Principal Component Analysis: *Tweets* *–Charles 4/23/2012*
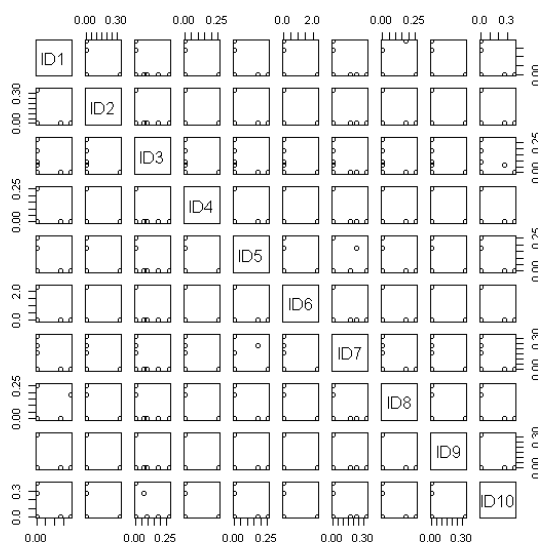
Approach 1   *rows: keywords, columns: IDs*

> data1 <- read.table("output2.csv", header=T, sep=",")

> print(data1)

> plot(data1)

> data1_pca <- prcomp(data1, scale = TRUE)

> print(data1_pca)

Standard deviations:

  [1] 1.1330103 1.1139063 1.0844549 1.0566032 1.0429877 1.0182214 1.0089421

  [8] 0.9985871 0.9875787 0.2607632

Rotation:

	PC1	PC2	PC3	PC4	PC5
ID1	-0.525141206	0.41115898	-0.13724385	0.09762030	-0.03542829
ID2	-0.009984241	-0.07219987	0.11114815	-0.34502396	0.87384147
ID3	-0.038824407	-0.54832568	-0.41732713	0.18265310	-0.06291871
ID4	-0.022219816	-0.19199513	0.79968275	0.39758398	-0.10135835
ID5	0.390207077	0.23851504	-0.09298418	0.06955273	-0.02583705
ID6	-0.003252189	-0.02203971	0.02819675	-0.05096152	0.03060483
ID7	0.616749002	0.37628792	-0.14603667	0.10831898	-0.03991910
ID8	-0.434143073	0.34196858	-0.11575204	0.08434133	-0.03124399
ID9	-0.012110434	-0.08996183	0.15188024	-0.79986244	-0.46417943
ID10	-0.029114114	-0.40522695	-0.29651613	0.11803615	-0.03573749

	PC6	PC7	PC8	PC9	PC10

ID1	-0.005353423	0.02771192	0.0004774272	0.634763348	0.3490457
ID2	-0.022487479	0.09186974	0.0014575756	-0.005449596	0.3011903
ID3	-0.596612585	0.02757737	0.0006595855	-0.003001343	0.3600448
ID4	-0.011645398	0.05832089	0.0010672803	-0.004418257	0.3889055
ID5	-0.004215972	0.02377100	0.8455286221	-0.001654196	0.2473734
ID6	0.020583282	-0.98752169	0.0029405843	-0.006283757	0.1396567
ID7	-0.006244021	0.03294042	-0.5339157001	-0.003051017	0.3964536
ID8	-0.005132340	0.02879122	0.0006353945	-0.772612026	0.2744159
ID9	-0.017753990	0.07916795	0.0013275996	-0.005148750	0.3268625
ID10	0.801599329	0.05181154	0.0007965665	-0.003143571	0.2942032

> summary(data1_pca)

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.1330	1.1139	1.0845	1.0566	1.0430	1.0182	1.0089
Proportion of Variance	0.1284	0.1241	0.1176	0.1116	0.1088	0.1037	0.1018
Cumulative Proportion	0.1284	0.2525	0.3700	0.4817	0.5905	0.6942	0.7960

	PC8	PC9	PC10
Standard deviation	0.99859	0.98758	0.2608
Proportion of Variance	0.09972	0.09753	0.0068
Cumulative Proportion	0.89567	0.99320	1.0000

The linear combination for the first principal component is
> a1 <- data1_pca$rotation[,1]
> a1

ID1	ID2	ID3	ID4	ID5
-0.525141206	-0.009984241	-0.038824407	-0.022219816	0.390207077

ID6	ID7	ID8	ID9	ID10
-0.003252189	0.616749002	-0.434143073	-0.012110434	-0.029114114

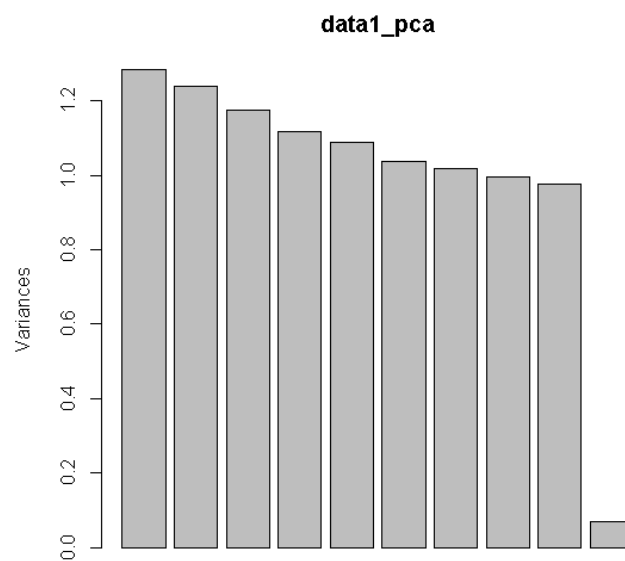We compute the first principal component score for each word.
> predict(data1_pca)[,1]

```
 [1]   1.35983980   1.35983980   1.53655876 -0.11719715 -1.44935990 -0.17731974
 [7] -0.07123756 -1.57286091 -0.07123756 -0.11719715 -1.44935990   2.98264163
[13]   1.35983980 -0.11719715   1.35983980 -1.44935990 -0.06376128 -1.44935990
[19]   1.53655876 -1.57286091 -0.04878790 -0.12054051 -0.04878790   1.35983980
[25] -0.04442704   1.35983980   1.35983980   1.35983980 -0.07123756   1.35983980
[31] -0.06376128   1.53655876 -0.07123756 -0.04878790 -0.12370608 -0.17731974
```

[37] 1.35983980 -1.44935990 -0.07123756 -1.44935990 2.72666165 -0.12054051
[43] -1.44935990 -0.04442704 -0.07123756 -1.57286091 -0.06376128 -0.07123756
[49] -0.03340296 -0.04878790 -0.04442704 -1.57286091 -0.07123756 -0.04442704
[55] -1.44935990 -0.07123756 -0.06376128 -0.11719715 -1.57286091 -0.04878790
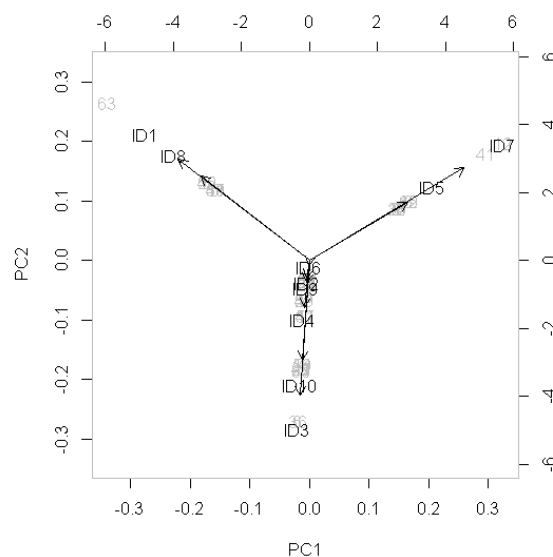[61] -0.12054051 -0.04442704 -3.11784212 -0.12370608 1.53655876 -0.04878790

Barplot of the variances explained by the principal components.

> **plot(data1_pca)**

Biplot of the (scaled) first two principal components.

> **biplot(data1_pca, col = c("gray", "black"))**

Approach 2 *columns: keywords, rows: IDs*

```
> data2 <- read.table("output2-2.csv", header=T, sep=",")
> print(data2)
> plot(data2)
Error in plot.new() : figure margins too large

> data2_pca <- prcomp(data2, scale = TRUE)
> print(data2_pca)
Standard deviations:
 [1] 3.537317e+00 3.179270e+00 3.134357e+00 2.849473e+00 2.572210e+00
 [6] 2.497146e+00 2.404012e+00 2.249802e+00 1.320251e+00 1.191307e-15
```

Rotation:

	PC1	PC2	PC3	PC4
gituuuuu	0.276513418	-0.008338855	0.012150887	0.057214081
oooooh	0.276513418	-0.008338855	0.012150887	0.057214081
meeeeeee	0.003977118	0.003289388	-0.005664495	-0.064213382
xjd1bnfg4w	-0.029119708	-0.032314086	0.006444253	-0.110312194
app	-0.051696746	0.055635714	-0.243265471	0.194411081
alive	-0.063771437	-0.251602404	0.104612151	0.148886170
belapa	-0.054381761	0.205247016	0.199919965	0.127824420
chef	-0.026522465	0.014338996	-0.050798269	-0.060306892
ajadah	-0.054381761	0.205247016	0.199919965	0.127824420
photo	-0.029119708	-0.032314086	0.006444253	-0.110312194
fine	-0.051696746	0.055635714	-0.243265471	0.194411081
rt	0.233647500	-0.004830305	0.006472088	0.006075307
iyaaaa	0.276513418	-0.008338855	0.012150887	0.057214081
tumblr	-0.029119708	-0.032314086	0.006444253	-0.110312194
lo	0.276513418	-0.008338855	0.012150887	0.057214081
working	-0.051696746	0.055635714	-0.243265471	0.194411081
tv	-0.063771437	-0.251602404	0.104612151	0.148886170
its	-0.051696746	0.055635714	-0.243265471	0.194411081
tweet	0.003977118	0.003289388	-0.005664495	-0.064213382
hungry	-0.026522465	0.014338996	-0.050798269	-0.060306892
lol	-0.023373500	0.006427332	-0.011131299	-0.165453014
jan	-0.063771437	-0.251602404	0.104612151	0.148886170
feel	-0.023373500	0.006427332	-0.011131299	-0.165453014
gituuu	0.276513418	-0.008338855	0.012150887	0.057214081

Word				
much	-0.019640518	0.004858535	-0.008233861	-0.093752112
online	0.276513418	-0.008338855	0.012150887	0.057214081
renyfebry	0.276513418	-0.008338855	0.012150887	0.057214081
sih	0.276513418	-0.008338855	0.012150887	0.057214081
mandi	-0.054381761	0.205247016	0.199919965	0.127824420
ngapain	0.276513418	-0.008338855	0.012150887	0.057214081
shows	-0.063771437	-0.251602404	0.104612151	0.148886170
wants	0.003977118	0.003289388	-0.005664495	-0.064213382
cekalang	-0.054381761	0.205247016	0.199919965	0.127824420
kissmytweetmuah	-0.023373500	0.006427332	-0.011131299	-0.165453014
http	-0.044330382	-0.092501536	0.031484075	-0.074420918
get	-0.063771437	-0.251602404	0.104612151	0.148886170
nif	0.276513418	-0.008338855	0.012150887	0.057214081
ipod	-0.051696746	0.055635714	-0.243265471	0.194411081
males	-0.054381761	0.205247016	0.199919965	0.127824420
haha	-0.051696746	0.055635714	-0.243265471	0.194411081
hanifpeng	0.276513418	-0.008338855	0.012150887	0.057214081
abduction	-0.063771437	-0.251602404	0.104612151	0.148886170
tumblr.1	-0.051696746	0.055635714	-0.243265471	0.194411081
chalk_flew_up	-0.019640518	0.004858535	-0.008233861	-0.093752112
nantian	-0.054381761	0.205247016	0.199919965	0.127824420
name	-0.026522465	0.014338996	-0.050798269	-0.060306892
bsstracing	-0.063771437	-0.251602404	0.104612151	0.148886170
jam	-0.054381761	0.205247016	0.199919965	0.127824420
stadd	-0.011984403	0.002458364	-0.004033862	-0.034298159
word	-0.023373500	0.006427332	-0.011131299	-0.165453014
longer	-0.019640518	0.004858535	-0.008233861	-0.093752112
salad	-0.026522465	0.014338996	-0.050798269	-0.060306892
mandinyah	-0.054381761	0.205247016	0.199919965	0.127824420
maybe	-0.019640518	0.004858535	-0.008233861	-0.093752112
biebercrombie	-0.051696746	0.055635714	-0.243265471	0.194411081
akuh	-0.054381761	0.205247016	0.199919965	0.127824420
html	-0.063771437	-0.251602404	0.104612151	0.148886170
seco	-0.029119708	-0.032314086	0.006444253	-0.110312194
calling	-0.026522465	0.014338996	-0.050798269	-0.060306892
u	-0.023373500	0.006427332	-0.011131299	-0.165453014
jungle	-0.063771437	-0.251602404	0.104612151	0.148886170
yup	-0.019640518	0.004858535	-0.008233861	-0.093752112

im	-0.059173674	0.053338756	-0.224557561	0.105982325
com	-0.044330382	-0.092501536	0.031484075	-0.074420918
misslovelace_	0.003977118	0.003289388	-0.005664495	-0.064213382
thats	-0.023373500	0.006427332	-0.011131299	-0.165453014

	PC5	PC6	PC7	PC8
gituuuuu	-0.002845652	0.015817650	0.016762350	0.042740886
oooooh	-0.002845652	0.015817650	0.016762350	0.042740886
meeeeeee	0.007130886	-0.050597033	-0.075489743	-0.417191110
xjd1bnfg4w	0.298804766	0.189108241	0.081460780	0.050839642
app	-0.008564860	0.079102779	-0.019716986	0.008982776
alive	-0.049140791	-0.006166308	0.007402843	0.011934725
belapa	-0.004095272	0.020304576	0.017876088	0.017709971
chef	0.018057003	-0.297651980	0.240098778	0.075571625
ajadah	-0.004095272	0.020304576	0.017876088	0.017709971
photo	0.298804766	0.189108241	0.081460780	0.050839642
fine	-0.008564860	0.079102779	-0.019716986	0.008982776
rt	0.002256693	-0.019667029	-0.035056220	-0.235431031
iyaaaa	-0.002845652	0.015817650	0.016762350	0.042740886
tumblr	0.298804766	0.189108241	0.081460780	0.050839642
lo	-0.002845652	0.015817650	0.016762350	0.042740886
working	-0.008564860	0.079102779	-0.019716986	0.008982776
tv	-0.049140791	-0.006166308	0.007402843	0.011934725
its	-0.008564860	0.079102779	-0.019716986	0.008982776
tweet	0.007130886	-0.050597033	-0.075489743	-0.417191110
hungry	0.018057003	-0.297651980	0.240098778	0.075571625
lol	-0.274927461	0.177365803	0.085024387	0.054489450
jan	-0.049140791	-0.006166308	0.007402843	0.011934725
feel	-0.274927461	0.177365803	0.085024387	0.054489450
gituuu	-0.002845652	0.015817650	0.016762350	0.042740886
much	0.013064227	-0.112369143	-0.337255802	0.177060718
online	-0.002845652	0.015817650	0.016762350	0.042740886
renyfebry	-0.002845652	0.015817650	0.016762350	0.042740886
sih	-0.002845652	0.015817650	0.016762350	0.042740886
mandi	-0.004095272	0.020304576	0.017876088	0.017709971
ngapain	-0.002845652	0.015817650	0.016762350	0.042740886
shows	-0.049140791	-0.006166308	0.007402843	0.011934725
wants	0.007130886	-0.050597033	-0.075489743	-0.417191110
cekalang	-0.004095272	0.020304576	0.017876088	0.017709971

kissmytweetmuah	-0.274927461	0.177365803	0.085024387	0.054489450
http	0.286400638	0.187229330	0.083060455	0.053589773
get	-0.049140791	-0.006166308	0.007402843	0.011934725
nif	-0.002845652	0.015817650	0.016762350	0.042740886
ipod	-0.008564860	0.079102779	-0.019716986	0.008982776
males	-0.004095272	0.020304576	0.017876088	0.017709971
haha	-0.008564860	0.079102779	-0.019716986	0.008982776
hanifpeng	-0.002845652	0.015817650	0.016762350	0.042740886
abduction	-0.049140791	-0.006166308	0.007402843	0.011934725
tumblr.1	-0.008564860	0.079102779	-0.019716986	0.008982776
chalk_flew_up	0.013064227	-0.112369143	-0.337255802	0.177060718
nantian	-0.004095272	0.020304576	0.017876088	0.017709971
name	0.018057003	-0.297651980	0.240098778	0.075571625
bsstracing	-0.049140791	-0.006166308	0.007402843	0.011934725
jam	-0.004095272	0.020304576	0.017876088	0.017709971
stadd	0.002517156	-0.014914585	-0.016162696	-0.022138683
word	-0.274927461	0.177365803	0.085024387	0.054489450
longer	0.013064227	-0.112369143	-0.337255802	0.177060718
salad	0.018057003	-0.297651980	0.240098778	0.075571625
mandinyah	-0.004095272	0.020304576	0.017876088	0.017709971
maybe	0.013064227	-0.112369143	-0.337255802	0.177060718
biebercrombie	-0.008564860	0.079102779	-0.019716986	0.008982776
akuh	-0.004095272	0.020304576	0.017876088	0.017709971
html	-0.049140791	-0.006166308	0.007402843	0.011934725
seco	0.298804766	0.189108241	0.081460780	0.050839642
calling	0.018057003	-0.297651980	0.240098778	0.075571625
u	-0.274927461	0.177365803	0.085024387	0.054489450
jungle	-0.049140791	-0.006166308	0.007402843	0.011934725
yup	0.013064227	-0.112369143	-0.337255802	0.177060718
im	0.006545300	-0.155759136	0.159637437	0.061957294
com	0.286400638	0.187229330	0.083060455	0.053589773
misslovelace_	0.007130886	-0.050597033	-0.075489743	-0.417191110
thats	-0.274927461	0.177365803	0.085024387	0.054489450

	PC9	PC10
gituuuuu	-0.03449278	-0.304093892
oooooh	-0.03449278	-0.275404854
meeeeeee	-0.14310944	-0.636119246
xjd1bnfg4w	-0.09107964	0.077414166

app	-0.05277487	0.447775520
alive	-0.05083334	0.007520000
belapa	-0.05747605	-0.009326983
chef	-0.10033704	0.050873195
ajadah	-0.05747605	-0.199630549
photo	-0.09107964	-0.070211873
fine	-0.05277487	-0.070120659
rt	-0.12183481	0.108200510
iyaaaa	-0.03449278	0.046333742
tumblr	-0.09107964	-0.070211873
lo	-0.03449278	0.046333742
working	-0.05277487	-0.070120659
tv	-0.05083334	-0.011784204
its	-0.05277487	-0.070120659
tweet	-0.14310944	0.164572168
hungry	-0.10033704	-0.023888495
lol	-0.09638795	-0.012013264
jan	-0.05083334	-0.011784204
feel	-0.09638795	-0.012013264
gituuu	-0.03449278	0.046333742
much	-0.12447920	-0.014415917
online	-0.03449278	0.046333742
renyfebry	-0.03449278	0.046333742
sih	-0.03449278	0.046333742
mandi	-0.05747605	0.019553992
ngapain	-0.03449278	0.046333742
shows	-0.05083334	-0.011784204
wants	-0.14310944	0.164572168
cekalang	-0.05747605	0.019553992
kissmytweetmuah	-0.09638795	-0.012013264
http	-0.10306001	0.030636082
get	-0.05083334	-0.011784204
nif	-0.03449278	0.046333742
ipod	-0.05277487	-0.070120659
males	-0.05747605	0.019553992
haha	-0.05277487	-0.070120659
hanifpeng	-0.03449278	0.046333742
abduction	-0.05083334	-0.011784204

tumblr.1	-0.05277487	-0.070120659
chalk_flew_up	-0.12447920	-0.014415917
nantian	-0.05747605	0.019553992
name	-0.10033704	-0.023888495
bsstracing	-0.05083334	-0.011784204
jam	-0.05747605	0.019553992
stadd	0.75097032	-0.072079585
word	-0.09638795	-0.012013264
longer	-0.12447920	-0.014415917
salad	-0.10033704	-0.023888495
mandinyah	-0.05747605	0.019553992
maybe	-0.12447920	-0.014415917
biebercrombie	-0.05277487	-0.070120659
akuh	-0.05747605	0.019553992
html	-0.05083334	-0.011784204
seco	-0.09107964	-0.070211873
calling	-0.10033704	-0.023888495
u	-0.09638795	-0.012013264
jungle	-0.05083334	-0.011784204
yup	-0.12447920	-0.014415917
im	-0.11375673	-0.037625379
com	-0.10306001	0.030636082
misslovelace_	-0.14310944	0.164572168
thats	-0.09638795	-0.012013264

> summary(data2_pca)
Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	3.5373	3.1793	3.1344	2.8495	2.5722	2.49715	2.40401
Proportion of Variance	0.1896	0.1532	0.1489	0.1230	0.1003	0.09448	0.08756
Cumulative Proportion	0.1896	0.3427	0.4916	0.6146	0.7148	0.80933	0.89690

	PC8	PC9	PC10
Standard deviation	2.24980	1.32025	1.191e-15
Proportion of Variance	0.07669	0.02641	0.000e+00
Cumulative Proportion	0.97359	1.00000	1.000e+00

The linear combination for the first principal component is
> a2 <- data2_pca$rotation[,1]

```
> a2
        gituuuuu            oooooh           meeeeeee         xjd1bnfg4w
      0.276513418        0.276513418        0.003977118       -0.029119708
             app              alive             belapa               chef
     -0.051696746       -0.063771437       -0.054381761       -0.026522465
          ajadah              photo               fine                 rt
     -0.054381761       -0.029119708       -0.051696746        0.233647500
           iyaaaa             tumblr                 lo            working
      0.276513418       -0.029119708        0.276513418       -0.051696746
               tv                its              tweet             hungry
     -0.063771437       -0.051696746        0.003977118       -0.026522465
              lol                jan               feel             gituuu
     -0.023373500       -0.063771437       -0.023373500        0.276513418
             much             online           renyfebry                sih
     -0.019640518        0.276513418        0.276513418        0.276513418
            mandi            ngapain              shows              wants
     -0.054381761        0.276513418       -0.063771437        0.003977118
         cekalang  kissmytweetmuah               http                get
     -0.054381761       -0.023373500       -0.044330382       -0.063771437
              nif               ipod              males               haha
      0.276513418       -0.051696746       -0.054381761       -0.051696746
         hanifpeng         abduction           tumblr.1     chalk_flew_up
      0.276513418       -0.063771437       -0.051696746       -0.019640518
           nantian               name          bsstracing               jam
     -0.054381761       -0.026522465       -0.063771437       -0.054381761
            stadd               word             longer             salad
     -0.011984403       -0.023373500       -0.019640518       -0.026522465
        mandinyah              maybe       biebercrombie              akuh
     -0.054381761       -0.019640518       -0.051696746       -0.054381761
             html               seco            calling                 u
     -0.063771437       -0.029119708       -0.026522465       -0.023373500
           jungle                yup                 im               com
     -0.063771437       -0.019640518       -0.059173674       -0.044330382
      misslovelace_             thats
       0.003977118       -0.023373500
```

We compute the first principal component score for each word.

> predict(data2_pca)[,1]

[1] -1.8409993 -0.6994286 -2.2709973 -1.9366167   0.1416312 -0.4267827
 [7]   9.8470610 -0.9445051 -0.8323657 -1.0369968

Barplot of the variances explained by the principal components.

> plot(data2_pca)

Biplot of the (scaled) first two principal components.

> biplot(data2_pca, col = c("gray", "black"))