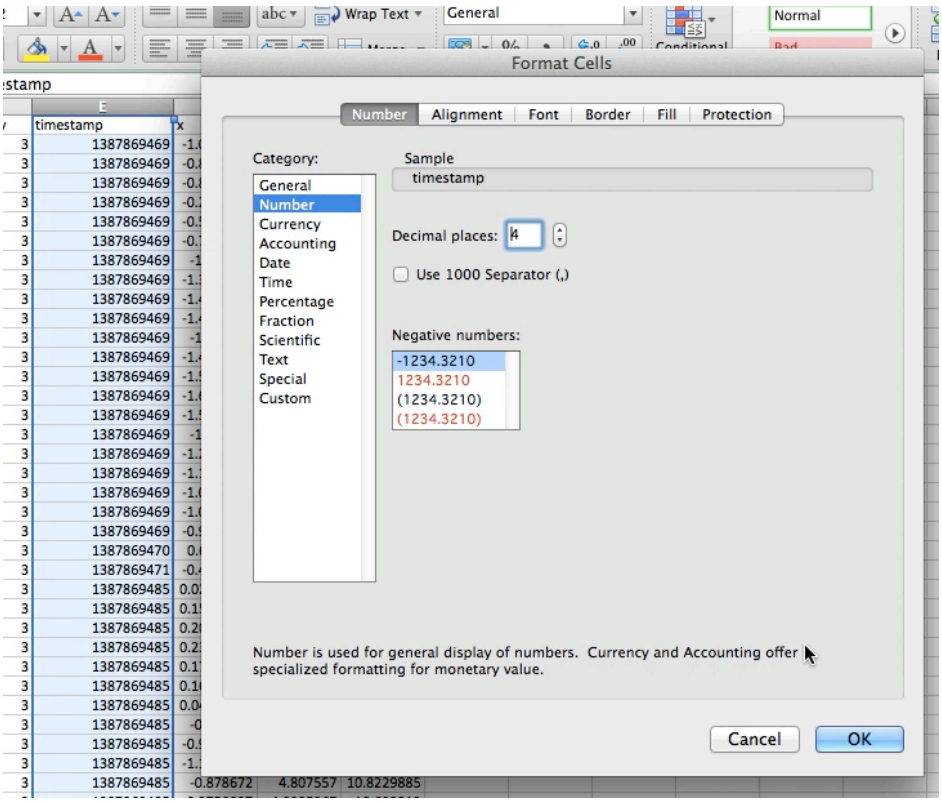
Data Processing using Excel

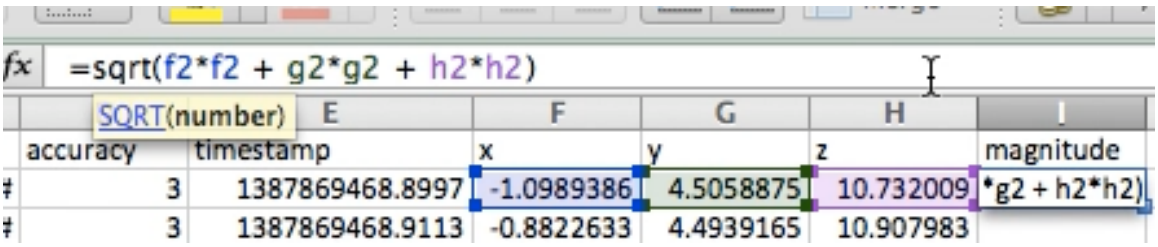
The script directory contains the python scripts to calculate the DFT and labeling the services. We use Excel to perform various transformations, cleaning, and join operations.

1. Merge the LocationProbe.csv file from each day into a single file.

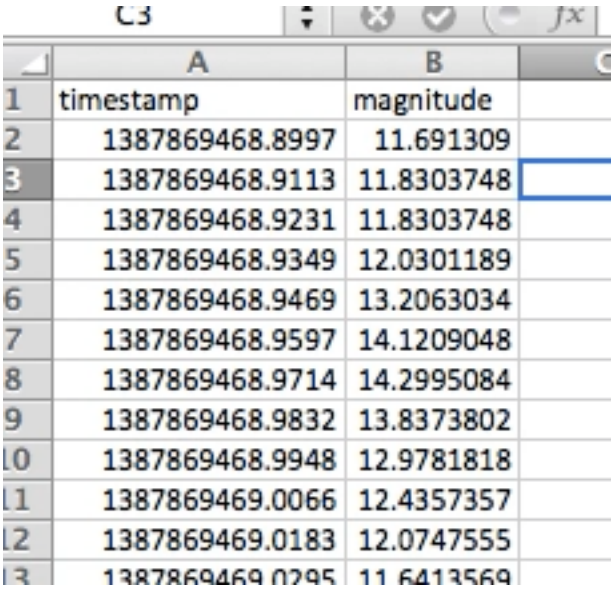
2. Processing AccelerometerSensor.csv
    a. Open the file in Excel
    b. Transform the timestamp column to have 4 decimal places



    c. Add a new column - Magnitude based on the squared sum of x, y, and z coordinate values.

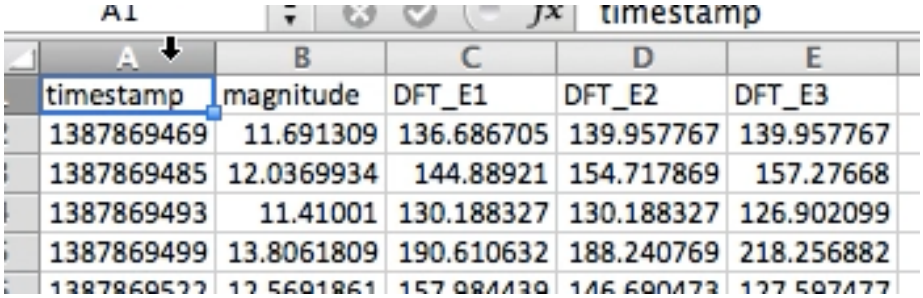


d. Extract the timestamp and Magnitude values into a new file to be used as input for DFT calculation. Name it 'forDFT.csv'



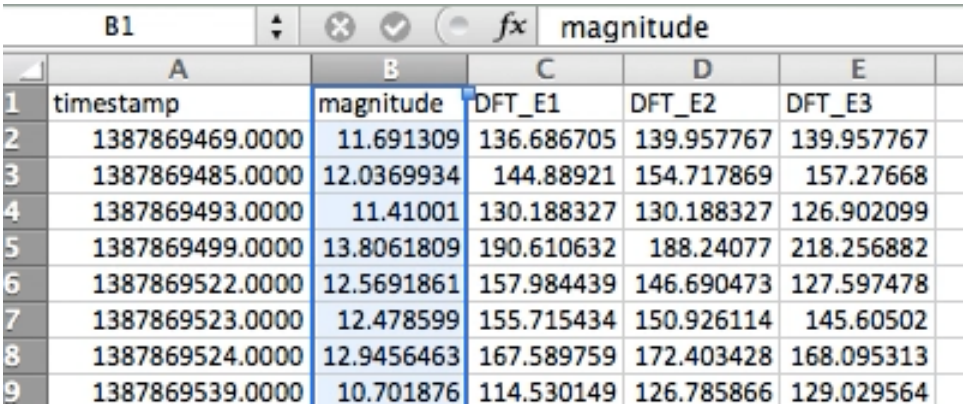
	A	B
1	timestamp	magnitude
2	1387869468.8997	11.691309
3	1387869468.9113	11.8303748
4	1387869468.9231	11.8303748
5	1387869468.9349	12.0301189
6	1387869468.9469	13.2063034
7	1387869468.9597	14.1209048
8	1387869468.9714	14.2995084
9	1387869468.9832	13.8373802
10	1387869468.9948	12.9781818
11	1387869469.0066	12.4357357
12	1387869469.0183	12.0747555
13	1387869469.0295	11.6413569

3. Invoke DFT python script. The output is a new file with timestamp, magnitude and DFT coefficients for 1Hz, 2Hz, and 3Hz. Rename the output file as AccewithDFT.csv



	A	B	C	D	E
	timestamp	magnitude	DFT_E1	DFT_E2	DFT_E3
	1387869469	11.691309	136.686705	139.957767	139.957767
	1387869485	12.0369934	144.88921	154.717869	157.27668
	1387869493	11.41001	130.188327	130.188327	126.902099
	1387869499	13.8061809	190.610632	188.240769	218.256882
	1387869522	12.5691861	157.984439	146.690473	127.597477

4. Transform the timestamp column format to have 4 decimal places in the AccewithDFT.csv



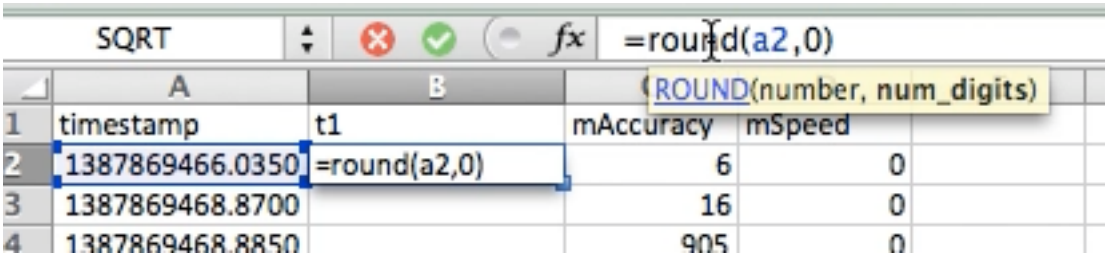
	A	B	C	D	E
1	timestamp	magnitude	DFT_E1	DFT_E2	DFT_E3
2	1387869469.0000	11.691309	136.686705	139.957767	139.957767
3	1387869485.0000	12.0369934	144.88921	154.717869	157.27668
4	1387869493.0000	11.41001	130.188327	130.188327	126.902099
5	1387869499.0000	13.8061809	190.610632	188.24077	218.256882
6	1387869522.0000	12.5691861	157.984439	146.690473	127.597478
7	1387869523.0000	12.478599	155.715434	150.926114	145.60502
8	1387869524.0000	12.9456463	167.589759	172.403428	168.095313
9	1387869539.0000	10.701876	114.530149	126.785866	129.029564

5. Add a new worksheet in AccewithDFT.csv and name it “Location”

6. Open the LocationProbe.csv file and copy 3 columns – “Timestamp”, “mSpeed” and “mAccuracy” contents into the “Location” worksheet

	A	B	C
1	timestamp	mAccuracy	mSpeed
2	1387869466	6	0
3	1387869469	16	0
4	1387869469	905	0
5	1387869469	905	0
6	1387869469	905	0
7	1387869470	16	0
8	1387869470	905	0
9	1387869471	16	0
10	1387869474	16	0
11	1387869478	16	0
12	1387869478	905	0
13	1387869480	16	0
14	1387869484	16	0
15	1387869485	16	0
16	1387869485	905	0
17	1387869485	905	0
18	1387869485	905	0

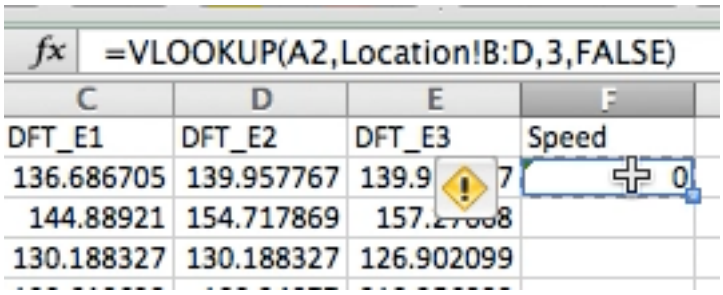
7. Format the timestamp column in the “Location” worksheet to have 4 decimal places
8. Add a new column “t1” to contain rounded values from the timestamp column. We use a *=ROUND(A2,0)* formulae. Repeat it for all rows.



9. Going back the first worksheet containing the accelerometer and DFT data, we add a new column – Speed.

10. The speed column values are fetched from the “Location” worksheet, for the corresponding match in timestamp values. This is the join operation that we perform in Excel.

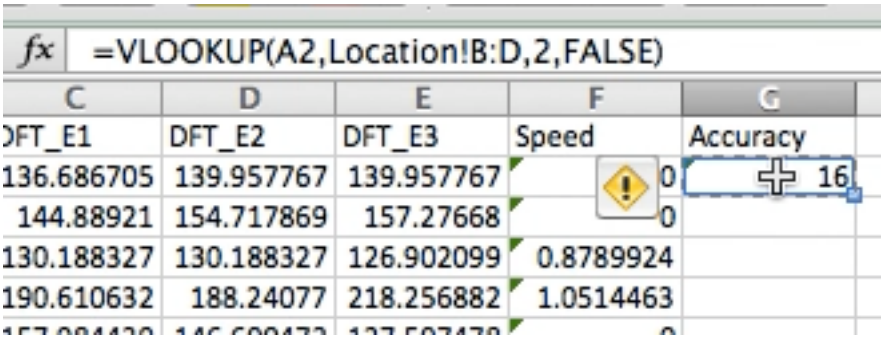
11. Add a formula =VLookUp() to fetch values from the “speed” column in the “Location” worksheet by matching “timestamp” column in the Acceleration worksheet and “t1” column in the “Location” worksheet.



12. Repeat this formulae for the entire speed column

13. Add a new column “Accuracy”

14. Add a formula =VLookUp() to fetch values from the “accuracy” column in the “Location” worksheet by matching “timestamp” column in the Acceleration worksheet and “t1” column in the “Location” worksheet.

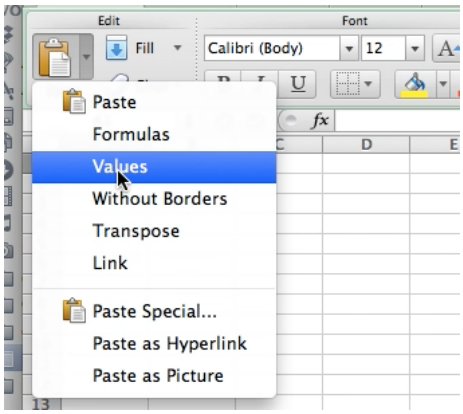


15. Repeat this formulae for the entire accuracy column

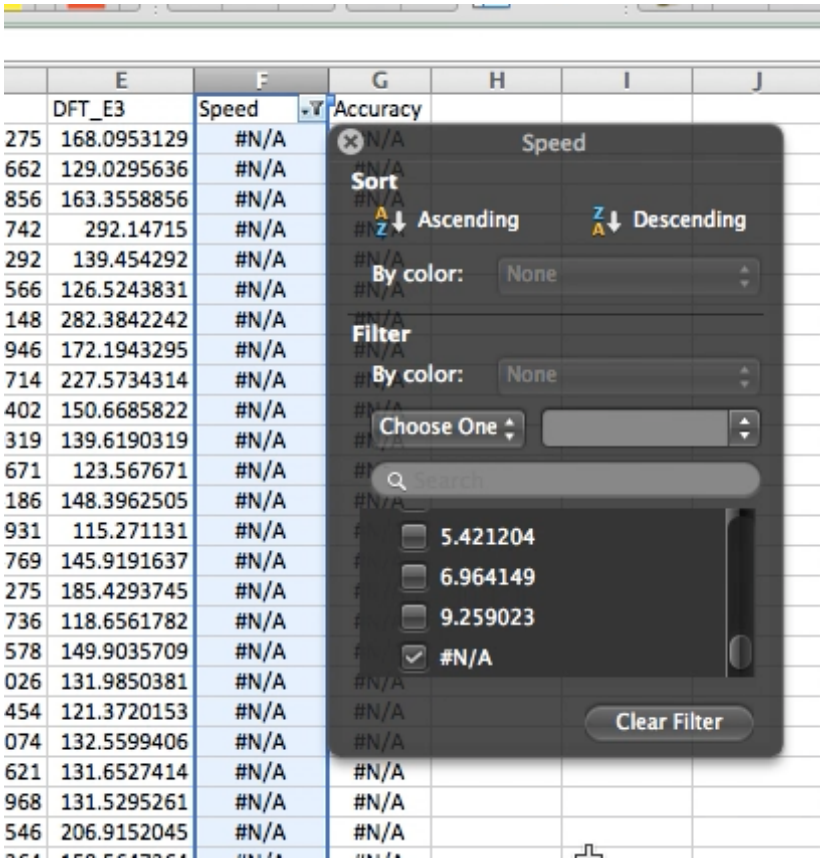
16. Now our columns should be in the order Timestamp, Magnitude, DFT_E1, DFT_E2, DFT_E3, Speed, Accuracy.

	A	B	C	D	E	F	G
1	timestamp	magnitude	DFT_E1	DFT_E2	DFT_E3	Speed	Accuracy
2	1387869469.0000	11.691309	136.686705	139.957767	139.957767	0	16
3	1387869485.0000	12.0369934	144.88921	154.717869	157.27668	0	16
4	1387869493.0000	11.41001	130.188327	130.188327	126.902099	0.8789924	6
5	1387869499.0000	13.8061809	190.610632	188.24077	218.256882	1.0514463	6
6	1387869522.0000	12.5691861	157.984439	146.690473	127.597478	0	905

17. We copy all rows and used the “paste special” feature to paste only the values into a new file. This is used to remove all references and formula mapping that existed in the previous worksheet.



18. Add a filter on the “Speed” column and select only “NA” values to be displayed



19. Delete all “NA” rows. These are rows for which data could not be joined between the LocationProbe and AccelerometerSensor files.

20. Remove all filters and export worksheet as a CSV file. Lets name it “AcceJoinedData.csv”

21. Invoke the addLabel Python script and use “AcceJoinedData.csv” file in the script as input.

22. The output of the script is another CSV file with the “label” column appended to our joined CSV file. Let us name the resultant file “ProcessedData.csv”

	A	B	C	D	E	F	G	H
1	timestamp	magnitude	DFT_E1	DFT_E2	DFT_E3	Speed	Accuracy	mode
2	1387869469	11.691309	136.686705	139.957767	139.957767	0	16	walking
3	1387869485	12.0369934	144.88921	154.717869	157.27668	0	16	walking
4	1387869493	11.41001	130.188327	130.188327	126.902099	0.8789924	6	walking
5	1387869499	13.8061809	190.610632	188.24077	218.256882	1.0514463	6	walking
6	1387869522	12.5691861	157.984439	146.690473	127.597478	0	905	walking
7	1387869523	12.478599	155.715434	150.926114	145.60502	1.1312865	8	walking
8	1387869552	13.4970618	182.170676	198.911571	217.430248	1.1908313	6	walking
9	1387869584	10.5916402	112.182842	480.853665	435.038933	1.5446255	12	walking
10	1387869644	8.85123024	78.3442768	36.5938333	67.0036808	1.3220195	16	walking
11	1387869651	14.2504924	203.076534	176.143405	158.784958	1.3220195	16	walking
12	1387869653	12.6943826	161.14735	161.1475	166.848715	1.3220195	16	walking

23. Open the “ProcessedData.csv” and apply filters on the “label” column. Select all “NA” rows and delete them.

24. Save this file as ProcessedAllData1.csv

25. Invoke SVM:

    1. If this is the first file, then we directly train our model using it as the training set. From the command line, we invoke the R-script svmTraining.R with the file path and other parameters as below

    Rscript svmTraining.R <processed_Data_file> <kernel> <classification_type> <model_file_name>

    Eg. *Rscript svmTraining.R acee_combined.csv linear C-classification mode_of_transport_model*

```
Shrikanths-MacBook-Air:jetty-distribution-9.1.0.v20131115 shri$
_model
Loading required package: class
Warning message:
package 'e1071' was built under R version 3.0.2
[1] "temp_data_dm_service/ProcessedAllData.csv"
[2] "linear"
[3] "C-classification"
[4] "mode_of_transport_model"
[1] "modelName: "
[1] "models/mode_of_transport_model.RData"

Call:
svm.default(x = x, y = y, type = c_type, kernel = kerneltype)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1
      gamma:  0.1428571

Number of Support Vectors:  156

 ( 68 56 21 11 )


Number of Classes:  4

Levels:
 auto bus stationary walking
```

2. If this is not the first file, we first invoke the SVM_test script and evaluate our model for the accuracy.

3. Now merge this file with the previous training data and train the SVM model again.

   Rscript Rscripts/svmTesting.R <test_Data_Set> <model_file_path> <prediction_out_file> <matrix_file_path>

   Eg. Rscript Rscripts/svmTesting.R Data2.csv models/mode_of_transport_model.RData predict.csv matrix.csv