Research Title Analysis (KeyATM)

Charles J. Gomez, Harshvardhan Singh

2023-06-13

Loading data for quanteda

```
# Load the library
library(keyATM)
```

```
## keyATM 0.5.0 successfully loaded.
##  Papers, examples, resources, and other materials are at
##  https://keyatm.github.io/keyATM/
```

```
library(quanteda)
```

```
## Package version: 3.3.1
## Unicode version: 14.0
## ICU version: 70.1
```

```
## Parallel computing: 8 of 8 threads used.
```

```
## See https://quanteda.io for tutorials and examples.
```

```
library(readtext)
```

```
##
## Attaching package: 'readtext'
```

```
## The following object is masked from 'package:quanteda':
##
##     texts
```

```
# Read the CSV file
data <- read.csv("INPUT_SQL_Text_Data_Astronomy_and_Astrophysics.csv")
text_data <- data$title

# Preprocessing
key_corpus <- corpus(text_data)

# Covariate data
covariate_data <- data[, c("publication_year", "country")]
key_corpus <- corpus(text_data, docvars = covariate_data)
#docvars(key_corpus, c("publication_year", "country")) <- covariate_data

#Creating token object
key_token <- tokens(key_corpus)

# Createing a document-feature matrix (dfm object) from the token object
key_dfm <- dfm(key_token)
```

Preprocessing data

```
library(keyATM)
library(quanteda)
library(magrittr)

#remove punctuations and unnecessary characters
data_tokens <- tokens(
    data$title,
    remove_numbers = FALSE,
    remove_punct = TRUE,
    remove_symbols = TRUE,
    remove_separators = TRUE,
    remove_url = TRUE
  ) %>%
  tokens_tolower() %>%    #converts all characters into lower cases
  tokens_remove(
    c(stopwords("english"),
      "may", "shall", "can",
      "must", "upon", "with", "without"
    )
  ) %>%
  tokens_select(min_nchar = 3)

#Before loading data into the keyATM, construct a document-feature matrix (dfm object)
data_dfm <- dfm(data_tokens) %>%
              dfm_trim(min_termfreq = 5, min_docfreq = 2)
ncol(data_dfm)  # the number of unique words
```

```
## [1] 6327
```

```
# Filter out documents with length 0
#data_dfm <- data_dfm[ndoc(data_dfm) > 0, ]
data_dfm <- dfm_subset(data_dfm, ntoken(data_dfm) > 0)

# Read the document-feature matrix using keyATM_read()
keyATM_docs <- keyATM_read(texts = data_dfm)
```

```
## i Using quanteda dfm.
```

```
## Loading documents ========>---------------------- 26% | ETA: 3s
## Loading documents =========>--------------------- 28% | ETA: 3s
## Loading documents =========>--------------------- 31% | ETA: 3s
## Loading documents ==========>-------------------- 33% | ETA: 4s
## Loading documents ==========>-------------------- 35% | ETA: 4s
## Loading documents ===========>------------------- 36% | ETA: 4s
## Loading documents ===========>------------------- 38% | ETA: 4s
## Loading documents ============>------------------ 40% | ETA: 4s
## Loading documents =============>----------------- 42% | ETA: 4s
## Loading documents =============>----------------- 43% | ETA: 4s
## Loading documents =============>----------------- 45% | ETA: 4s
## Loading documents ==============>---------------- 46% | ETA: 4s
## Loading documents ==============>---------------- 48% | ETA: 4s
## Loading documents ===============>--------------- 49% | ETA: 4s
## Loading documents ===============>--------------- 50% | ETA: 4s
```

```
## Loading documents ================>-------------- 52% | ETA: 4s
## Loading documents ================>-------------- 53% | ETA: 4s
## Loading documents ================>-------------- 54% | ETA: 4s
## Loading documents =================>------------- 56% | ETA: 4s
## Loading documents =================>------------- 57% | ETA: 4s
## Loading documents =================>------------- 58% | ETA: 4s
## Loading documents ==================>------------ 59% | ETA: 4s
## Loading documents ==================>------------ 60% | ETA: 4s
## Loading documents ==================>------------ 61% | ETA: 4s
## Loading documents ===================>----------- 62% | ETA: 4s
## Loading documents ===================>----------- 63% | ETA: 4s
## Loading documents ===================>----------- 64% | ETA: 4s
## Loading documents ===================>----------- 65% | ETA: 4s
## Loading documents ====================>---------- 66% | ETA: 4s
## Loading documents ====================>---------- 67% | ETA: 4s
## Loading documents ====================>---------- 68% | ETA: 3s
## Loading documents =====================>--------- 69% | ETA: 3s
## Loading documents =====================>--------- 70% | ETA: 3s
## Loading documents =====================>--------- 70% | ETA: 3s
## Loading documents =====================>--------- 71% | ETA: 3s
## Loading documents ======================>-------- 72% | ETA: 3s
## Loading documents ======================>-------- 73% | ETA: 3s
## Loading documents ======================>-------- 74% | ETA: 3s
## Loading documents ======================>-------- 75% | ETA: 3s
## Loading documents =======================>------- 76% | ETA: 3s
## Loading documents =======================>------- 76% | ETA: 3s
## Loading documents =======================>------- 76% | ETA: 3s
## Loading documents =======================>------- 77% | ETA: 3s
## Loading documents =======================>------- 78% | ETA: 3s
## Loading documents ========================>------ 79% | ETA: 3s
## Loading documents ========================>------ 80% | ETA: 3s
## Loading documents ========================>------ 80% | ETA: 3s
## Loading documents ========================>------ 81% | ETA: 2s
## Loading documents =========================>----- 82% | ETA: 2s
## Loading documents =========================>----- 83% | ETA: 2s
## Loading documents =========================>----- 84% | ETA: 2s
## Loading documents =========================>----- 84% | ETA: 2s
## Loading documents ==========================>---- 85% | ETA: 2s
## Loading documents ==========================>---- 86% | ETA: 2s
## Loading documents ==========================>---- 87% | ETA: 2s
## Loading documents ==========================>---- 87% | ETA: 2s
## Loading documents ==========================>---- 88% | ETA: 2s
## Loading documents ===========================>--- 89% | ETA: 2s
## Loading documents ===========================>--- 90% | ETA: 1s
## Loading documents ===========================>--- 91% | ETA: 1s
## Loading documents ===========================>--- 91% | ETA: 1s
## Loading documents ============================>-- 92% | ETA: 1s
## Loading documents ============================>-- 93% | ETA: 1s
## Loading documents ============================>-- 93% | ETA: 1s
## Loading documents ============================>-- 94% | ETA: 1s
## Loading documents ============================>-- 95% | ETA: 1s
## Loading documents =============================>- 96% | ETA: 1s
## Loading documents =============================>- 96% | ETA: 1s
## Loading documents =============================>- 97% | ETA: 0s
```

```
## Loading documents ============================>- 98% | ETA: 0s
## Loading documents ============================>- 98% | ETA: 0s Loading
## documents ============================> 99% | ETA: 0s Loading documents
## ============================> 100% | ETA: 0s
```

```
# Summary of keyATM_docs
summary(keyATM_docs)
```

```
## keyATM_docs object of 63982 documents.
## * Average (min/max) document length: 7.19 (1/22) words
##
## * Number of unique words: 6327
```

Preparing keywords

```
keywords <- list(
 Topic1 = c("burst", "gamma-ray", "quasar", "line", "discoveri"),
 Topic2 = c("variabl", "studi", "ngc", "galaxi", "ultraviolet"),
 Topic3 = c("solar", "use", "measur", "infrar", "variat"),
 Topic4 = c("binari", "hole", "black", "accret", "system"),
 Topic5 = c("dwarf", "planet", "star", "disk", "around"),
 Topic6 = c("star", "cluster", "format", "massiv", "globular"),
 Topic7 = c("supernova", "type", "remnant", "spectra", "light"),
 Topic8 = c("nebula", "planetari", "wind", "young", "outflow"),
 Topic9 = c("galaxi", "redshift", "survey", "sky", "high"),
 Topic10 = c("galaxi", "structur", "gas", "spiral", "local"),
 Topic11 = c("model", "distribut", "univers", "energi", "cosmic"),
 Topic12 = c("dust", "abund", "interstellar", "comet", "chemic"),
 Topic13 = c("magnet", "field", "solar", "coron", "shock"),
 Topic14 = c("gravit", "effect", "lens", "background", "cosmic"),
 Topic15 = c("x-ray", "radio", "emiss", "pulsar", "sourc"),
 Topic16 = c("galact", "activ", "stellar", "region", "luminos"),
 Topic17 = c("observ", "cloud", "chandra", "magellan", "molecular"),
 Topic18 = c("survey", "telescop", "imag", "space", "hubbl"),
 Topic19 = c("dark", "halo", "matter", "galaxi", "mass"),
 Topic20 = c("new", "evid", "origin", "rotat", "dynam")
)
```

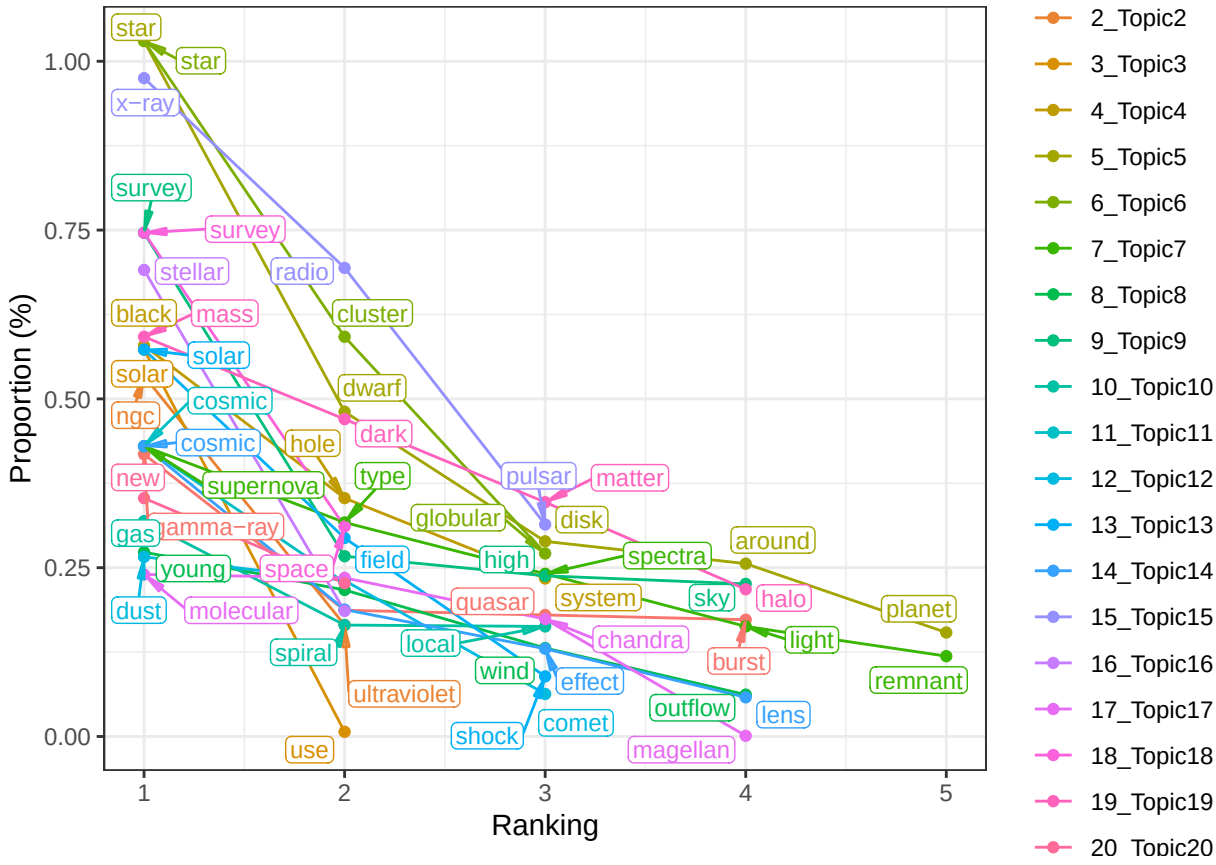
Checking Keywords

```
key_viz <- visualize_keywords(docs = keyATM_docs, keywords = keywords)
```

```
## Warning: Keywords are pruned because they do not appear in the documents: discoveri,
## variabl, studi, galaxi, measur, infrar, variat, binari, accret, format, massiv,
## planetari, galaxi, galaxi, structur, distribut, univers, energi, ..., rotat,
## and dynam
```

```
key_viz
```

```
## Warning: ggrepel: 9 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
#save figure
key_viz <- visualize_keywords(docs = keyATM_docs, keywords = keywords)
```

```
## Warning: Keywords are pruned because they do not appear in the documents: discoveri,
## variabl, studi, galaxi, measur, infrar, variat, binari, accret, format, massiv,
## planetari, galaxi, galaxi, structur, distribut, univers, energi, ..., rotat,
## and dynam
```

```
save_fig(key_viz, "/Users/harshvardhansingh/Documents/keyword.pdf", width = 6.5, height = 4)
```

```
## Warning: ggrepel: 12 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

```
#get the actual values
values_fig(key_viz)
```

```
## # A tibble: 64 x 5
## # Groups:   Topic [20]
##    Word        WordCount `Proportion(%)` Ranking Topic
##    <chr>           <int>           <dbl>   <int> <fct>
##  1 gamma-ray        1924           0.418       1 1_Topic1
##  2 line              859           0.187       2 1_Topic1
##  3 quasar            828           0.18        3 1_Topic1
##  4 burst             796           0.173       4 1_Topic1
##  5 ngc              2455           0.534       1 2_Topic2
##  6 ultraviolet       760           0.165       2 2_Topic2
##  7 solar            2634           0.573       1 3_Topic3
##  8 use                33           0.007       2 3_Topic3
```

```
##  9 black            2662           0.579       1 4_Topic4
## 10 hole             1625           0.353       2 4_Topic4
## # i 54 more rows
```

Preparing Covariates

```
vars <- docvars(key_dfm)
head(vars)
```

```
##   publication_year
## 1             2004
## 2             1991
## 3             2003
## 4             2003
## 5             2002
## 6             1999
##
## 1
## 2
## 3
## 4
## 5 DE+middle US+middle US+last US+middle DE+middle US+middle DE+middle US+middle US+middle US+first US
## 6
```