diff.end`

Definition at line 184 of file `diff.py`.

#### 3.2.2.8 `diff.file`

Definition at line 197 of file `diff.py`.

#### 3.2.2.9 `diff.file_index = os.listdir(folder_path)`

Definition at line 170 of file `diff.py`.

#### 3.2.2.10 `diff.file_index_one = file_index[index_one]`

Definition at line 175 of file `diff.py`.

#### 3.2.2.11 `diff.file_index_two = file_index[index_two]`

Definition at line 176 of file `diff.py`.

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

Definition at line 139 of file `diff.py`.

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

Definition at line 163 of file `diff.py`.

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

Definition at line 141 of file `diff.py`.

#### 3.2.2.15 `diff.result = calculate_distance(document_one_dictionary, document_two_dictionary)`

Definition at line 182 of file `diff.py`.

#### 3.2.2.16 `dictionary diff.result_dictionary = {}`

Definition at line 145 of file `diff.py`.



3.2.2.17 `diff.result_file = open(result_file_name, 'w')`

Definition at line 195 of file diff.py.

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

Definition at line 194 of file diff.py.

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

Definition at line 190 of file diff.py.

3.2.2.20 `diff.start = time.clock()`

Definition at line 137 of file diff.py.

3.2.2.21 `dictionary diff.temp_dictionary = {}`

Definition at line 159 of file diff.py.

3.2.2.22 `diff.text_line = text_line.rstrip()`

Definition at line 152 of file diff.py.

3.2.2.23 `diff.text_line_divided = text_line.split(' ')`

Definition at line 153 of file diff.py.

3.2.2.24 `diff.tf_key = document_dictionary[filename][key]`

Definition at line 164 of file diff.py.

## 3.3 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`

### 3.3.1 Function Documentation

#### 3.3.1.1 `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`.

### 3.3.2 Variable Documentation

#### 3.3.2.1 `result.cur_home_dir = os.getcwd()`

Definition at line 49 of file `result.py`.

#### 3.3.2.2 `result.end`

Definition at line 59 of file `result.py`.

#### 3.3.2.3 `result.file`

Definition at line 59 of file `result.py`.

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

Definition at line 52 of file `result.py`.

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

Definition at line 48 of file `result.py`.

#### 3.3.2.6 result.key

Definition at line 59 of file result.py.

#### 3.3.2.7 result.res\_out\_file = open('eval\_res-nearest', 'w')

Definition at line 57 of file result.py.

#### 3.3.2.8 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](#)

#### Functions

- def [create\\_dict.remove\\_delimiters](#) (text\_line)
- def [create\\_dict.format\\_text](#) (text\_line)
- def [create\\_dict.create\\_dictionary](#) (file\_path)
- def [create\\_dict.text\\_formatting](#) (my\_dictionary)

#### Variables

- string [create\\_dict.delimiters](#) = "\\\"/<?>,:\*~!@#^&()\_ ;{}|]"
- dictionary [create\\_dict.document\\_word\\_count](#) = {}
- [create\\_dict.folder\\_path](#) = raw\_input("Please input the path to the folder\t")
- dictionary [create\\_dict.document\\_dictionary](#) = {}
- [create\\_dict.dict\\_file](#) = open('dict.txt','w')
- [create\\_dict.key](#)
- [create\\_dict.end](#)
- [create\\_dict.file](#)

### 4.2 diff.py File Reference

#### Namespaces

- [diff](#)

## Functions

- def `diff.remove_delimiters` (text\_line)
- def `diff.format_text` (text\_line)
- def `diff.create_dictionary` (file\_path)
- def `diff.inner_product` (document\_one\_dictionary, document\_two\_dictionary)
- def `diff.calculate_distance` (document\_one\_dictionary, document\_two\_dictionary)
- def `diff.text_formatting` (my\_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` = {}
- `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`

## 4.3 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
  - create\_dictionary, 5
  - delimiters, 6
  - dict\_file, 6
  - document\_dictionary, 6
  - document\_word\_count, 6
  - end, 6
  - file, 7
  - folder\_path, 7
  - format\_text, 5
  - key, 7
  - remove\_delimiters, 6
  - text\_formatting, 6
- create\_dict.py, 15
- create\_dictionary
  - create\_dict, 5
  - diff, 8
- cur\_home\_dir
  - result, 12
- delimiters
  - create\_dict, 6
  - diff, 9
- dict\_file
  - create\_dict, 6
  - diff, 9
- diff, 7
  - calculate\_distance, 8
  - create\_dictionary, 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, 10
  - file\_index, 10
  - file\_index\_one, 10
  - file\_index\_two, 10
  - folder\_path, 10
  - format\_text, 8
  - idf\_key, 10
  - inner\_product, 8
  - output\_folder\_path, 10
  - remove\_delimiters, 8
  - result, 10
  - result\_dictionary, 10
  - result\_file, 10
  - result\_file\_name, 11
  - sorted\_result, 11
  - start, 11
  - temp\_dictionary, 11
  - text\_formatting, 9
  - text\_line, 11
  - text\_line\_divided, 11
  - tf\_key, 11
- diff.py, 15
- document\_dictionary
  - create\_dict, 6
  - diff, 9
- document\_one\_dictionary
  - diff, 9
- document\_two\_dictionary
  - diff, 9
- document\_word\_count
  - create\_dict, 6
  - diff, 9
- end
  - create\_dict, 6
  - diff, 9
  - result, 12
- ensure\_dir
  - diff, 8
- file
  - create\_dict, 7
  - diff, 10
  - result, 12
- file\_index
  - diff, 10
- file\_index\_one
  - diff, 10
- file\_index\_two
  - diff, 10
- file\_name
  - result, 12
- folder\_path
  - create\_dict, 7
  - diff, 10
  - result, 12
- format\_text
  - create\_dict, 5
  - diff, 8

- idf\_key
  - diff, [10](#)
- inner\_product
  - diff, [8](#)
- key
  - create\_dict, [7](#)
  - result, [12](#)
- output\_folder\_path
  - diff, [10](#)
- remove\_delimiters
  - create\_dict, [6](#)
  - diff, [8](#)
- res\_out\_file
  - result, [13](#)
- 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, [13](#)
  - result\_dict, [13](#)
- result.py, [16](#)
- result\_dict
  - result, [13](#)
- result\_dictionary
  - diff, [10](#)
- result\_file
  - diff, [10](#)
- result\_file\_name
  - diff, [11](#)
- sorted\_result
  - diff, [11](#)
- start
  - diff, [11](#)
- temp\_dictionary
  - diff, [11](#)
- text\_formatting
  - create\_dict, [6](#)
  - diff, [9](#)
- text\_line
  - diff, [11](#)
- text\_line\_divided
  - diff, [11](#)
- tf\_key
  - diff, [11](#)