Help on module Co\_Occurrence:

NAME

Co\_Occurrence

CLASSES

builtins.object

CoOccur

class CoOccur(builtins.object)

| CoOccur(dataframe)

|

| This class contains methods for updating the document frequency and

| tokenizing the article.

|

| Methods defined here:

|

| \_\_init\_\_(self, dataframe)

| This is the constructor for this class which initialises the instance

| variable.

| :param dataframe: This is a pandas dataframe storing every news article.

|

| find(self)

| This method updates the document frequency by iterating through each

| document in the dataframe.

| :return: It returns a list which contains all unique words across all

| documents(document corpus).

|

| spell\_check(self, query)

| This method checks for spelling errors in the query, and updates any

| misspelt word.It uses the python library 'pyspellchecker'

| :param query: query is a list of words which may contain misspelt words.

| :return: It returns a list which contains correct words in query

| as well as the incorrect words which have now been modified.

|

| tokenize(self, article)

| This method tokenises a string into important words.It first case folds

| the string, creates tokens, removes stop words and then stems word using

| Porter Stemmer.

| :param article: It is a news article in form of string.

| :return: It returns a list of tokenized words.

|

| ----------------------------------------------------------------------

| Data descriptors defined here:

|

| \_\_dict\_\_

| dictionary for instance variables (if defined)

|

| \_\_weakref\_\_

| list of weak references to the object (if defined)

|

| ----------------------------------------------------------------------

| Data and other attributes defined here:

|

| df = Empty DataFrame

| Columns: []

| Index: []

DATA

docFreq = {}

ps = <PorterStemmer>

stop\_words = {'!', '"', "'", '(', ')', ',', ...}

stopwords = <WordListCorpusReader in 'C:\\Users\\Smit\\AppData\\Roamin...

FILE

c:\users\smit\pycharmprojects\searchengine\co\_occurrence.py