Preprocessing Data

1. Collect data

In [2]:

```
!git clone https://github.com/wchill/HMP_Dataset.git
```

```
Cloning into 'HMP_Dataset'...
remote: Enumerating objects: 865, done.
remote: Total 865 (delta 0), reused 0 (delta 0), pack-reused 865
Receiving objects: 100% (865/865), 1010.96 KiB | 4.14 MiB/s, done.
Updating files: 100% (848/848), done.
```

In [23]:

```
! ls HMP_Dataset
```

```
Brush_teeth     Drink_glass  Getup_bed    Pour_water     Use_telephone
Climb_stairs    Eat_meat     impdata.py   README.txt     Walk
Comb_hair       Eat_soup     Liedown_bed  Sitdown_chair
Descend_stairs  final.py     MANUAL.txt   Standup_chair
```

In [24]:

```
! ls HMP_Dataset/Brush_teeth
```

```
Accelerometer-2011-04-11-13-28-18-brush_teeth-f1.txt
Accelerometer-2011-04-11-13-29-54-brush_teeth-f1.txt
Accelerometer-2011-05-30-08-35-11-brush_teeth-f1.txt
Accelerometer-2011-05-30-09-36-50-brush_teeth-f1.txt
Accelerometer-2011-05-30-10-34-16-brush_teeth-m1.txt
Accelerometer-2011-05-30-21-10-57-brush_teeth-f1.txt
Accelerometer-2011-05-30-21-55-04-brush_teeth-m2.txt
Accelerometer-2011-05-31-15-16-47-brush_teeth-f1.txt
Accelerometer-2011-06-02-10-42-22-brush_teeth-f1.txt
Accelerometer-2011-06-02-10-45-50-brush_teeth-f1.txt
Accelerometer-2011-06-06-10-45-27-brush_teeth-f1.txt
Accelerometer-2011-06-06-10-48-05-brush_teeth-f1.txt
```

In [25]:

```
from pyspark.sql.types import StructType, StructField, IntegerType

schema = StructType([
    StructField('x', IntegerType(), True),
    StructField('y', IntegerType(), True),
    StructField('z', IntegerType(), True)
])
```

In [19]:

```
import os
```

In [30]:

```
file_list = os.listdir('HMP_Dataset')
```

In [31]:

```
file_list
```

Out[31]:

```
['.git',
 '.idea',
 'Brush_teeth',
 'Climb_stairs',
 'Comb_hair',
 'Descend_stairs',
 'Drink_glass',
 'Eat_meat',
 'Eat_soup',
 'Getup_bed',
 'Liedown_bed',
 'MANUAL.txt',
 'Pour_water',
 'README.txt',
 'Sitdown_chair',
 'Standup_chair',
 'Use_telephone',
 'Walk',
 'final.py',
 'impdata.py']
```

In [32]:

```
file_list_filtered = [s for s in file_list if '_' in s]
```

In [33]:

```
file_list_filtered
```

Out[33]:

```
['Brush_teeth',
 'Climb_stairs',
 'Comb_hair',
 'Descend_stairs',
 'Drink_glass',
 'Eat_meat',
 'Eat_soup',
 'Getup_bed',
 'Liedown_bed',
 'Pour_water',
 'Sitdown_chair',
 'Standup_chair',
 'Use_telephone']
```

2. Make Dataframe

In [34]:

```
df = None

from pyspark.sql.functions import lit

for category in file_list_filtered:
    data_files = os.listdir('HMP_Dataset/' + category)

    for data_file in data_files:
        temp_df = spark.read.option('header', 'false').option('delimiter', ' ').csv('HM
P_Dataset/' + category + '/' + data_file, schema = schema)
        temp_df = temp_df.withColumn('class', lit(category))
        temp_df = temp_df.withColumn('source', lit(data_file))

        if df is None:
            df = temp_df
        else:
            df = df.union(temp_df)
```

In [35]:

```
df.show()
```

```
+---+---+---+-----------+--------------------+
|  x|  y|  z|      class|              source|
+---+---+---+-----------+--------------------+
| 22| 49| 35|Brush_teeth|Accelerometer-201...|
| 22| 49| 35|Brush_teeth|Accelerometer-201...|
| 22| 52| 35|Brush_teeth|Accelerometer-201...|
| 22| 52| 35|Brush_teeth|Accelerometer-201...|
| 21| 52| 34|Brush_teeth|Accelerometer-201...|
| 22| 51| 34|Brush_teeth|Accelerometer-201...|
| 20| 50| 35|Brush_teeth|Accelerometer-201...|
| 22| 52| 34|Brush_teeth|Accelerometer-201...|
| 22| 50| 34|Brush_teeth|Accelerometer-201...|
| 22| 51| 35|Brush_teeth|Accelerometer-201...|
| 21| 51| 33|Brush_teeth|Accelerometer-201...|
| 20| 50| 34|Brush_teeth|Accelerometer-201...|
| 21| 49| 33|Brush_teeth|Accelerometer-201...|
| 21| 49| 33|Brush_teeth|Accelerometer-201...|
| 20| 51| 35|Brush_teeth|Accelerometer-201...|
| 18| 49| 34|Brush_teeth|Accelerometer-201...|
| 19| 48| 34|Brush_teeth|Accelerometer-201...|
| 16| 53| 34|Brush_teeth|Accelerometer-201...|
| 18| 52| 35|Brush_teeth|Accelerometer-201...|
| 18| 51| 32|Brush_teeth|Accelerometer-201...|
+---+---+---+-----------+--------------------+
only showing top 20 rows
```

3. Pipeline 1 transform - indexer

In [36]:

```
#transform data
from pyspark.ml.feature import StringIndexer

indexer = StringIndexer(inputCol = 'class', outputCol = 'classIndex')
indexed = indexer.fit(df).transform(df)
indexed.show()
```

```
+---+---+---+-----------+--------------------+----------+
|  x|  y|  z|      class|              source|classIndex|
+---+---+---+-----------+--------------------+----------+
| 22| 49| 35|Brush_teeth|Accelerometer-201...|       5.0|
| 22| 49| 35|Brush_teeth|Accelerometer-201...|       5.0|
| 22| 52| 35|Brush_teeth|Accelerometer-201...|       5.0|
| 22| 52| 35|Brush_teeth|Accelerometer-201...|       5.0|
| 21| 52| 34|Brush_teeth|Accelerometer-201...|       5.0|
| 22| 51| 34|Brush_teeth|Accelerometer-201...|       5.0|
| 20| 50| 35|Brush_teeth|Accelerometer-201...|       5.0|
| 22| 52| 34|Brush_teeth|Accelerometer-201...|       5.0|
| 22| 50| 34|Brush_teeth|Accelerometer-201...|       5.0|
| 22| 51| 35|Brush_teeth|Accelerometer-201...|       5.0|
| 21| 51| 33|Brush_teeth|Accelerometer-201...|       5.0|
| 20| 50| 34|Brush_teeth|Accelerometer-201...|       5.0|
| 21| 49| 33|Brush_teeth|Accelerometer-201...|       5.0|
| 21| 49| 33|Brush_teeth|Accelerometer-201...|       5.0|
| 20| 51| 35|Brush_teeth|Accelerometer-201...|       5.0|
| 18| 49| 34|Brush_teeth|Accelerometer-201...|       5.0|
| 19| 48| 34|Brush_teeth|Accelerometer-201...|       5.0|
| 16| 53| 34|Brush_teeth|Accelerometer-201...|       5.0|
| 18| 52| 35|Brush_teeth|Accelerometer-201...|       5.0|
| 18| 51| 32|Brush_teeth|Accelerometer-201...|       5.0|
+---+---+---+-----------+--------------------+----------+
only showing top 20 rows
```

4. Pipeline 2 transform - Encoder

In [37]:

```
from pyspark.ml.feature import OneHotEncoder
encoder = OneHotEncoder(inputCol = 'classIndex', outputCol = 'categoryVec')
encoded = encoder.transform(indexed)
```

In [38]:

```
encoded.show()
```

```
+---+---+---+----------+--------------------+----------+--------------+
|  x|  y|  z|     class|              source|classIndex|   categoryVec|
+---+---+---+----------+--------------------+----------+--------------+
| 22| 49| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
| 22| 49| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
| 22| 52| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
| 22| 52| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
| 21| 52| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
| 22| 51| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
| 20| 50| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
| 22| 52| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
| 22| 50| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
| 22| 51| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
| 21| 51| 33|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
| 20| 50| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
| 21| 49| 33|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
| 21| 49| 33|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
| 20| 51| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
| 18| 49| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
| 19| 48| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
| 16| 53| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
| 18| 52| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
| 18| 51| 32|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|
+---+---+---+----------+--------------------+----------+--------------+
only showing top 20 rows
```

5. Pipeline 3 transform - vectorAssembler

In [40]:

```
from pyspark.ml.linalg import Vectors
from pyspark.ml.feature import VectorAssembler

vectorAssembler = VectorAssembler(inputCols = ['x', 'y', 'z'],
                                  outputCol = 'features')

features_vectorized = vectorAssembler.transform(encoded)
```

In [41]:

```
features_vectorized.show()
```

```
+---+---+---+----------+-------------------+---------+-------------+--
--------------+
|  x|  y|  z|     class|             source|classIndex|  categoryVec|
features|
+---+---+---+----------+-------------------+---------+-------------+--
--------------+
| 22| 49| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,49.0,35.0]|
| 22| 49| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,49.0,35.0]|
| 22| 52| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,52.0,35.0]|
| 22| 52| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,52.0,35.0]|
| 21| 52| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
1.0,52.0,34.0]|
| 22| 51| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,51.0,34.0]|
| 20| 50| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
0.0,50.0,35.0]|
| 22| 52| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,52.0,34.0]|
| 22| 50| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,50.0,34.0]|
| 22| 51| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,51.0,35.0]|
| 21| 51| 33|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
1.0,51.0,33.0]|
| 20| 50| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
0.0,50.0,34.0]|
| 21| 49| 33|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
1.0,49.0,33.0]|
| 21| 49| 33|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
1.0,49.0,33.0]|
| 20| 51| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
0.0,51.0,35.0]|
| 18| 49| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[1
8.0,49.0,34.0]|
| 19| 48| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[1
9.0,48.0,34.0]|
| 16| 53| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[1
6.0,53.0,34.0]|
| 18| 52| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[1
8.0,52.0,35.0]|
| 18| 51| 32|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[1
8.0,51.0,32.0]|
+---+---+---+----------+-------------------+---------+-------------+--
--------------+
only showing top 20 rows
```

6. Pipeline 4 transform - Normalizer

In [42]:

```
from pyspark.ml.feature import Normalizer

normalizer = Normalizer(inputCol = 'features', outputCol = 'features_norm', p = 1.0)
normalized_data = normalizer.transform(features_vectorized)
normalized_data.show()
```

```
+---+---+---+----------+--------------------+---------+-------------+--
--------------+--------------------+
|  x|  y|  z|     class|              source|classIndex|   categoryVec|
features|       features_norm|
+---+---+---+----------+--------------------+---------+-------------+--
--------------+--------------------+
| 22| 49| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,49.0,35.0]|[0.20754716981132...|
| 22| 49| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,49.0,35.0]|[0.20754716981132...|
| 22| 52| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,52.0,35.0]|[0.20183486238532...|
| 22| 52| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,52.0,35.0]|[0.20183486238532...|
| 21| 52| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
1.0,52.0,34.0]|[0.19626168224299...|
| 22| 51| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,51.0,34.0]|[0.20560747663551...|
| 20| 50| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
0.0,50.0,35.0]|[0.19047619047619...|
| 22| 52| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,52.0,34.0]|[0.20370370370370...|
| 22| 50| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,50.0,34.0]|[0.20754716981132...|
| 22| 51| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,51.0,35.0]|[0.20370370370370...|
| 21| 51| 33|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
1.0,51.0,33.0]|[0.2,0.4857142857...|
| 20| 50| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
0.0,50.0,34.0]|[0.19230769230769...|
| 21| 49| 33|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
1.0,49.0,33.0]|[0.20388349514563...|
| 21| 49| 33|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
1.0,49.0,33.0]|[0.20388349514563...|
| 20| 51| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
0.0,51.0,35.0]|[0.18867924528301...|
| 18| 49| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[1
8.0,49.0,34.0]|[0.17821782178217...|
| 19| 48| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[1
9.0,48.0,34.0]|[0.18811881188118...|
| 16| 53| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[1
6.0,53.0,34.0]|[0.15533980582524...|
| 18| 52| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[1
8.0,52.0,35.0]|[0.17142857142857...|
| 18| 51| 32|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[1
8.0,51.0,32.0]|[0.17821782178217...|
+---+---+---+----------+--------------------+---------+-------------+--
--------------+--------------------+
only showing top 20 rows
```

7. Stack pipeline

In [43]:

```
from pyspark.ml import Pipeline

pipeline = Pipeline(stages = [indexer, encoder, vectorAssembler, normalizer])

model = pipeline.fit(df)

prediction = model.transform(df)

prediction.show()
```

```
+---+---+---+----------+--------------------+----------+-------------+--
--------------+--------------------+
|  x|  y|  z|     class|              source|classIndex|  categoryVec|
features|       features_norm|
+---+---+---+----------+--------------------+----------+-------------+--
--------------+--------------------+
| 22| 49| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,49.0,35.0]|[0.20754716981132...|
| 22| 49| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,49.0,35.0]|[0.20754716981132...|
| 22| 52| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,52.0,35.0]|[0.20183486238532...|
| 22| 52| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,52.0,35.0]|[0.20183486238532...|
| 21| 52| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
1.0,52.0,34.0]|[0.19626168224299...|
| 22| 51| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,51.0,34.0]|[0.20560747663551...|
| 20| 50| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
0.0,50.0,35.0]|[0.19047619047619...|
| 22| 52| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,52.0,34.0]|[0.20370370370370...|
| 22| 50| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,50.0,34.0]|[0.20754716981132...|
| 22| 51| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
2.0,51.0,35.0]|[0.20370370370370...|
| 21| 51| 33|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
1.0,51.0,33.0]|[0.2,0.4857142857...|
| 20| 50| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
0.0,50.0,34.0]|[0.19230769230769...|
| 21| 49| 33|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
1.0,49.0,33.0]|[0.20388349514563...|
| 21| 49| 33|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
1.0,49.0,33.0]|[0.20388349514563...|
| 20| 51| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[2
0.0,51.0,35.0]|[0.18867924528301...|
| 18| 49| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[1
8.0,49.0,34.0]|[0.17821782178217...|
| 19| 48| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[1
9.0,48.0,34.0]|[0.18811881188118...|
| 16| 53| 34|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[1
6.0,53.0,34.0]|[0.15533980582524...|
| 18| 52| 35|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[1
8.0,52.0,35.0]|[0.17142857142857...|
| 18| 51| 32|Brush_teeth|Accelerometer-201...|       5.0|(12,[5],[1.0])|[1
8.0,51.0,32.0]|[0.17821782178217...|
+---+---+---+----------+--------------------+----------+-------------+--
--------------+--------------------+
only showing top 20 rows
```

In [45]:

```
#make structure ==> features | classes
df_train = prediction.drop('x').drop('y').drop('z').drop('class').drop('source').drop(
'features')
```

In [49]:

```
df_train = df_train.drop('classIndex')
```

In [50]:

```
df_train.show()
```

```
+--------------+--------------------+
|   categoryVec|       features_norm|
+--------------+--------------------+
|(12,[5],[1.0])|[0.20754716981132...|
|(12,[5],[1.0])|[0.20754716981132...|
|(12,[5],[1.0])|[0.20183486238532...|
|(12,[5],[1.0])|[0.20183486238532...|
|(12,[5],[1.0])|[0.19626168224299...|
|(12,[5],[1.0])|[0.20560747663551...|
|(12,[5],[1.0])|[0.19047619047619...|
|(12,[5],[1.0])|[0.20370370370370...|
|(12,[5],[1.0])|[0.20754716981132...|
|(12,[5],[1.0])|[0.20370370370370...|
|(12,[5],[1.0])|[0.2,0.4857142857...|
|(12,[5],[1.0])|[0.19230769230769...|
|(12,[5],[1.0])|[0.20388349514563...|
|(12,[5],[1.0])|[0.20388349514563...|
|(12,[5],[1.0])|[0.18867924528301...|
|(12,[5],[1.0])|[0.17821782178217...|
|(12,[5],[1.0])|[0.18811881188118...|
|(12,[5],[1.0])|[0.15533980582524...|
|(12,[5],[1.0])|[0.17142857142857...|
|(12,[5],[1.0])|[0.17821782178217...|
+--------------+--------------------+
only showing top 20 rows
```