app.Cluster

index
d:\projects\summer-project-msc\pythonserver\app\cluster.py

Modules

jellyfish numpy pandas

Classes

abc.ABC(builtins.object)

Cluster

StringCluster(Cluster, abc.ABC)

LinkageBasedStringCluster

class **Cluster**(abc.ABC)

```
Cluster(dataList, targetNumberOfCluster)

Abstract base class representing a cluster of data points.

Attributes:
    dataList (list): A list of data points with the same format.
    targetNumberOfCluster (int): The desired number of clusters.

Behaviors:
    __init__: Validates the datalist and numberofcluster, construct the object from args
    _validateDataList: Validates the data, the data is valid if all the data points share the same format
    _validateNumberOfCluster: Validates the number of cluster, the number is valid if 1<=number<=len(data)
    getClusterId: Returns a list of cluster IDs corresponding to the data in the dataList.
```

Method resolution order:

Cluster
abc.ABC
builtins.object

Methods defined here:

__init__(self, dataList, targetNumberOfCluster)

```
Initializes the Cluster instance.

Parameters:
    dataList (list): A list of data points.
    targetNumberOfCluster (int): The desired number of clusters.

Raises:
    ValueError: If the data list or number of clusters is not valid.
```

getClusterIdList(self) -> list

```
Returns a list of cluster IDs corresponding to the data in the dataList.

Returns:
    list: A list of cluster IDs.
```

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:

__abstractmethods__ = frozenset({'_validateDataList', '_validateNumberOfCluster', 'getClusterIdList'})

class **LinkageBasedStringCluster**(StringCluster)

```
LinkageBasedStringCluster(dataList: list[str], targetNumberOfCluster: int, distanceMetric: str, linkageMethod: str, stringPreprocessor: Callable[[str], str], testMode=False)

Cluster that can using LinkageBased Clustering Algorithm to cluster a list of string
WARNING: targetNumberOfCluster doesn't work, ignore it
Attributes:
    dataList (list): A list of string

    targetNumberOfCluster (int): The desired number of clusters, should between 1 and the number of unique preprocessed string processed by stringPreprocessor from the dataList

    distanceMetric (str): distanceMetric for different strings, used for generating distance matrix.
        it should be one of 'levenshtein' or 'damerauLevenshtein' or 'hamming' or 'jaroSimilarity' or 'jaroWinklerSimilarity' or 'MatchRatingApproach'

    linkageMethod (str): the linkage algorithm to use. see: https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html
        it should be one of 'average' or 'single' or 'complete' or 'weigthed' or 'centroid' or 'median' or 'ward'

    stringPreprocessor (function:str->str): a function to preprocess the string, whose input and output is string

Behaviors:
    __init__: Validates the args, construct the object from args or raise error.
    _validateDataList(private): Validates the data
    _validateNumberOfCluster(private): Validates the number of cluster
    _validateDistanceMetric(private): Validates the distanceMetric
    _validateLinkageMethod(private): Validates linkageMethods
    _validateStringPreprocessor(private): Validates preprocessor
    setNumberOfCluster (targetNumberOfCluster): set the number of cluster if the number is valid.
    getClusterId: Returns a list of cluster IDs corresponding to the data in the dataList.

# reference: Algorithm to Cluster Similar Strings in Python | Saturn Cloud Blog. (2023, July 18). https://saturncloud.io/blog/algorithm-to-cluster-similar-strings-in-python/
```

Method resolution order:
LinkageBasedStringCluster
StringCluster
Cluster
abc.ABC
builtins.object

Methods defined here:

__init__(self, dataList: list[str], targetNumberOfCluster: int, distanceMetric: str, linkageMethod: str, stringPreprocessor: Callable[[str], str], testMode=False)

```
Initialise the object
dataList (list): A list of string

targetNumberOfCluster (int): The desired number of clusters, should between 1 and the number of unique preprocessed string processed by stringPreprocessor from the dataList

distanceMetric (str): distanceMetric for different strings, used for generating distance matrix.
    it should be one of 'levenshtein' or 'damerauLevenshtein' or 'hamming' or 'jaroSimilarity' or 'jaroWinklerSimilarity' or 'MatchRatingApproach'

linkageMethod (str): the linkage algorithm to use. see: https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html
```

```
          it should be one of 'average' or 'single' or 'complete' or 'weigthed' or 'centroid' or 'median' or 'ward'

     stringPreprocessor (function:str->str): a function to preprocess the string, whose input and output is string
```

__str__(self)
```
     Return str(self).
```

getClusterIdList(self, targetNumberOfCluster: int) -> list[int]
```
     Raise ValueError if targetNumberOfCluster is invalid (not implemented)
     Returns a list of cluster IDs corresponding to the data in dataList, distanceMetrics, linkageMethod and Preprocessor
```

getClusterInfo(self)

getDataList(self) -> list[str]
```
     return a list of string which is aligned to the cluster id list
```

getDistanceMatrix(self)

getLinkageMatrix(self)

getPreprocessedData(self)
```
     # getters
```

setDataList(self, dataList: list[str], updateChainning: bool = True)
```
     Validate the datalist, update the datalist,
     UpdateChainning (bool): If True, update the preprocessedStringArray, update the distanceMatrix, update the linkageMatrix
```

setDistanceMetric(self, distanceMetric: str, updateChainning: bool = True)
```
     Validate the distanceMetric, update the distanceMatrix
     updateChainning (bool): If True, update the linkageMatrix
```

setLinkageMethod(self, linkageMethod: str)
```
     Validate the linkageMethod, update the linkageMatrix
```

setStringPreprocessor(self, stringPreprocessor: Callable[[str], str], updateChainning: bool = True)
```
     Validate the StringPreprocessor, update the preprocessedStringArray
     updateChainning (bool): If True, update the distanceMatrix, update the linkageMatrix
```

Data and other attributes defined here:

VALID_DISTANCE_METRIC = ['levenshtein', 'damerauLevenshtein', 'hamming', 'jaroSimilarity', 'jaroWinklerSimilarity', 'MatchRatingApproach']

VALID_LINKAGE_METHOD = ['average', 'single', 'complete', 'weighted', 'centroid', 'median', 'ward']

__abstractmethods__ = frozenset()

Data descriptors inherited from Cluster:

__dict__
```
     dictionary for instance variables (if defined)
```

__weakref__
```
     list of weak references to the object (if defined)
```

class **StringCluster**(Cluster, abc.ABC)
```
   StringCluster(dataList, targetNumberOfCluster)

   Abstract base class representing a cluster of string

   Attributes:
```

```
    dataList (list): A list of string
    targetNumberOfCluster (int): The desired number of clusters.
    stringPreprocessor (function:str->str): a function to preprocess the string, whose input and output is string

Behaviors:
    __init__: Validates the datalist and numberofcluster, construct the object from args
    _validateDataList: Validates the data, the data is valid if all the data points string
    _validateNumberOfCluster: Validates the number of cluster, the number is valid if 1<=number<=len(set(data))
    getClusterId: Returns a list of cluster IDs corresponding to the data in the dataList.
```

Method resolution order:

StringCluster
Cluster
abc.ABC
builtins.object

Data and other attributes defined here:

__abstractmethods__ = frozenset({'_validateDataList', '_validateNumberOfCluster', 'getClusterIdList'})

Methods inherited from Cluster:

__init__(self, dataList, targetNumberOfCluster)

```
    Initializes the Cluster instance.

    Parameters:
        dataList (list): A list of data points.
        targetNumberOfCluster (int): The desired number of clusters.

    Raises:
        ValueError: If the data list or number of clusters is not valid.
```

getClusterIdList(self) -> list

```
    Returns a list of cluster IDs corresponding to the data in the dataList.

    Returns:
        list: A list of cluster IDs.
```

Data descriptors inherited from Cluster:

__dict__

```
    dictionary for instance variables (if defined)
```

__weakref__

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

Data

Callable = typing.Callable
Union = typing.Union