

DICTIONARY:

1. Compute the document frequency of all words (how many documents each word occurred in)

A. <<DocumentTermMatrix (documents: 10, terms: 15)>>

Non-/sparse entries: 18/132

Sparsity : 88%

Maximal term length: 10

Weighting : term frequency (tf)

Sample :

Terms

Docs able absence accept acceptable accepted according act actions acts actually

1	1	1	1	3	1	3	2	3	4	3
10	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	2	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
5	0	1	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	1	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0

Accuracy of train model:

We got about 68.8% percent accuracy on training model.

Accuracy : 0.688

95% CI : (0.6805, 0.6954)

```
> confusionMatrix(pred_train,Final.train$Folder_name)
```

Confusion Matrix and Statistics

	Reference				
Prediction	alt.atheism	comp.graphics	comp.os.ms-windows.misc	comp.sys.ibm.pc.hardware	comp.sys.mac.hardware
alt.atheism	363	0	0	0	0
comp.graphics	0	643	0	0	0
comp.os.ms-windows.misc	9	5	716	2	0
comp.sys.ibm.pc.hardware	22	6	3	730	3
comp.sys.mac.hardware	187	6	0	2	733
comp.windows.x	2	3	0	0	0
misc.forsale	83	85	31	28	26
rec.autos	0	0	0	0	0
rec.motorcycles	23	0	0	0	0
rec.sport.baseball	3	0	0	0	0
rec.sport.hockey	2	0	0	0	0
sci.crypt	0	2	2	0	1
sci.electronics	0	0	0	0	0
sci.med	0	1	0	0	0
sci.space	0	1	0	0	0
soc.religion.christian	0	0	0	0	0
talk.politics.guns	0	0	0	0	0
talk.politics.mideast	2	0	0	0	0
talk.politics.misc	0	2	1	0	0
talk.religion.misc	38	0	0	0	0

		Reference					
Prediction		comp.windows.x	misc.forsale	rec.autos	rec.motorcycles	rec.sport.baseball	
rec.sport.hockey	sci.crypt						
alt.atheism		0	0	0	0	0	0
comp.graphics		2	0	0	1	0	2

comp.os.ms-windows.misc	0	4	0	0	0	1	22
comp.sys.ibm.pc.hardware	2	4	6	0	1	3	24
comp.sys.mac.hardware	0	6	101	19	16	20	158
comp.windows.x	702	0	1	2	0	0	8
misc.forsale	55	732	216	68	135	97	124
rec.autos	0	0	385	0	0	0	0
rec.motorcycles	0	1	13	661	0	1	0
rec.sport.baseball	0	0	0	0	592	0	0
rec.sport.hockey	0	0	0	0	0	636	0
sci.crypt	1	0	0	0	0	0	419
sci.electronics	0	0	1	0	0	0	5
sci.med	0	0	0	1	0	0	0
sci.space	0	0	0	0	0	0	0
soc.religion.christian	0	0	0	0	0	0	0
talk.politics.guns	0	0	0	0	0	0	0
talk.politics.mideast	1	0	0	0	0	0	0
talk.politics.misc	1	0	2	1	0	0	1
talk.religion.misc	0	0	0	0	0	0	0

Reference

Prediction	sci.electronics sci.med sci.space soc.religion.christian talk.politics.guns						
talk.politics.mideast							
alt.atheism	0	0	0	0	0	4	
comp.graphics	2	5	3	1	0	0	
comp.os.ms-windows.misc	3	17	22	16	17	52	
comp.sys.ibm.pc.hardware	18	13	5	2	12	5	
comp.sys.mac.hardware	57	130	138	30	162	137	
comp.windows.x	1	1	5	1	3	4	
misc.forsale	203	180	182	136	162	182	
rec.autos	0	0	0	0	0	0	
rec.motorcycles	1	6	2	0	41	30	
rec.sport.baseball	0	0	0	0	2	4	
rec.sport.hockey	0	0	1	0	1	1	

sci.crypt	0	0	0	0	3	0
sci.electronics	449	0	0	0	2	0
sci.med	0	402	0	1	1	1
sci.space	0	0	390	0	0	0
soc.religion.christian	0	0	1	549	0	0
talk.politics.guns	1	0	0	0	319	0
talk.politics.mideast	0	1	1	0	1	322
talk.politics.misc	0	1	0	0	6	12
talk.religion.misc	0	0	0	0	11	1

Reference

Prediction talk.politics.misc talk.religion.misc

alt.atheism	1	103
comp.graphics	1	3
comp.os.ms-windows.misc	37	20
comp.sys.ibm.pc.hardware	10	8
comp.sys.mac.hardware	132	113
comp.windows.x	0	1
misc.forsale	202	168
rec.autos	1	0
rec.motorcycles	21	13
rec.sport.baseball	3	1
rec.sport.hockey	1	0
sci.crypt	1	0
sci.electronics	0	0
sci.med	1	0
sci.space	1	0
soc.religion.christian	0	1
talk.politics.guns	14	4
talk.politics.mideast	9	1
talk.politics.misc	307	30
talk.religion.misc	27	270

Overall Statistics

Accuracy : 0.688

95% CI : (0.6805, 0.6954)

No Information Rate : 0.0513

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.6715

McNemar's Test P-Value : NA

Statistics by Class:

Class: alt.atheism Class: comp.graphics Class: comp.os.ms-windows.misc Class:
comp.sys.ibm.pc.hardware

Sensitivity	0.49455	0.85279	0.95086	0.95801
Specificity	0.99243	0.99860	0.98407	0.98968
Pos Pred Value	0.77070	0.96983	0.75928	0.83238
Neg Pred Value	0.97446	0.99226	0.99737	0.99773
Prevalence	0.04893	0.05027	0.05020	0.05080
Detection Rate	0.02420	0.04287	0.04773	0.04867
Detection Prevalence	0.03140	0.04420	0.06287	0.05847
Balanced Accuracy	0.74349	0.92569	0.96747	0.97384

Class: comp.sys.mac.hardware Class: comp.windows.x Class: misc.forsale Class: rec.autos
Class: rec.motorcycles

Sensitivity	0.96068	0.91885	0.9799	0.53103	0.87782
Specificity	0.90068	0.99775	0.8342	0.99993	0.98933
Pos Pred Value 0.81304	0.34141	0.95640	0.2365	0.99741	
Neg Pred Value 0.99352	0.99767	0.99565	0.9987	0.97673	
Prevalence	0.05087	0.05093	0.0498	0.04833	0.05020
Detection Rate 0.04407	0.04887	0.04680	0.0488	0.02567	
Detection Prevalence 0.05420	0.14313	0.04893	0.2063	0.02573	

Balanced Accuracy	0.93068	0.95830	0.9071	0.76548
0.93358				

Class: rec.sport.baseball Class: rec.sport.hockey Class: sci.crypt Class: sci.electronics Class: sci.med

Sensitivity	0.79570	0.83905	0.54915	0.61088	0.5317
Specificity	0.99909	0.99958	0.99930	0.99944	0.9996
Pos Pred Value	0.97851	0.99065	0.97669	0.98249	0.9853
Neg Pred Value	0.98944	0.99150	0.97639	0.98033	0.9757
Prevalence	0.04960	0.05053	0.05087	0.04900	0.0504
Detection Rate	0.03947	0.04240	0.02793	0.02993	0.0268
Detection Prevalence	0.04033	0.04280	0.02860	0.03047	0.0272
Balanced Accuracy	0.89739	0.91931	0.77422	0.80516	0.7657

Class: sci.space Class: soc.religion.christian Class: talk.politics.guns Class: talk.politics.mideast

Sensitivity	0.52000	0.74592	0.42934	0.42649
Specificity	0.99986	0.99986	0.99867	0.99888
Pos Pred Value	0.99490	0.99637	0.94379	0.95266
Neg Pred Value	0.97536	0.98706	0.97108	0.97047
Prevalence	0.05000	0.04907	0.04953	0.05033
Detection Rate	0.02600	0.03660	0.02127	0.02147
Detection Prevalence	0.02613	0.03673	0.02253	0.02253
Balanced Accuracy	0.75993	0.87289	0.71400	0.71268

Class: talk.politics.misc Class: talk.religion.misc

Sensitivity	0.39922	0.36685
Specificity	0.99599	0.99460
Pos Pred Value	0.84341	0.77810
Neg Pred Value	0.96843	0.96820
Prevalence	0.05127	0.04907
Detection Rate	0.02047	0.01800
Detection Prevalence	0.02427	0.02313
Balanced Accuracy	0.69761	0.68072

And about 68.29% accuracy on test model.

Accuracy : 0.6829

95% CI : (0.6698, 0.6958)

```
> confusionMatrix(pred_test, test_labels)
```

Confusion Matrix and Statistics

	Reference				
Prediction	alt.atheism	comp.graphics	comp.os.ms-windows.misc	comp.sys.ibm.pc.hardware	comp.sys.mac.hardware
alt.atheism	118	1	0	0	0
comp.graphics	0	185	0	0	0
comp.os.ms-windows.misc	6	5	231	0	0
comp.sys.ibm.pc.hardware	6	2	3	220	2
comp.sys.mac.hardware	6	2	0	1	224
comp.windows.x	1	1	2	0	0
misc.forsale	57	36	14	21	24
rec.autos	1	0	0	0	0
rec.motorcycles	8	0	0	0	0
rec.sport.baseball	52	0	0	1	0
rec.sport.hockey	0	0	0	0	0
sci.crypt	0	1	0	0	0
sci.electronics	2	0	0	0	0
sci.med	0	1	0	0	0
sci.space	0	0	0	0	0
soc.religion.christian	2	0	0	0	0
talk.politics.guns	0	0	0	0	0
talk.politics.mideast	0	0	0	0	0
talk.politics.misc	0	0	0	0	0
talk.religion.misc	23	0	0	0	0

	Reference				
Prediction	comp.windows.x	misc.forsale	rec.autos	rec.motorcycles	rec.sport.baseball
rec.sport.hockey	0	0	0	0	0
alt.atheism	0	0	0	0	0
comp.graphics	4	0	0	0	8
comp.os.ms-windows.misc	4	2	0	0	8
comp.sys.ibm.pc.hardware	1	1	0	0	4

comp.sys.mac.hardware	1	0	0	1	0	0	1
comp.windows.x	176	0	1	0	0	0	2
misc.forsale	68	261	52	20	15	42	53
rec.autos	0	1	186	0	0	0	0
rec.motorcycles	0	0	1	237	0	0	2
rec.sport.baseball	0	1	0	0	191	14	6
rec.sport.hockey	0	0	0	0	0	203	0
sci.crypt	0	0	0	0	0	0	148
sci.electronics	0	0	0	0	0	0	4
sci.med	0	0	0	0	0	0	0
sci.space	0	0	0	0	0	0	0
soc.religion.christian	0	0	0	0	0	0	0
talk.politics.guns	0	0	0	0	0	0	0
talk.politics.mideast	0	0	0	0	0	1	0
talk.politics.misc	0	0	0	0	0	0	0
talk.religion.misc	0	0	0	0	0	0	0

Reference

Prediction	sci.electronics sci.med sci.space soc.religion.christian talk.politics.guns						
talk.politics.mideast							
alt.atheism	0	0	0	0	0	1	
comp.graphics	0	8	1	1	0	0	
comp.os.ms-windows.misc		2	6	3	4	1	7
comp.sys.ibm.pc.hardware		5	4	2	1	4	1
comp.sys.mac.hardware		6	11	7	0	7	5
comp.windows.x	0	0	1	0	0	0	
misc.forsale	83	74	65	50	62	63	
rec.autos	3	1	3	0	14	6	
rec.motorcycles	0	3	1	0	7	7	
rec.sport.baseball	0	6	9	3	16	33	
rec.sport.hockey	0	0	0	0	0	0	
sci.crypt	0	0	0	0	1	0	
sci.electronics	164	0	0	0	1	0	
sci.med	0	117	0	0	1	0	
sci.space	0	0	139	0	0	0	
soc.religion.christian	0	0	0	189	0	0	
talk.politics.guns	0	0	0	0	139	0	
talk.politics.mideast	0	1	1	0	1	105	
talk.politics.misc	0	3	1	0	1	5	
talk.religion.misc	0	0	0	0	5	1	

Reference

Prediction	talk.politics.misc	talk.religion.misc
alt.atheism	0	47
comp.graphics	0	1
comp.os.ms-windows.misc	9	5
comp.sys.ibm.pc.hardware	2	2
comp.sys.mac.hardware	2	2
comp.windows.x	0	0
misc.forsale	74	68
rec.autos	31	5
rec.motorcycles	7	0
rec.sport.baseball	27	13
rec.sport.hockey	0	0
sci.crypt	0	0
sci.electronics	0	1
sci.med	3	0
sci.space	3	1
soc.religion.christian	0	6
talk.politics.guns	16	4
talk.politics.mideast	8	1
talk.politics.misc	76	5
talk.religion.misc	17	103

Overall Statistics

Accuracy : 0.6829

95% CI : (0.6698, 0.6958)

No Information Rate : 0.0564

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.6662

Mcnemar's Test P-Value : NA

Statistics by Class:

	Class: alt.atheism	Class: comp.graphics	Class: comp.os.ms-windows.misc	Class: comp.sys.ibm.pc.hardware
Sensitivity	0.41844	0.79060	0.92400	0.90535
Specificity	0.98961	0.99517	0.98694	0.99158

Pos Pred value	0.70659	0.88942	0.78840	0.84615	
Neg Pred Value	0.96604	0.98977	0.99596	0.99514	
Prevalence	0.05645	0.04684	0.05004	0.04864	
Detection Rate	0.02362	0.03703	0.04624	0.04404	
Detection Prevalence	0.03343	0.04163	0.05865	0.05204	
Balanced Accuracy	0.70402	0.89288	0.95547	0.94847	
Class: comp.sys.mac.hardware Class: comp.windows.x Class: misc.forsale Class: rec.autos Class: rec.motorcycles					
Sensitivity	0.89600	0.69291	0.98120	0.77500	0.91860
Specificity	0.98778	0.99831	0.80106	0.98633	0.99240
Pos Pred Value	0.79433	0.95652	0.21714	0.74104	
Neg Pred Value	0.99448	0.98379	0.99868	0.98862	
Prevalence	0.05004	0.05084	0.05324	0.04804	
Detection Rate	0.04484	0.03523	0.05224	0.03723	
Detection Prevalence	0.05645	0.03683	0.24059	0.05024	
Balanced Accuracy	0.94189	0.84561	0.89113	0.88067	
Class: rec.sport.baseball Class: rec.sport.hockey Class: sci.crypt Class: sci.electronics Class: sci.med					
Sensitivity	0.92718	0.78077	0.61157	0.62357	0.50000
Specificity	0.96221	1.00000	0.99958	0.99831	0.99895
Pos Pred Value	0.51344	1.00000	0.98667	0.95349	0.95902
Neg Pred Value	0.99676	0.98811	0.98060	0.97948	0.97600
Prevalence	0.04123	0.05204	0.04844	0.05264	0.04684
Detection Rate	0.03823	0.04063	0.02962	0.03283	0.02342
Detection Prevalence	0.07446	0.04063	0.03002	0.03443	
Balanced Accuracy	0.94470	0.89038	0.80557	0.81094	0.74948
Class: sci.space Class: soc.religion.christian Class: talk.politics.guns Class: talk.politics.mideast					
Sensitivity	0.59657	0.76210	0.53462	0.44872	
Specificity	0.99916	0.99832	0.99578	0.99727	
Pos Pred Value	0.97203	0.95939	0.87421	0.88983	
Neg Pred Value	0.98063	0.98771	0.97498	0.97355	
Prevalence	0.04664	0.04964	0.05204	0.04684	
Detection Rate	0.02782	0.03783	0.02782	0.02102	
Detection Prevalence	0.02862	0.03943	0.03183	0.02362	
Balanced Accuracy	0.79786	0.88021	0.76520	0.72299	
Class: talk.politics.misc Class: talk.religion.misc					
Sensitivity	0.27636	0.39015			

specificity	0.99682	0.99028
Pos Pred Value	0.83516	0.69128
Neg Pred Value	0.95943	0.96678
Prevalence	0.05504	0.05284
Detection Rate	0.01521	0.02062
Detection Prevalence	0.01821	0.02982
Balanced Accuracy	0.63659	0.69022