SA2 BONUS

Ablian, Andrei Jon A., Cuerdo, Naomi Hannah A., Percia, Kyte Daiter M.

2025-05-20

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.4
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(ggplot2)
```

PCA Analysis for New York Times Articles (TF - IDF)

The dataset contains TF-IDF-normalized word frequencies for a collection of New York Times articles. Each row represents an article, and each column (except type) corresponds to the TF-IDF score of a specific word. The type column indicates the article's category (e.g., "art" or "music"). The data is suitable for text analysis and dimensionality reduction using techniques like Principal Component Analysis (PCA) to explore patterns and differences in word usage across article types.

```
df <- read.csv("C:\\Users\\naomi\\Downloads\\nyt_articles.csv")
str(df)
```

```
## 'data.frame':	102 obs. of  4432 variables:
##  $ class.labels    : chr  "art" "art" "art" "art" ...
##  $ X.              : num  0.00871 0.00585 0.01604 0.02641 0.00729 ...
##  $ X.d             : num  0 0 0 0 0 ...
##  $ X.nd            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ X.s             : num  0 0 0.0114 0 0.011 ...
##  $ X.th            : num  0.00925 0 0 0 0 ...
##  $ X.this          : num  0 0 0 0 0 ...
##  $ a               : num  0.00756 0.00142 0.01006 0.00868 0.00839 ...
##  $ abandoned       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ abc             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ability         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ able            : num  0 0.0399 0 0 0 ...
##  $ about           : num  0.0533 0 0 0.0125 0 ...
##  $ above           : num  0 0 0.0536 0 0 ...
##  $ abroad          : num  0 0 0 0.041 0 ...
##  $ absorbed        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ absorbing       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ abstract        : num  0.0216 0.0435 0 0 0 ...
##  $ abstraction     : num  0 0 0 0 0 ...
##  $ abstractions    : num  0 0 0 0 0 ...
##  $ abundance       : num  0.0349 0 0 0 0 ...
##  $ academic        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ academy         : num  0 0.273 0 0 0 ...
##  $ accents         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ accept          : num  0 0 0 0 0 ...
##  $ access          : num  0 0 0 0 0 ...
##  $ accessible      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ acclaimed       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ accommodate     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ accompanied     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ accompanying    : num  0.0268 0 0 0 0 ...
##  $ according       : num  0 0 0 0.0736 0 ...
##  $ accordingly     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ account         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ accounted       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ accused         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ achieved        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ achievement     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ acknowledge     : num  0 0 0 0.041 0.0438 ...
##  $ acknowledged    : num  0 0 0 0.0315 0 ...
##  $ acquired        : num  0 0 0.0348 0 0 ...
##  $ acquisition     : num  0 0 0 0 0 ...
##  $ acquisitions    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ acre            : num  0 0 0.0454 0 0 ...
##  $ across          : num  0 0 0.02 0 0 ...
##  $ acrylics        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ act             : num  0.0198 0 0 0 0 ...
##  $ acted           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ acting          : num  0 0 0 0 0 ...
##  $ action          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ actions         : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ active          : num  0.0349 0 0 0 0 ...
##  $ activities      : num  0 0 0 0 0 ...
##  $ actor           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ actors          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ actress         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ acts            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ actually        : num  0.0198 0 0 0 0.0248 ...
##  $ adam            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adams           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adamss          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adaptation      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ add             : num  0 0 0 0.0736 0 ...
##  $ added           : num  0 0 0 0 0.0443 ...
##  $ adding          : num  0 0 0 0 0 ...
##  $ addition        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ additional      : num  0 0 0 0 0 ...
##  $ address         : num  0.0349 0 0 0 0 ...
##  $ addresses       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adds            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adhering        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adjacent        : num  0 0 0.0454 0 0 ...
##  $ administration  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ admired         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ admission       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ admits          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adopted         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ads             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adults          : num  0 0 0 0 0.0361 ...
##  $ advance         : num  0 0 0 0 0 ...
##  $ advanced        : num  0 0 0 0 0.0438 ...
##  $ advantage       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adventure       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adventurous     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ advertisements  : num  0 0 0 0 0.0438 ...
##  $ advertising     : num  0 0 0 0 0.118 ...
##  $ advice          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ advised         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adviser         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ advising        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ advocates       : num  0 0 0 0.041 0 ...
##  $ aesthetic       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ affair          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ affairs         : num  0 0 0 0.101 0 ...
##  $ affect          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ affected        : num  0.0627 0 0 0 0 ...
##  $ affection       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ afford          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ afraid          : num  0 0 0 0 0 0 0 0 0 0 ...
##   [list output truncated]
```

The summary above show the variables and observation of the dataset. The dataset has 102 observations and 4,432 variables.

We now then have to remove the non-numeric or irrelevant columns (assuming ‘type is the label column’) and zero-variance columns

```
article_type <- df$type
df_features <- df %>% select(-type)


df_features_numeric <- df_features %>% mutate(across(everything(), ~as.numeric(.)))
```

```
## Warning: There was 1 warning in `mutate()`.
## ℹ In argument: `across(everything(), ~as.numeric(.))`.
## Caused by warning:
## ! NAs introduced by coercion
```

```
df_features_numeric <- df_features_numeric[, colSums(!is.na(df_features_numeric)) > 0]


nzv_cols <- sapply(df_features_numeric, function(col) var(col, na.rm = TRUE) > 0)
df_features_numeric <- df_features_numeric[, nzv_cols]


df_features_numeric <- na.omit(df_features_numeric)

article_type <- article_type[as.numeric(rownames(df_features_numeric))]
```

Now, we are ready for the PCA analysis proper.

```
pca_scaled <- prcomp(df_features_numeric, center = TRUE, scale. = TRUE)

summary(pca_scaled)
```

```
## Importance of components:
##                             PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     10.02711 8.83791 8.70079 8.56437 8.39037 8.30627 8.23678
## Proportion of Variance  0.02271 0.01764 0.01710 0.01656 0.01590 0.01558 0.01532
## Cumulative Proportion   0.02271 0.04035 0.05744 0.07401 0.08991 0.10549 0.12081
##                            PC8     PC9    PC10    PC11   PC12    PC13    PC14
## Standard deviation     8.14085 8.06555 8.01430 7.96833 7.8742 7.83155 7.78354
## Proportion of Variance 0.01497 0.01469 0.01451 0.01434 0.0140 0.01385 0.01368
## Cumulative Proportion  0.13578 0.15047 0.16497 0.17931 0.1933 0.20716 0.22085
##                           PC15    PC16    PC17   PC18    PC19    PC20    PC21
## Standard deviation     7.74893 7.64753 7.60092 7.5582 7.49017 7.47999 7.46155
## Proportion of Variance 0.01356 0.01321 0.01305 0.0129 0.01267 0.01264 0.01257
## Cumulative Proportion  0.23441 0.24762 0.26066 0.2736 0.28623 0.29887 0.31144
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     7.42031 7.36110 7.36099 7.29801 7.24208 7.22373 7.20321
## Proportion of Variance 0.01243 0.01224 0.01224 0.01203 0.01184 0.01178 0.01172
## Cumulative Proportion  0.32388 0.33611 0.34835 0.36038 0.37222 0.38401 0.39573
##                           PC29    PC30    PC31    PC32    PC33    PC34    PC35
## Standard deviation     7.15370 7.14271 7.12152 7.10166 7.05704 6.99153 6.98176
## Proportion of Variance 0.01156 0.01152 0.01145 0.01139 0.01125 0.01104 0.01101
## Cumulative Proportion  0.40728 0.41881 0.43026 0.44165 0.45290 0.46393 0.47494
##                           PC36    PC37    PC38    PC39    PC40    PC41    PC42
## Standard deviation     6.95750 6.91884 6.86616 6.83073 6.80027 6.76417 6.75531
## Proportion of Variance 0.01093 0.01081 0.01065 0.01054 0.01044 0.01033 0.01031
## Cumulative Proportion  0.48587 0.49669 0.50733 0.51787 0.52831 0.53865 0.54895
##                           PC43    PC44    PC45    PC46    PC47    PC48    PC49
## Standard deviation     6.68411 6.66343 6.63219 6.61655 6.59921 6.58135 6.56005
## Proportion of Variance 0.01009 0.01003 0.00993 0.00989 0.00984 0.00978 0.00972
## Cumulative Proportion  0.55904 0.56907 0.57900 0.58889 0.59872 0.60851 0.61823
##                           PC50    PC51    PC52    PC53   PC54    PC55    PC56
## Standard deviation     6.53292 6.50285 6.47720 6.45645 6.4162 6.39226 6.36251
## Proportion of Variance 0.00964 0.00955 0.00947 0.00941 0.0093 0.00923 0.00914
## Cumulative Proportion  0.62786 0.63741 0.64689 0.65630 0.6656 0.67483 0.68397
##                           PC57    PC58    PC59    PC60    PC61    PC62    PC63
## Standard deviation     6.33970 6.31055 6.26798 6.25158 6.24543 6.22232 6.21414
## Proportion of Variance 0.00908 0.00899 0.00887 0.00883 0.00881 0.00874 0.00872
## Cumulative Proportion  0.69305 0.70204 0.71091 0.71974 0.72855 0.73729 0.74601
##                           PC64    PC65    PC66   PC67    PC68    PC69    PC70
## Standard deviation     6.18828 6.16359 6.14565 6.0993 6.08588 6.04540 6.02312
## Proportion of Variance 0.00865 0.00858 0.00853 0.0084 0.00836 0.00825 0.00819
## Cumulative Proportion  0.75466 0.76324 0.77177 0.7802 0.78853 0.79679 0.80498
##                           PC71    PC72    PC73   PC74    PC75    PC76    PC77
## Standard deviation     5.99227 5.98371 5.90079 5.8764 5.84273 5.82193 5.80377
## Proportion of Variance 0.00811 0.00809 0.00786 0.0078 0.00771 0.00765 0.00761
## Cumulative Proportion  0.81309 0.82118 0.82904 0.8368 0.84455 0.85220 0.85981
##                           PC78    PC79    PC80    PC81    PC82    PC83    PC84
## Standard deviation     5.74554 5.69275 5.68180 5.64362 5.57634 5.53235 5.50532
## Proportion of Variance 0.00746 0.00732 0.00729 0.00719 0.00702 0.00691 0.00684
## Cumulative Proportion  0.86726 0.87458 0.88187 0.88907 0.89609 0.90300 0.90985
##                           PC85    PC86    PC87    PC88    PC89    PC90    PC91
## Standard deviation     5.47152 5.41465 5.41122 5.39074 5.29275 5.26054 5.22600
## Proportion of Variance 0.00676 0.00662 0.00661 0.00656 0.00633 0.00625 0.00617
```

```
## Cumulative Proportion  0.91661 0.92323 0.92984 0.93640 0.94273 0.94898 0.95515
##                           PC92    PC93    PC94    PC95    PC96    PC97    PC98
## Standard deviation     5.1547 5.03510 4.99400 4.90218 4.75703 4.34636 4.22475
## Proportion of Variance 0.0060 0.00573 0.00563 0.00543 0.00511 0.00427 0.00403
## Cumulative Proportion  0.9611 0.96687 0.97251 0.97793 0.98304 0.98731 0.99134
##                           PC99   PC100   PC101     PC102
## Standard deviation     3.86143 3.46998 3.37492 1.245e-14
## Proportion of Variance 0.00337 0.00272 0.00257 0.000e+00
## Cumulative Proportion  0.99471 0.99743 1.00000 1.000e+00
```

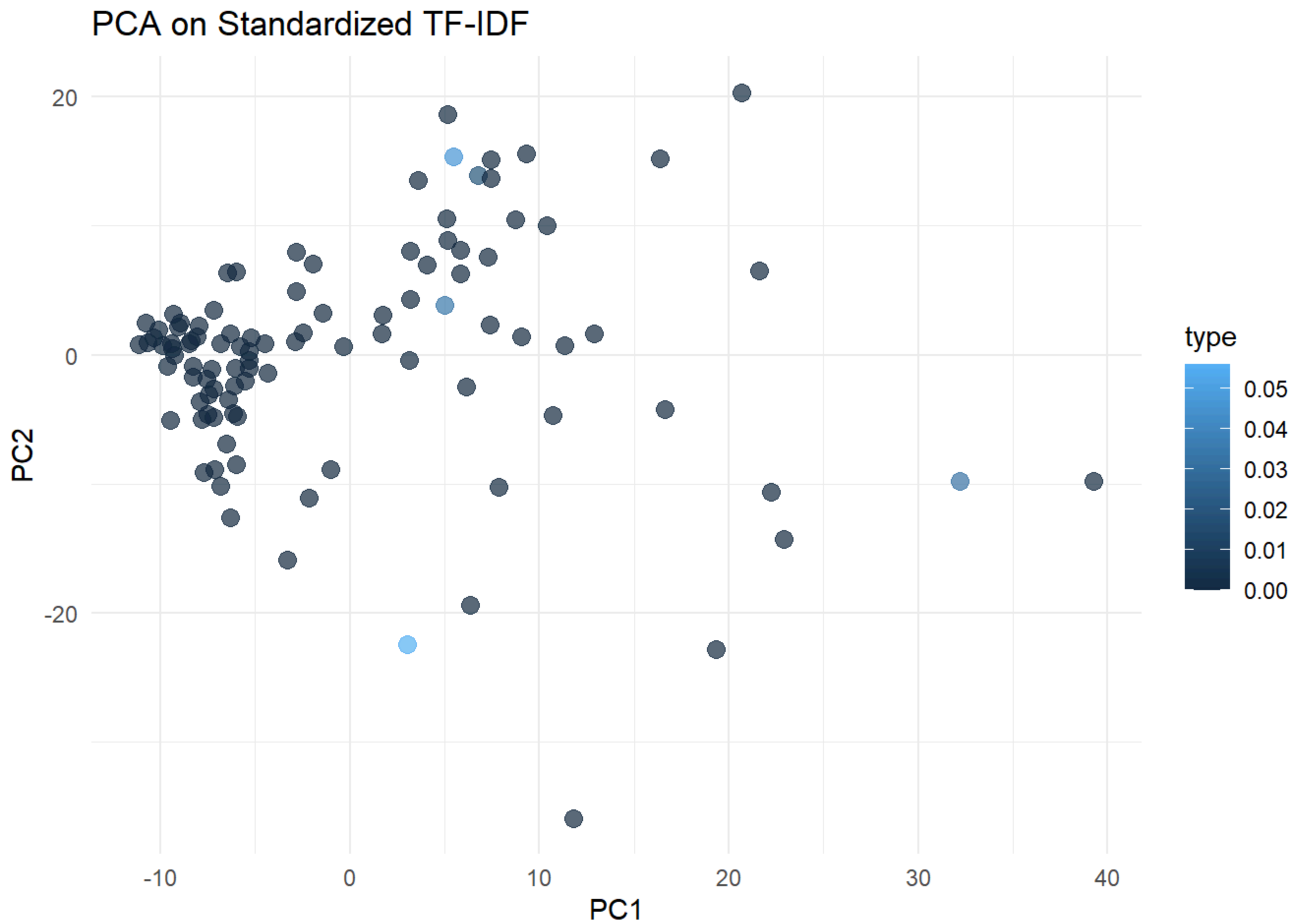
The principal component analysis (PCA) was performed on a dataset with 102 observations and 4432 variables. The first principal component (PC1) has the highest standard deviation (10.03) but explains only about 2.27% of the total variance. Subsequent components explain progressively smaller proportions of variance, with the first 10 components cumulatively accounting for approximately 16.5% of the variance. The gradual increase in cumulative variance suggests that many components are needed to capture most of the data's variability, indicating a complex, high-dimensional dataset.

Plots

Plotting the Scaled PCA:

```
df_pca_scaled <- as.data.frame(pca_scaled$x)
df_pca_scaled$type <- article_type

ggplot(df_pca_scaled, aes(PC1, PC2, color = type)) +
  geom_point(size = 3, alpha = 0.7) +
  labs(title = "PCA on Standardized TF-IDF", x = "PC1", y = "PC2") +
  theme_minimal()
```



The scatter plot displays the first two principal components (PC1 and PC2) from the standardized TF-IDF matrix. Each point represents an article, colored by its article type. From the graph, the points clustered together (especially near the origin) are more similar based on their TF-IDF features.

Points farther out (e.g., around PC1 = 40) are likely outliers or documents with unique term usage.

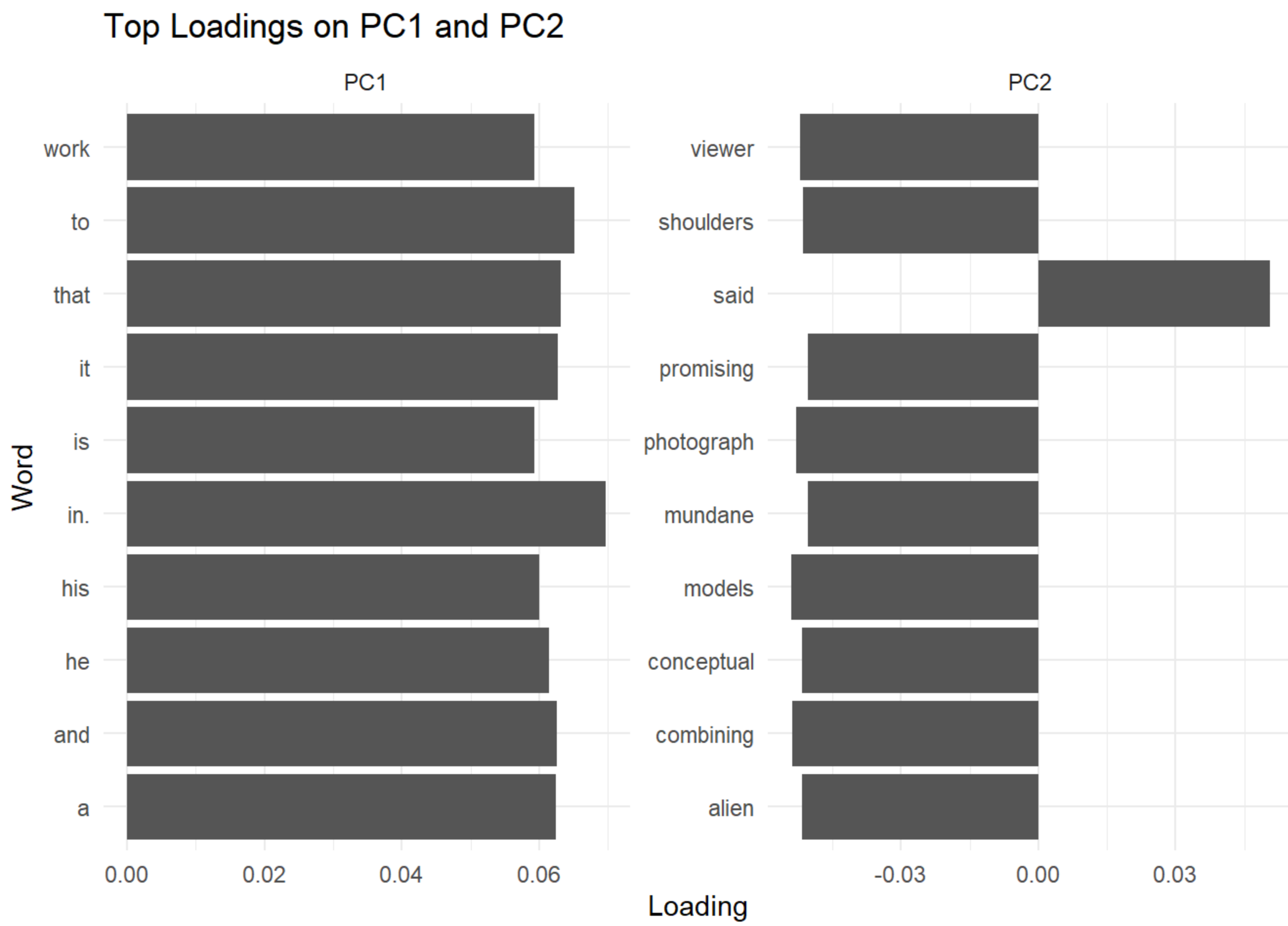
Now we determine the top loadings for visualization:

```
top_loadings <- function(rotation_matrix, pc = 1, n = 10) {
  loadings <- rotation_matrix[, pc]
  idx <- order(abs(loadings), decreasing = TRUE)[1:n]
  tibble(word = names(loadings)[idx], loading = loadings[idx])
}
top_pc1 <- top_loadings(pca_scaled$rotation, pc = 1, n = 10)
top_pc2 <- top_loadings(pca_scaled$rotation, pc = 2, n = 10)
```

Combining and plot vectors for a subset biplot to check the top loadings of the dataset:

```
biplot_data <- rbind(
  top_pc1 %>% mutate(PC = "PC1"),
  top_pc2 %>% mutate(PC = "PC2")
)

ggplot(biplot_data, aes(x = loading, y = word)) +
  geom_col() +
  facet_wrap(~PC, scales = "free") +
  labs(title = "Top Loadings on PC1 and PC2", x = "Loading", y = "Word") +
  theme_minimal()
```



The chart shows the top words influencing the first two principal components (PC1 and PC2) from a PCA on TF-IDF text data. PC1 is dominated by common words (e.g., “work”, “that”, “he”), likely reflecting general language usage. PC2 highlights more content-specific terms (e.g., “said”, “viewer”, “conceptual”), suggesting it captures thematic or topical variation across documents.

Results and Discussion

From the results above, we can say that the standardized variables makes a difference when the variables have different scales of variation, which is almost always the case in TF-IDF matrices. Furthermore, it gives better separation by article type, as can be seein the PC1 and PC2 plot. Thus, we can say that **the standardized PCA is more useful.** It allows intepretable inspection via loadings, which are key words driving variance.