STM\_Model Great Britain to Africa k-value analysis

Charlie J. Gomez, Harshvardhan Singh

2023-08-12

library('stm')

## stm v1.3.6 successfully loaded. See ?stm for help.   
## Papers, resources, and other materials at structuraltopicmodel.com

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(stringr)  
library(wordcloud)

## Loading required package: RColorBrewer

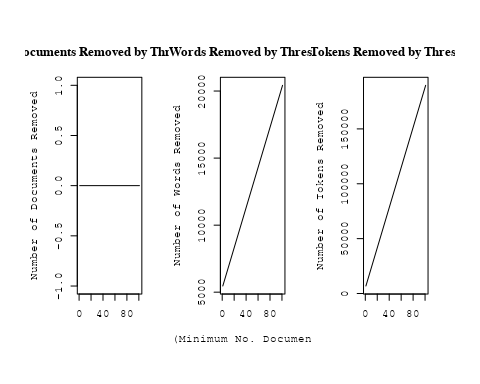
#Read csv file  
data = read.csv("preprocessed\_data\_Jul14.csv")  
  
  
##Topic generation for GB (in collaboration) publications  
  
data\_collab <- data[data[["GB"]] != 0,]  
  
# Save the original title data for future use  
data\_collab$original\_concatenated\_title\_abstract <- data\_collab$concatenated\_title\_abstract  
  
#pre-processing the titles using textProcessor from the stm package  
processed\_text <- textProcessor(data\_collab$concatenated\_title\_abstract, metadata = data\_collab)

## Building corpus...   
## Converting to Lower Case...   
## Removing punctuation...   
## Removing stopwords...   
## Removing numbers...   
## Stemming...   
## Creating Output...

# Further prepare the data by removing low-frequency terms  
out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta)

## Removing 5462 of 21718 terms (5462 of 775142 tokens) due to frequency   
## Your corpus now has 9054 documents, 16256 terms and 769680 tokens.

docs\_text <- out\_text$documents  
vocab\_text <- out\_text$vocab  
meta\_text <- out\_text$meta  
  
  
#Prepare data  
plotRemoved(processed\_text$documents, lower.thresh = seq(1, 200, by = 100))



out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta, lower.thresh = 5)

## Removing 15774 of 21718 terms (34858 of 775142 tokens) due to frequency   
## Your corpus now has 9054 documents, 5944 terms and 740284 tokens.

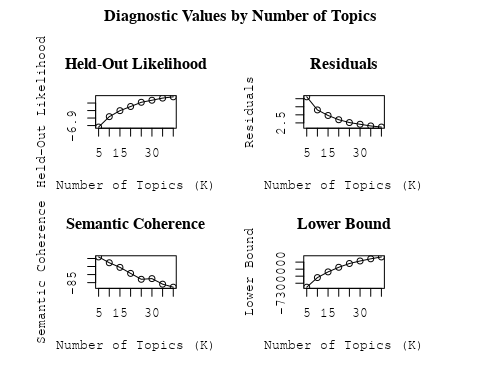
str(out\_text$meta)

## 'data.frame': 9054 obs. of 39 variables:  
## $ concept\_id : chr "https://openalex.org/C1276947" "https://openalex.org/C44870925" "https://openalex.org/C44870925" "https://openalex.org/C44870925" ...  
## $ work\_id : chr "https://openalex.org/W2031560129" "https://openalex.org/W2091953534" "https://openalex.org/W2155870677" "https://openalex.org/W2030196560" ...  
## $ publication\_year : int 2010 2004 2010 2009 2002 1998 2021 2009 1993 2011 ...  
## $ title : chr "Lunar Palaeoregolith Deposits as Recorders of the Galactic Environment of the Solar System and Implications for Astrobiology" "Inner Polar Rings in Regular Lenticular Galaxies" "The Detection of a Population of Submillimeter-Bright, Strongly Lensed Galaxies" "The Effect of Stellar Rotation on Colour-Magnitude Diagrams: On the apparent presence of multiple populations i"| \_\_truncated\_\_ ...  
## $ paperabstract : chr "One of the principal scientific reasons for wanting to resume in situ exploration of the lunar surface is to ga"| \_\_truncated\_\_ "We have investigated a sample of S0 galaxies, mostly with circumnuclear dust lanes orthogonal to their major ax"| \_\_truncated\_\_ "Through a Lens Brightly Astronomical sources detected in the submillimeter range are generally thought to be di"| \_\_truncated\_\_ "A significant number of intermediate age clusters (1 − 2 Gyr) in the Magellanic Clouds appear to have multiple "| \_\_truncated\_\_ ...  
## $ country : chr "GB GB US GB US GB" "RU RU GB" "US GB US GB" "GB NL GB NL" ...  
## $ year\_concept : chr "2010+https://openalex.org/C1276947" "2004+https://openalex.org/C44870925" "2010+https://openalex.org/C44870925" "2009+https://openalex.org/C44870925" ...  
## $ concatenated\_title\_abstract : chr "Lunar Palaeoregolith Deposits as Recorders of the Galactic Environment of the Solar System and Implications for"| \_\_truncated\_\_ "Inner Polar Rings in Regular Lenticular Galaxies We have investigated a sample of S0 galaxies, mostly with circ"| \_\_truncated\_\_ "The Detection of a Population of Submillimeter-Bright, Strongly Lensed Galaxies Through a Lens Brightly Astrono"| \_\_truncated\_\_ "The Effect of Stellar Rotation on Colour-Magnitude Diagrams: On the apparent presence of multiple populations i"| \_\_truncated\_\_ ...  
## $ US : num 33.3 0 50 0 50 ...  
## $ IN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DE : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CH : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ GB : num 66.7 33.3 50 50 50 ...  
## $ CN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ FR : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IT : num 0 0 0 0 0 ...  
## $ RU : num 0 66.7 0 0 0 ...  
## $ CA : num 0 0 0 0 0 0 0 0 0 25 ...  
## $ NL : num 0 0 0 50 0 0 0 50 0 0 ...  
## $ AU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ JP : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ES : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Americas : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Europe : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Africa : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AsiaAndOceania : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2020\_2022 : int 0 0 0 0 0 0 1 0 0 0 ...  
## $ pub\_interval\_2015\_2019 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2010\_2014 : int 1 0 1 0 0 0 0 0 0 1 ...  
## $ pub\_interval\_2005\_2009 : int 0 0 0 1 0 0 0 1 0 0 ...  
## $ pub\_interval\_2000\_2004 : int 0 1 0 0 1 0 0 0 0 0 ...  
## $ pub\_interval\_1995\_1999 : int 0 0 0 0 0 1 0 0 0 0 ...  
## $ pub\_interval\_1985\_1994 : int 0 0 0 0 0 0 0 0 1 0 ...  
## $ pub\_interval\_1975\_1984 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1965\_1974 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1900\_1964 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1824\_1899 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ original\_concatenated\_title\_abstract: chr "Lunar Palaeoregolith Deposits as Recorders of the Galactic Environment of the Solar System and Implications for"| \_\_truncated\_\_ "Inner Polar Rings in Regular Lenticular Galaxies We have investigated a sample of S0 galaxies, mostly with circ"| \_\_truncated\_\_ "The Detection of a Population of Submillimeter-Bright, Strongly Lensed Galaxies Through a Lens Brightly Astrono"| \_\_truncated\_\_ "The Effect of Stellar Rotation on Colour-Magnitude Diagrams: On the apparent presence of multiple populations i"| \_\_truncated\_\_ ...

# Initialize an empty formula string  
prevalence\_formula\_str <- "~"  
  
# Define the publication intervals  
pub\_intervals <- c("pub\_interval\_2020\_2022", "pub\_interval\_2015\_2019", "pub\_interval\_2010\_2014",   
 "pub\_interval\_2005\_2009", "pub\_interval\_2000\_2004", "pub\_interval\_1995\_1999",  
 "pub\_interval\_1985\_1994", "pub\_interval\_1975\_1984", "pub\_interval\_1965\_1974",  
 "pub\_interval\_1900\_1964", "pub\_interval\_1824\_1899")  
  
# Add each publication interval to the formula string  
for (interval in pub\_intervals) {  
 # add an if statement to handle the first addition (without '+')  
 if (prevalence\_formula\_str == "~") {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, interval)  
 } else {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, "+", interval)  
 }  
}  
  
  
# Convert the string to a formula  
prevalence\_formula <- as.formula(prevalence\_formula\_str)  
print(prevalence\_formula)

## ~pub\_interval\_2020\_2022 + pub\_interval\_2015\_2019 + pub\_interval\_2010\_2014 +   
## pub\_interval\_2005\_2009 + pub\_interval\_2000\_2004 + pub\_interval\_1995\_1999 +   
## pub\_interval\_1985\_1994 + pub\_interval\_1975\_1984 + pub\_interval\_1965\_1974 +   
## pub\_interval\_1900\_1964 + pub\_interval\_1824\_1899

Kvals <- seq(from = 5, to = 40, by = 5) # K values to try  
search\_results <- searchK(documents = out\_text$documents,   
 vocab = out\_text$vocab,   
 K = Kvals,   
 prevalence = prevalence\_formula,   
 data = out\_text$meta,   
 init.type = "Spectral",   
 verbose = FALSE)  
  
# Plot the results  
plot(search\_results)



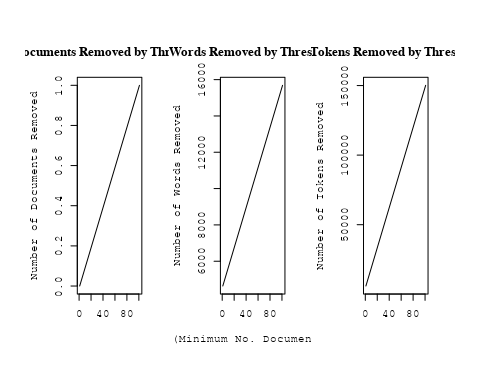
##Topic generation for DE (in collaboration) publications  
  
data\_collab <- data[data[["DE"]] != 0,]  
  
# Save the original title data for future use  
data\_collab$original\_concatenated\_title\_abstract <- data\_collab$concatenated\_title\_abstract  
  
#pre-processing the titles using textProcessor from the stm package  
processed\_text <- textProcessor(data\_collab$concatenated\_title\_abstract, metadata = data\_collab)

## Building corpus...   
## Converting to Lower Case...   
## Removing punctuation...   
## Removing stopwords...   
## Removing numbers...   
## Stemming...   
## Creating Output...

# Further prepare the data by removing low-frequency terms  
out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta)

## Removing 4638 of 16647 terms (4638 of 476346 tokens) due to frequency   
## Your corpus now has 5572 documents, 12009 terms and 471708 tokens.

docs\_text <- out\_text$documents  
vocab\_text <- out\_text$vocab  
meta\_text <- out\_text$meta  
  
  
#Prepare data  
plotRemoved(processed\_text$documents, lower.thresh = seq(1, 200, by = 100))



out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta, lower.thresh = 5)

## Removing 12095 of 16647 terms (25318 of 476346 tokens) due to frequency   
## Your corpus now has 5572 documents, 4552 terms and 451028 tokens.

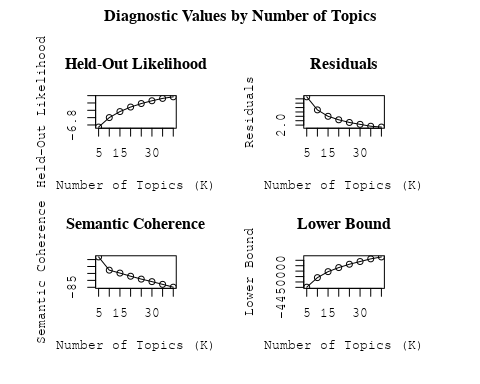
str(out\_text$meta)

## 'data.frame': 5572 obs. of 39 variables:  
## $ concept\_id : chr "https://openalex.org/C44870925" "https://openalex.org/C44870925" "https://openalex.org/C44870925" "https://openalex.org/C1276947" ...  
## $ work\_id : chr "https://openalex.org/W2073719541" "https://openalex.org/W3102133339" "https://openalex.org/W3101291558" "https://openalex.org/W2153493790" ...  
## $ publication\_year : int 1997 2006 2005 2011 1997 2014 2021 2005 2005 2017 ...  
## $ title : chr "Effects of Disks on Gravitational Lensing by Spiral Galaxies" "The hierarchical formation of the brightest cluster galaxies" "The Formation of Fossil Galaxy Groups in the hierarchical Universe" "Effects of the explosion asymmetry and viewing angle on the Type Ia supernova colour and luminosity calibration" ...  
## $ paperabstract : chr "Gravitational lensing of a quasar by a spiral galaxy should often be accompanied by damped Lyα absorption and d"| \_\_truncated\_\_ "We use semi-analytic techniques to study the formation and evolution of brightest cluster galaxies (BCGs). We s"| \_\_truncated\_\_ "We use a set of 12 high-resolution N-body/hydrodynamical simulations in the ΛCDM cosmology to investigate the o"| \_\_truncated\_\_ "Phenomenological relations exist between the peak luminosity and other observables of type Ia supernovae (SNe I"| \_\_truncated\_\_ ...  
## $ country : chr "US DE" "DE DE DE DE" "GB DE" "JP DE JP DE" ...  
## $ year\_concept : chr "1997+https://openalex.org/C44870925" "2006+https://openalex.org/C44870925" "2005+https://openalex.org/C44870925" "2011+https://openalex.org/C1276947" ...  
## $ concatenated\_title\_abstract : chr "Effects of Disks on Gravitational Lensing by Spiral Galaxies Gravitational lensing of a quasar by a spiral gala"| \_\_truncated\_\_ "The hierarchical formation of the brightest cluster galaxies We use semi-analytic techniques to study the forma"| \_\_truncated\_\_ "The Formation of Fossil Galaxy Groups in the hierarchical Universe We use a set of 12 high-resolution N-body/hy"| \_\_truncated\_\_ "Effects of the explosion asymmetry and viewing angle on the Type Ia supernova colour and luminosity calibration"| \_\_truncated\_\_ ...  
## $ US : num 50 0 0 0 0 0 0 50 0 0 ...  
## $ IN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DE : num 50 100 50 50 100 100 100 50 50 100 ...  
## $ CH : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ GB : num 0 0 50 0 0 0 0 0 0 0 ...  
## $ CN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ FR : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IT : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ RU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CA : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ NL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AU : num 0 0 0 0 0 0 0 0 50 0 ...  
## $ JP : num 0 0 0 50 0 0 0 0 0 0 ...  
## $ ES : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Americas : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Europe : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Africa : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AsiaAndOceania : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2020\_2022 : int 0 0 0 0 0 0 1 0 0 0 ...  
## $ pub\_interval\_2015\_2019 : int 0 0 0 0 0 0 0 0 0 1 ...  
## $ pub\_interval\_2010\_2014 : int 0 0 0 1 0 1 0 0 0 0 ...  
## $ pub\_interval\_2005\_2009 : int 0 1 1 0 0 0 0 1 1 0 ...  
## $ pub\_interval\_2000\_2004 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1995\_1999 : int 1 0 0 0 1 0 0 0 0 0 ...  
## $ pub\_interval\_1985\_1994 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1975\_1984 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1965\_1974 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1900\_1964 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1824\_1899 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ original\_concatenated\_title\_abstract: chr "Effects of Disks on Gravitational Lensing by Spiral Galaxies Gravitational lensing of a quasar by a spiral gala"| \_\_truncated\_\_ "The hierarchical formation of the brightest cluster galaxies We use semi-analytic techniques to study the forma"| \_\_truncated\_\_ "The Formation of Fossil Galaxy Groups in the hierarchical Universe We use a set of 12 high-resolution N-body/hy"| \_\_truncated\_\_ "Effects of the explosion asymmetry and viewing angle on the Type Ia supernova colour and luminosity calibration"| \_\_truncated\_\_ ...

# Initialize an empty formula string  
prevalence\_formula\_str <- "~"  
  
# Define the publication intervals  
pub\_intervals <- c("pub\_interval\_2020\_2022", "pub\_interval\_2015\_2019", "pub\_interval\_2010\_2014",   
 "pub\_interval\_2005\_2009", "pub\_interval\_2000\_2004", "pub\_interval\_1995\_1999",  
 "pub\_interval\_1985\_1994", "pub\_interval\_1975\_1984", "pub\_interval\_1965\_1974",  
 "pub\_interval\_1900\_1964", "pub\_interval\_1824\_1899")  
  
# Add each publication interval to the formula string  
for (interval in pub\_intervals) {  
 # add an if statement to handle the first addition (without '+')  
 if (prevalence\_formula\_str == "~") {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, interval)  
 } else {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, "+", interval)  
 }  
}  
  
  
# Convert the string to a formula  
prevalence\_formula <- as.formula(prevalence\_formula\_str)  
print(prevalence\_formula)

## ~pub\_interval\_2020\_2022 + pub\_interval\_2015\_2019 + pub\_interval\_2010\_2014 +   
## pub\_interval\_2005\_2009 + pub\_interval\_2000\_2004 + pub\_interval\_1995\_1999 +   
## pub\_interval\_1985\_1994 + pub\_interval\_1975\_1984 + pub\_interval\_1965\_1974 +   
## pub\_interval\_1900\_1964 + pub\_interval\_1824\_1899

Kvals <- seq(from = 5, to = 40, by = 5) # K values to try  
search\_results <- searchK(documents = out\_text$documents,   
 vocab = out\_text$vocab,   
 K = Kvals,   
 prevalence = prevalence\_formula,   
 data = out\_text$meta,   
 init.type = "Spectral",   
 verbose = FALSE)  
  
# Plot the results  
plot(search\_results)



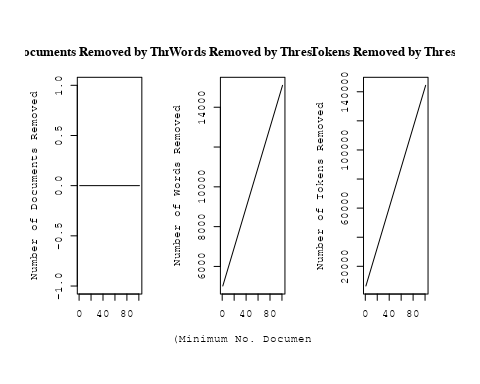
##Topic generation for Europe (in collaboration) publications  
  
data\_collab <- data[data[["Europe"]] != 0,]  
  
# Save the original title data for future use  
data\_collab$original\_concatenated\_title\_abstract <- data\_collab$concatenated\_title\_abstract  
  
#pre-processing the titles using textProcessor from the stm package  
processed\_text <- textProcessor(data\_collab$concatenated\_title\_abstract, metadata = data\_collab)

## Building corpus...   
## Converting to Lower Case...   
## Removing punctuation...   
## Removing stopwords...   
## Removing numbers...   
## Stemming...   
## Creating Output...

# Further prepare the data by removing low-frequency terms  
out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta)

## Removing 5015 of 15871 terms (5015 of 346655 tokens) due to frequency   
## Your corpus now has 4355 documents, 10856 terms and 341640 tokens.

docs\_text <- out\_text$documents  
vocab\_text <- out\_text$vocab  
meta\_text <- out\_text$meta  
  
  
#Prepare data  
plotRemoved(processed\_text$documents, lower.thresh = seq(1, 200, by = 100))



out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta, lower.thresh = 5)

## Removing 11837 of 15871 terms (23672 of 346655 tokens) due to frequency   
## Your corpus now has 4355 documents, 4034 terms and 322983 tokens.

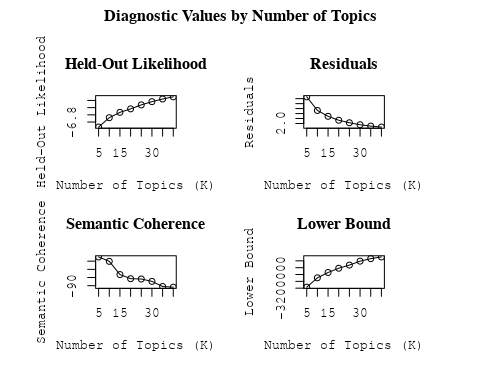
str(out\_text$meta)

## 'data.frame': 4355 obs. of 39 variables:  
## $ concept\_id : chr "https://openalex.org/C1276947" "https://openalex.org/C44870925" "https://openalex.org/C44870925" "https://openalex.org/C44870925" ...  
## $ work\_id : chr "https://openalex.org/W1967633346" "https://openalex.org/W1996505939" "https://openalex.org/W1994827287" "https://openalex.org/W1548323710" ...  
## $ publication\_year : int 2003 2014 2008 2002 1998 2008 2005 2005 2008 2008 ...  
## $ title : chr "Around-the-Clock Observations of the Q0957+561A,B Gravitationally Lensed Quasar. II. Results for the Second Observing Season" "A luminous, blue progenitor system for the type Iax supernova 2012Z" "On the Use of Blanketed Atmospheres as Boundary Conditions for Stellar Evolutionary Models" "Interpretation of the Core-Wing Anomaly of Balmer Line Profiles of Cool Ap Stars\*" ...  
## $ paperabstract : chr "We report on an observing campaign in 2001 March to monitor the brightness of the later arriving Q0957+561B ima"| \_\_truncated\_\_ "Type Iax supernovae are stellar explosions that are spectroscopically similar to some type Ia supernovae at the"| \_\_truncated\_\_ "Stellar models have been computed for stars having [ Fe/H ] = 0.0 (assuming both the Grevesse & Sauval and Aspl"| \_\_truncated\_\_ "A number of cool magnetic chemically peculiar stars exhibit abnormal profiles of hydrogen Balmer lines. This an"| \_\_truncated\_\_ ...  
## $ country : chr "US RS" "DK US US DK" "CA SE" "SE SE" ...  
## $ year\_concept : chr "2003+https://openalex.org/C1276947" "2014+https://openalex.org/C44870925" "2008+https://openalex.org/C44870925" "2002+https://openalex.org/C44870925" ...  
## $ concatenated\_title\_abstract : chr "Around-the-Clock Observations of the Q0957+561A,B Gravitationally Lensed Quasar. II. Results for the Second Obs"| \_\_truncated\_\_ "A luminous, blue progenitor system for the type Iax supernova 2012Z Type Iax supernovae are stellar explosions "| \_\_truncated\_\_ "On the Use of Blanketed Atmospheres as Boundary Conditions for Stellar Evolutionary Models Stellar models have "| \_\_truncated\_\_ "Interpretation of the Core-Wing Anomaly of Balmer Line Profiles of Cool Ap Stars\* A number of cool magnetic che"| \_\_truncated\_\_ ...  
## $ US : num 50 50 0 0 0 50 0 50 0 0 ...  
## $ IN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DE : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CH : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ GB : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ FR : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IT : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ RU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CA : num 0 0 50 0 0 0 0 0 0 0 ...  
## $ NL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ JP : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ES : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Americas : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Europe : num 50 50 50 100 100 50 100 50 100 100 ...  
## $ Africa : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AsiaAndOceania : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2020\_2022 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2015\_2019 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2010\_2014 : int 0 1 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2005\_2009 : int 0 0 1 0 0 1 1 1 1 1 ...  
## $ pub\_interval\_2000\_2004 : int 1 0 0 1 0 0 0 0 0 0 ...  
## $ pub\_interval\_1995\_1999 : int 0 0 0 0 1 0 0 0 0 0 ...  
## $ pub\_interval\_1985\_1994 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1975\_1984 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1965\_1974 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1900\_1964 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1824\_1899 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ original\_concatenated\_title\_abstract: chr "Around-the-Clock Observations of the Q0957+561A,B Gravitationally Lensed Quasar. II. Results for the Second Obs"| \_\_truncated\_\_ "A luminous, blue progenitor system for the type Iax supernova 2012Z Type Iax supernovae are stellar explosions "| \_\_truncated\_\_ "On the Use of Blanketed Atmospheres as Boundary Conditions for Stellar Evolutionary Models Stellar models have "| \_\_truncated\_\_ "Interpretation of the Core-Wing Anomaly of Balmer Line Profiles of Cool Ap Stars\* A number of cool magnetic che"| \_\_truncated\_\_ ...

# Initialize an empty formula string  
prevalence\_formula\_str <- "~"  
  
# Define the publication intervals  
pub\_intervals <- c("pub\_interval\_2020\_2022", "pub\_interval\_2015\_2019", "pub\_interval\_2010\_2014",   
 "pub\_interval\_2005\_2009", "pub\_interval\_2000\_2004", "pub\_interval\_1995\_1999",  
 "pub\_interval\_1985\_1994", "pub\_interval\_1975\_1984", "pub\_interval\_1965\_1974",  
 "pub\_interval\_1900\_1964", "pub\_interval\_1824\_1899")  
  
# Add each publication interval to the formula string  
for (interval in pub\_intervals) {  
 # add an if statement to handle the first addition (without '+')  
 if (prevalence\_formula\_str == "~") {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, interval)  
 } else {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, "+", interval)  
 }  
}  
  
  
# Convert the string to a formula  
prevalence\_formula <- as.formula(prevalence\_formula\_str)  
print(prevalence\_formula)

## ~pub\_interval\_2020\_2022 + pub\_interval\_2015\_2019 + pub\_interval\_2010\_2014 +   
## pub\_interval\_2005\_2009 + pub\_interval\_2000\_2004 + pub\_interval\_1995\_1999 +   
## pub\_interval\_1985\_1994 + pub\_interval\_1975\_1984 + pub\_interval\_1965\_1974 +   
## pub\_interval\_1900\_1964 + pub\_interval\_1824\_1899

Kvals <- seq(from = 5, to = 40, by = 5) # K values to try  
search\_results <- searchK(documents = out\_text$documents,   
 vocab = out\_text$vocab,   
 K = Kvals,   
 prevalence = prevalence\_formula,   
 data = out\_text$meta,   
 init.type = "Spectral",   
 verbose = FALSE)  
  
# Plot the results  
plot(search\_results)



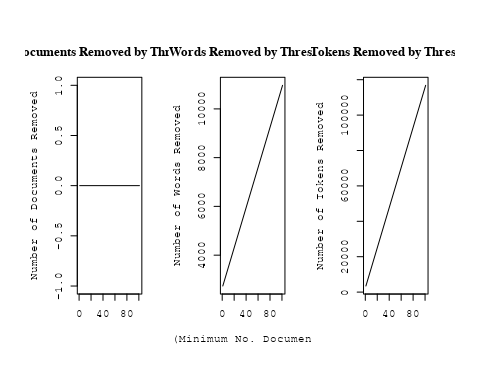
##Topic generation for IT (in collaboration) publications  
  
data\_collab <- data[data[["IT"]] != 0,]  
  
# Save the original title data for future use  
data\_collab$original\_concatenated\_title\_abstract <- data\_collab$concatenated\_title\_abstract  
  
#pre-processing the titles using textProcessor from the stm package  
processed\_text <- textProcessor(data\_collab$concatenated\_title\_abstract, metadata = data\_collab)

## Building corpus...   
## Converting to Lower Case...   
## Removing punctuation...   
## Removing stopwords...   
## Removing numbers...   
## Stemming...   
## Creating Output...

# Further prepare the data by removing low-frequency terms  
out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta)

## Removing 2735 of 11623 terms (2735 of 282817 tokens) due to frequency   
## Your corpus now has 3360 documents, 8888 terms and 280082 tokens.

docs\_text <- out\_text$documents  
vocab\_text <- out\_text$vocab  
meta\_text <- out\_text$meta  
  
  
#Prepare data  
plotRemoved(processed\_text$documents, lower.thresh = seq(1, 200, by = 100))



out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta, lower.thresh = 5)

## Removing 8109 of 11623 terms (17948 of 282817 tokens) due to frequency   
## Your corpus now has 3360 documents, 3514 terms and 264869 tokens.

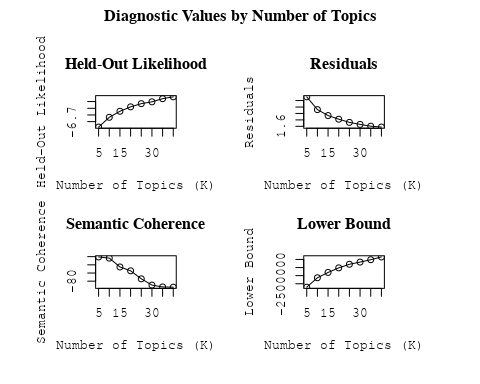
str(out\_text$meta)

## 'data.frame': 3360 obs. of 39 variables:  
## $ concept\_id : chr "https://openalex.org/C1276947" "https://openalex.org/C1276947" "https://openalex.org/C44870925" "https://openalex.org/C44870925" ...  
## $ work\_id : chr "https://openalex.org/W2068317303" "https://openalex.org/W3127045467" "https://openalex.org/W2086348127" "https://openalex.org/W2130580407" ...  
## $ publication\_year : int 2004 2021 2007 2011 1999 2001 1990 2009 2007 2011 ...  
## $ title : chr "Arp 299: A Second Merging System with Two Active Nuclei?" "A massive stellar bulge in a regularly rotating galaxy 1.2 billion years after the Big Bang" "Improving Stellar and Planetary Parameters of Transiting Planet Systems: The Case of TrES-2" "Creation of cosmic structure in the complex galaxy cluster merger Abell 2744" ...  
## $ paperabstract : chr "Recent BeppoSAX observations of Arp 299, a powerful far-IR merging starburst system composed of IC 694 and NGC "| \_\_truncated\_\_ "Early assembly of a galaxy disk and bulge Galaxy formation in the early Universe is thought to have been a chao"| \_\_truncated\_\_ "We report on a spectroscopic determination of the atmospheric parameters and chemical abundance of the parent s"| \_\_truncated\_\_ "We present a detailed strong lensing, weak lensing and X-ray analysis of Abell 2744 (z = 0:308), one of the mos"| \_\_truncated\_\_ ...  
## $ country : chr "IT IT IT IT" "GB GB IT GB GB IT" "IT US US US US IT" "US IT" ...  
## $ year\_concept : chr "2004+https://openalex.org/C1276947" "2021+https://openalex.org/C1276947" "2007+https://openalex.org/C44870925" "2011+https://openalex.org/C44870925" ...  
## $ concatenated\_title\_abstract : chr "Arp 299: A Second Merging System with Two Active Nuclei? Recent BeppoSAX observations of Arp 299, a powerful fa"| \_\_truncated\_\_ "A massive stellar bulge in a regularly rotating galaxy 1.2 billion years after the Big Bang Early assembly of a"| \_\_truncated\_\_ "Improving Stellar and Planetary Parameters of Transiting Planet Systems: The Case of TrES-2 We report on a spec"| \_\_truncated\_\_ "Creation of cosmic structure in the complex galaxy cluster merger Abell 2744 We present a detailed strong lensi"| \_\_truncated\_\_ ...  
## $ US : num 0 0 66.7 50 0 ...  
## $ IN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DE : num 0 0 0 0 0 ...  
## $ CH : num 0 0 0 0 0 0 0 50 0 0 ...  
## $ GB : num 0 66.7 0 0 0 ...  
## $ CN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ FR : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IT : num 100 33.3 33.3 50 100 ...  
## $ RU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CA : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ NL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ JP : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ES : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Americas : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Europe : num 0 0 0 0 0 ...  
## $ Africa : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AsiaAndOceania : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2020\_2022 : int 0 1 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2015\_2019 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2010\_2014 : int 0 0 0 1 0 0 0 0 0 1 ...  
## $ pub\_interval\_2005\_2009 : int 0 0 1 0 0 0 0 1 1 0 ...  
## $ pub\_interval\_2000\_2004 : int 1 0 0 0 0 1 0 0 0 0 ...  
## $ pub\_interval\_1995\_1999 : int 0 0 0 0 1 0 0 0 0 0 ...  
## $ pub\_interval\_1985\_1994 : int 0 0 0 0 0 0 1 0 0 0 ...  
## $ pub\_interval\_1975\_1984 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1965\_1974 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1900\_1964 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1824\_1899 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ original\_concatenated\_title\_abstract: chr "Arp 299: A Second Merging System with Two Active Nuclei? Recent BeppoSAX observations of Arp 299, a powerful fa"| \_\_truncated\_\_ "A massive stellar bulge in a regularly rotating galaxy 1.2 billion years after the Big Bang Early assembly of a"| \_\_truncated\_\_ "Improving Stellar and Planetary Parameters of Transiting Planet Systems: The Case of TrES-2 We report on a spec"| \_\_truncated\_\_ "Creation of cosmic structure in the complex galaxy cluster merger Abell 2744 We present a detailed strong lensi"| \_\_truncated\_\_ ...

# Initialize an empty formula string  
prevalence\_formula\_str <- "~"  
  
# Define the publication intervals  
pub\_intervals <- c("pub\_interval\_2020\_2022", "pub\_interval\_2015\_2019", "pub\_interval\_2010\_2014",   
 "pub\_interval\_2005\_2009", "pub\_interval\_2000\_2004", "pub\_interval\_1995\_1999",  
 "pub\_interval\_1985\_1994", "pub\_interval\_1975\_1984", "pub\_interval\_1965\_1974",  
 "pub\_interval\_1900\_1964", "pub\_interval\_1824\_1899")  
  
# Add each publication interval to the formula string  
for (interval in pub\_intervals) {  
 # add an if statement to handle the first addition (without '+')  
 if (prevalence\_formula\_str == "~") {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, interval)  
 } else {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, "+", interval)  
 }  
}  
  
  
# Convert the string to a formula  
prevalence\_formula <- as.formula(prevalence\_formula\_str)  
print(prevalence\_formula)

## ~pub\_interval\_2020\_2022 + pub\_interval\_2015\_2019 + pub\_interval\_2010\_2014 +   
## pub\_interval\_2005\_2009 + pub\_interval\_2000\_2004 + pub\_interval\_1995\_1999 +   
## pub\_interval\_1985\_1994 + pub\_interval\_1975\_1984 + pub\_interval\_1965\_1974 +   
## pub\_interval\_1900\_1964 + pub\_interval\_1824\_1899

Kvals <- seq(from = 5, to = 40, by = 5) # K values to try  
search\_results <- searchK(documents = out\_text$documents,   
 vocab = out\_text$vocab,   
 K = Kvals,   
 prevalence = prevalence\_formula,   
 data = out\_text$meta,   
 init.type = "Spectral",   
 verbose = FALSE)  
  
# Plot the results  
plot(search\_results)



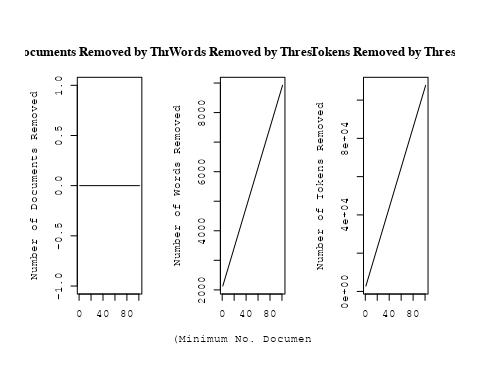
##Topic generation for AU (in collaboration) publications  
  
data\_collab <- data[data[["AU"]] != 0,]  
  
# Save the original title data for future use  
data\_collab$original\_concatenated\_title\_abstract <- data\_collab$concatenated\_title\_abstract  
  
#pre-processing the titles using textProcessor from the stm package  
processed\_text <- textProcessor(data\_collab$concatenated\_title\_abstract, metadata = data\_collab)

## Building corpus...   
## Converting to Lower Case...   
## Removing punctuation...   
## Removing stopwords...   
## Removing numbers...   
## Stemming...   
## Creating Output...

# Further prepare the data by removing low-frequency terms  
out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta)

## Removing 2131 of 9442 terms (2131 of 215077 tokens) due to frequency   
## Your corpus now has 2590 documents, 7311 terms and 212946 tokens.

docs\_text <- out\_text$documents  
vocab\_text <- out\_text$vocab  
meta\_text <- out\_text$meta  
  
  
#Prepare data  
plotRemoved(processed\_text$documents, lower.thresh = seq(1, 200, by = 100))



out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta, lower.thresh = 5)

## Removing 6360 of 9442 terms (14415 of 215077 tokens) due to frequency   
## Your corpus now has 2590 documents, 3082 terms and 200662 tokens.

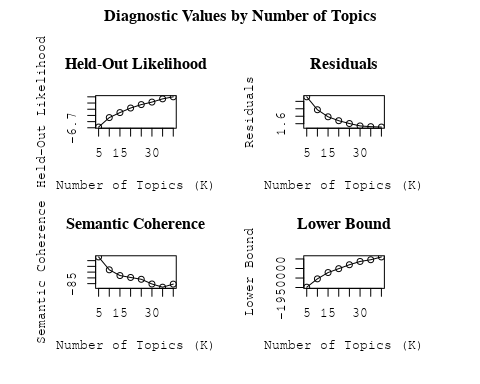
str(out\_text$meta)

## 'data.frame': 2590 obs. of 39 variables:  
## $ concept\_id : chr "https://openalex.org/C1276947" "https://openalex.org/C1276947" "https://openalex.org/C44870925" "https://openalex.org/C44870925" ...  
## $ work\_id : chr "https://openalex.org/W3100822581" "https://openalex.org/W3140362863" "https://openalex.org/W2019918353" "https://openalex.org/W2089705336" ...  
## $ publication\_year : int 2007 2021 1998 2001 2008 1999 2005 2001 1998 2014 ...  
## $ title : chr "Revealing Substructure in the Galactic Halo - The SEKBO RR Lyrae Survey" "Evidence for an intermediate-mass black hole from a gravitationally lensed gamma-ray burst" "The Nature of Bilateral Supernova Remnants" "NEW CLASS II METHANOL MASERS IN W3(OH)" ...  
## $ paperabstract : chr "We present a search for RR Lyrae variable stars from archival observations of the Southern Edgeworth-Kuiper Bel"| \_\_truncated\_\_ "If gamma-ray bursts are at cosmological distances, they must be gravitationally lensed occasionally [1, 2]. The"| \_\_truncated\_\_ "We present high-resolution radio images at 1.4 GHz of two Galactic supernova remnants (SNRs), G003.8–00.3 (form"| \_\_truncated\_\_ "We report interferometric observations of nine class II methanol maser candidate lines toward W3(OH). Narrow ma"| \_\_truncated\_\_ ...  
## $ country : chr "AU AU AU AU" "AU AU AU AU" "AU AU AU AU" "AU US AU US" ...  
## $ year\_concept : chr "2007+https://openalex.org/C1276947" "2021+https://openalex.org/C1276947" "1998+https://openalex.org/C44870925" "2001+https://openalex.org/C44870925" ...  
## $ concatenated\_title\_abstract : chr "Revealing Substructure in the Galactic Halo - The SEKBO RR Lyrae Survey We present a search for RR Lyrae variab"| \_\_truncated\_\_ "Evidence for an intermediate-mass black hole from a gravitationally lensed gamma-ray burst If gamma-ray bursts "| \_\_truncated\_\_ "The Nature of Bilateral Supernova Remnants We present high-resolution radio images at 1.4 GHz of two Galactic s"| \_\_truncated\_\_ "NEW CLASS II METHANOL MASERS IN W3(OH) We report interferometric observations of nine class II methanol maser c"| \_\_truncated\_\_ ...  
## $ US : num 0 0 0 50 0 ...  
## $ IN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DE : num 0 0 0 0 0 0 50 0 0 0 ...  
## $ CH : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ GB : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ FR : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IT : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ RU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CA : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ NL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AU : num 100 100 100 50 100 ...  
## $ JP : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ES : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Americas : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Europe : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Africa : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AsiaAndOceania : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2020\_2022 : int 0 1 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2015\_2019 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2010\_2014 : int 0 0 0 0 0 0 0 0 0 1 ...  
## $ pub\_interval\_2005\_2009 : int 1 0 0 0 1 0 1 0 0 0 ...  
## $ pub\_interval\_2000\_2004 : int 0 0 0 1 0 0 0 1 0 0 ...  
## $ pub\_interval\_1995\_1999 : int 0 0 1 0 0 1 0 0 1 0 ...  
## $ pub\_interval\_1985\_1994 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1975\_1984 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1965\_1974 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1900\_1964 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1824\_1899 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ original\_concatenated\_title\_abstract: chr "Revealing Substructure in the Galactic Halo - The SEKBO RR Lyrae Survey We present a search for RR Lyrae variab"| \_\_truncated\_\_ "Evidence for an intermediate-mass black hole from a gravitationally lensed gamma-ray burst If gamma-ray bursts "| \_\_truncated\_\_ "The Nature of Bilateral Supernova Remnants We present high-resolution radio images at 1.4 GHz of two Galactic s"| \_\_truncated\_\_ "NEW CLASS II METHANOL MASERS IN W3(OH) We report interferometric observations of nine class II methanol maser c"| \_\_truncated\_\_ ...

# Initialize an empty formula string  
prevalence\_formula\_str <- "~"  
  
# Define the publication intervals  
pub\_intervals <- c("pub\_interval\_2020\_2022", "pub\_interval\_2015\_2019", "pub\_interval\_2010\_2014",   
 "pub\_interval\_2005\_2009", "pub\_interval\_2000\_2004", "pub\_interval\_1995\_1999",  
 "pub\_interval\_1985\_1994", "pub\_interval\_1975\_1984", "pub\_interval\_1965\_1974",  
 "pub\_interval\_1900\_1964", "pub\_interval\_1824\_1899")  
  
# Add each publication interval to the formula string  
for (interval in pub\_intervals) {  
 # add an if statement to handle the first addition (without '+')  
 if (prevalence\_formula\_str == "~") {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, interval)  
 } else {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, "+", interval)  
 }  
}  
  
  
# Convert the string to a formula  
prevalence\_formula <- as.formula(prevalence\_formula\_str)  
print(prevalence\_formula)

## ~pub\_interval\_2020\_2022 + pub\_interval\_2015\_2019 + pub\_interval\_2010\_2014 +   
## pub\_interval\_2005\_2009 + pub\_interval\_2000\_2004 + pub\_interval\_1995\_1999 +   
## pub\_interval\_1985\_1994 + pub\_interval\_1975\_1984 + pub\_interval\_1965\_1974 +   
## pub\_interval\_1900\_1964 + pub\_interval\_1824\_1899

Kvals <- seq(from = 5, to = 40, by = 5) # K values to try  
search\_results <- searchK(documents = out\_text$documents,   
 vocab = out\_text$vocab,   
 K = Kvals,   
 prevalence = prevalence\_formula,   
 data = out\_text$meta,   
 init.type = "Spectral",   
 verbose = FALSE)  
  
# Plot the results  
plot(search\_results)



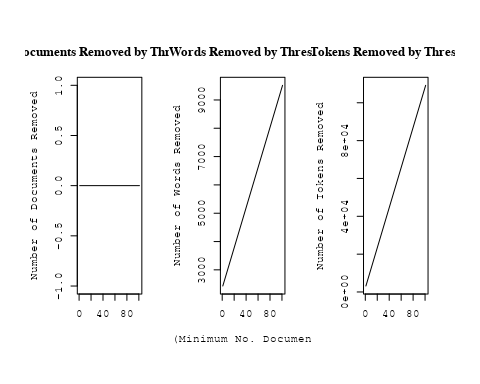
##Topic generation for CA (in collaboration) publications  
  
data\_collab <- data[data[["CA"]] != 0,]  
  
# Save the original title data for future use  
data\_collab$original\_concatenated\_title\_abstract <- data\_collab$concatenated\_title\_abstract  
  
#pre-processing the titles using textProcessor from the stm package  
processed\_text <- textProcessor(data\_collab$concatenated\_title\_abstract, metadata = data\_collab)

## Building corpus...   
## Converting to Lower Case...   
## Removing punctuation...   
## Removing stopwords...   
## Removing numbers...   
## Stemming...   
## Creating Output...

# Further prepare the data by removing low-frequency terms  
out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta)

## Removing 2430 of 10044 terms (2430 of 217233 tokens) due to frequency   
## Your corpus now has 2595 documents, 7614 terms and 214803 tokens.

docs\_text <- out\_text$documents  
vocab\_text <- out\_text$vocab  
meta\_text <- out\_text$meta  
  
  
#Prepare data  
plotRemoved(processed\_text$documents, lower.thresh = seq(1, 200, by = 100))



out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta, lower.thresh = 5)

## Removing 6959 of 10044 terms (15109 of 217233 tokens) due to frequency   
## Your corpus now has 2595 documents, 3085 terms and 202124 tokens.

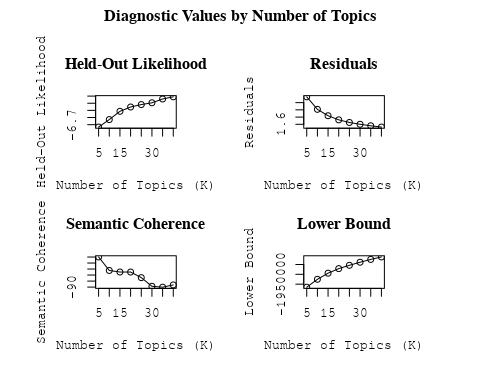
str(out\_text$meta)

## 'data.frame': 2595 obs. of 39 variables:  
## $ concept\_id : chr "https://openalex.org/C44870925" "https://openalex.org/C44870925" "https://openalex.org/C44870925" "https://openalex.org/C1276947" ...  
## $ work\_id : chr "https://openalex.org/W1993164994" "https://openalex.org/W1994827287" "https://openalex.org/W3099799912" "https://openalex.org/W3099799912" ...  
## $ publication\_year : int 1999 2008 1998 1998 2011 2001 2008 2001 2003 2001 ...  
## $ title : chr "Reconnection in a weakly stochastic field" "On the Use of Blanketed Atmospheres as Boundary Conditions for Stellar Evolutionary Models" "The Dependence of Cluster Galaxy Star Formation Rates on the Global Environment" "The Dependence of Cluster Galaxy Star Formation Rates on the Global Environment" ...  
## $ paperabstract : chr "We examine the effect of weak, small-scale magnetic field structure on the rate of reconnection in a strongly m"| \_\_truncated\_\_ "Stellar models have been computed for stars having [ Fe/H ] = 0.0 (assuming both the Grevesse & Sauval and Aspl"| \_\_truncated\_\_ "A comparison of star formation properties as a function of environment is made from the spectra of identically "| \_\_truncated\_\_ "A comparison of star formation properties as a function of environment is made from the spectra of identically "| \_\_truncated\_\_ ...  
## $ country : chr "US US CA US US CA" "CA SE" "US US CA US US CA" "US US CA US US CA" ...  
## $ year\_concept : chr "1999+https://openalex.org/C44870925" "2008+https://openalex.org/C44870925" "1998+https://openalex.org/C44870925" "1998+https://openalex.org/C1276947" ...  
## $ concatenated\_title\_abstract : chr "Reconnection in a weakly stochastic field We examine the effect of weak, small-scale magnetic field structure o"| \_\_truncated\_\_ "On the Use of Blanketed Atmospheres as Boundary Conditions for Stellar Evolutionary Models Stellar models have "| \_\_truncated\_\_ "The Dependence of Cluster Galaxy Star Formation Rates on the Global Environment A comparison of star formation "| \_\_truncated\_\_ "The Dependence of Cluster Galaxy Star Formation Rates on the Global Environment A comparison of star formation "| \_\_truncated\_\_ ...  
## $ US : num 66.7 0 66.7 66.7 25 ...  
## $ IN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DE : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CH : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ GB : num 0 0 0 0 50 0 0 0 50 0 ...  
## $ CN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ FR : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IT : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ RU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CA : num 33.3 50 33.3 33.3 25 ...  
## $ NL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ JP : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ES : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Americas : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Europe : num 0 50 0 0 0 0 0 0 0 0 ...  
## $ Africa : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AsiaAndOceania : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2020\_2022 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2015\_2019 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2010\_2014 : int 0 0 0 0 1 0 0 0 0 0 ...  
## $ pub\_interval\_2005\_2009 : int 0 1 0 0 0 0 1 0 0 0 ...  
## $ pub\_interval\_2000\_2004 : int 0 0 0 0 0 1 0 1 1 1 ...  
## $ pub\_interval\_1995\_1999 : int 1 0 1 1 0 0 0 0 0 0 ...  
## $ pub\_interval\_1985\_1994 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1975\_1984 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1965\_1974 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1900\_1964 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1824\_1899 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ original\_concatenated\_title\_abstract: chr "Reconnection in a weakly stochastic field We examine the effect of weak, small-scale magnetic field structure o"| \_\_truncated\_\_ "On the Use of Blanketed Atmospheres as Boundary Conditions for Stellar Evolutionary Models Stellar models have "| \_\_truncated\_\_ "The Dependence of Cluster Galaxy Star Formation Rates on the Global Environment A comparison of star formation "| \_\_truncated\_\_ "The Dependence of Cluster Galaxy Star Formation Rates on the Global Environment A comparison of star formation "| \_\_truncated\_\_ ...

# Initialize an empty formula string  
prevalence\_formula\_str <- "~"  
  
# Define the publication intervals  
pub\_intervals <- c("pub\_interval\_2020\_2022", "pub\_interval\_2015\_2019", "pub\_interval\_2010\_2014",   
 "pub\_interval\_2005\_2009", "pub\_interval\_2000\_2004", "pub\_interval\_1995\_1999",  
 "pub\_interval\_1985\_1994", "pub\_interval\_1975\_1984", "pub\_interval\_1965\_1974",  
 "pub\_interval\_1900\_1964", "pub\_interval\_1824\_1899")  
  
# Add each publication interval to the formula string  
for (interval in pub\_intervals) {  
 # add an if statement to handle the first addition (without '+')  
 if (prevalence\_formula\_str == "~") {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, interval)  
 } else {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, "+", interval)  
 }  
}  
  
  
# Convert the string to a formula  
prevalence\_formula <- as.formula(prevalence\_formula\_str)  
print(prevalence\_formula)

## ~pub\_interval\_2020\_2022 + pub\_interval\_2015\_2019 + pub\_interval\_2010\_2014 +   
## pub\_interval\_2005\_2009 + pub\_interval\_2000\_2004 + pub\_interval\_1995\_1999 +   
## pub\_interval\_1985\_1994 + pub\_interval\_1975\_1984 + pub\_interval\_1965\_1974 +   
## pub\_interval\_1900\_1964 + pub\_interval\_1824\_1899

Kvals <- seq(from = 5, to = 40, by = 5) # K values to try  
search\_results <- searchK(documents = out\_text$documents,   
 vocab = out\_text$vocab,   
 K = Kvals,   
 prevalence = prevalence\_formula,   
 data = out\_text$meta,   
 init.type = "Spectral",   
 verbose = FALSE)  
  
# Plot the results  
plot(search\_results)



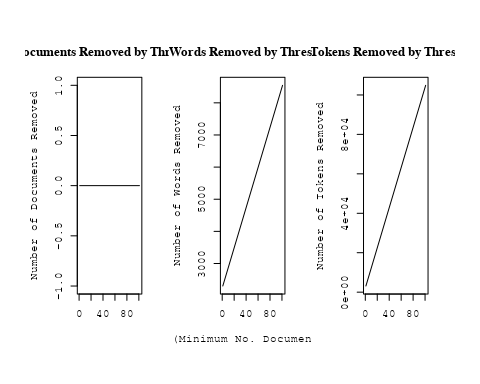
##Topic generation for JP (in collaboration) publications  
  
data\_collab <- data[data[["JP"]] != 0,]  
  
# Save the original title data for future use  
data\_collab$original\_concatenated\_title\_abstract <- data\_collab$concatenated\_title\_abstract  
  
#pre-processing the titles using textProcessor from the stm package  
processed\_text <- textProcessor(data\_collab$concatenated\_title\_abstract, metadata = data\_collab)

## Building corpus...   
## Converting to Lower Case...   
## Removing punctuation...   
## Removing stopwords...   
## Removing numbers...   
## Stemming...   
## Creating Output...

# Further prepare the data by removing low-frequency terms  
out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta)

## Removing 2300 of 8995 terms (2300 of 178949 tokens) due to frequency   
## Your corpus now has 2285 documents, 6695 terms and 176649 tokens.

docs\_text <- out\_text$documents  
vocab\_text <- out\_text$vocab  
meta\_text <- out\_text$meta  
  
  
#Prepare data  
plotRemoved(processed\_text$documents, lower.thresh = seq(1, 200, by = 100))



out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta, lower.thresh = 5)

## Removing 6142 of 8995 terms (13004 of 178949 tokens) due to frequency   
## Your corpus now has 2285 documents, 2853 terms and 165945 tokens.

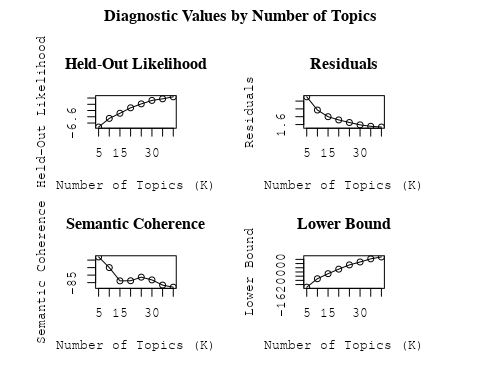
str(out\_text$meta)

## 'data.frame': 2285 obs. of 39 variables:  
## $ concept\_id : chr "https://openalex.org/C44870925" "https://openalex.org/C1276947" "https://openalex.org/C44870925" "https://openalex.org/C44870925" ...  
## $ work\_id : chr "https://openalex.org/W3100494265" "https://openalex.org/W2153493790" "https://openalex.org/W2905780457" "https://openalex.org/W1993936459" ...  
## $ publication\_year : int 2002 2011 2019 2006 1998 2009 1997 1998 2018 2016 ...  
## $ title : chr "Runaway Merging of Black Holes: Analytical Constraint on the Timescale" "Effects of the explosion asymmetry and viewing angle on the Type Ia supernova colour and luminosity calibration" "A warped disk around an infant protostar" "The escape fraction of ionizing photons from galaxies at z= 0–6" ...  
## $ paperabstract : chr "Following the discovery of a black hole (BH) with a mass of 103-106 M☉ in the starburst galaxy M82, we study th"| \_\_truncated\_\_ "Phenomenological relations exist between the peak luminosity and other observables of type Ia supernovae (SNe I"| \_\_truncated\_\_ "Recent exoplanet studies have revealed that the orbital planes of planets are not always aligned with one anoth"| \_\_truncated\_\_ "The escape fraction of ionizing photons from galaxies is a crucial quantity controlling the cosmic ionizing bac"| \_\_truncated\_\_ ...  
## $ country : chr "JP" "JP DE JP DE" "US JP JP US" "JP" ...  
## $ year\_concept : chr "2002+https://openalex.org/C44870925" "2011+https://openalex.org/C1276947" "2019+https://openalex.org/C44870925" "2006+https://openalex.org/C44870925" ...  
## $ concatenated\_title\_abstract : chr "Runaway Merging of Black Holes: Analytical Constraint on the Timescale Following the discovery of a black hole "| \_\_truncated\_\_ "Effects of the explosion asymmetry and viewing angle on the Type Ia supernova colour and luminosity calibration"| \_\_truncated\_\_ "A warped disk around an infant protostar Recent exoplanet studies have revealed that the orbital planes of plan"| \_\_truncated\_\_ "The escape fraction of ionizing photons from galaxies at z= 0–6 The escape fraction of ionizing photons from ga"| \_\_truncated\_\_ ...  
## $ US : num 0 0 50 0 0 ...  
## $ IN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DE : num 0 50 0 0 0 0 0 0 0 0 ...  
## $ CH : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ GB : num 0 0 0 0 0 0 0 0 0 50 ...  
## $ CN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ FR : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IT : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ RU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CA : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ NL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ JP : num 100 50 50 100 100 ...  
## $ ES : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Americas : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Europe : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Africa : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AsiaAndOceania : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2020\_2022 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2015\_2019 : int 0 0 1 0 0 0 0 0 1 1 ...  
## $ pub\_interval\_2010\_2014 : int 0 1 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2005\_2009 : int 0 0 0 1 0 1 0 0 0 0 ...  
## $ pub\_interval\_2000\_2004 : int 1 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1995\_1999 : int 0 0 0 0 1 0 1 1 0 0 ...  
## $ pub\_interval\_1985\_1994 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1975\_1984 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1965\_1974 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1900\_1964 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1824\_1899 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ original\_concatenated\_title\_abstract: chr "Runaway Merging of Black Holes: Analytical Constraint on the Timescale Following the discovery of a black hole "| \_\_truncated\_\_ "Effects of the explosion asymmetry and viewing angle on the Type Ia supernova colour and luminosity calibration"| \_\_truncated\_\_ "A warped disk around an infant protostar Recent exoplanet studies have revealed that the orbital planes of plan"| \_\_truncated\_\_ "The escape fraction of ionizing photons from galaxies at z= 0–6 The escape fraction of ionizing photons from ga"| \_\_truncated\_\_ ...

# Initialize an empty formula string  
prevalence\_formula\_str <- "~"  
  
# Define the publication intervals  
pub\_intervals <- c("pub\_interval\_2020\_2022", "pub\_interval\_2015\_2019", "pub\_interval\_2010\_2014",   
 "pub\_interval\_2005\_2009", "pub\_interval\_2000\_2004", "pub\_interval\_1995\_1999",  
 "pub\_interval\_1985\_1994", "pub\_interval\_1975\_1984", "pub\_interval\_1965\_1974",  
 "pub\_interval\_1900\_1964", "pub\_interval\_1824\_1899")  
  
# Add each publication interval to the formula string  
for (interval in pub\_intervals) {  
 # add an if statement to handle the first addition (without '+')  
 if (prevalence\_formula\_str == "~") {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, interval)  
 } else {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, "+", interval)  
 }  
}  
  
  
# Convert the string to a formula  
prevalence\_formula <- as.formula(prevalence\_formula\_str)  
print(prevalence\_formula)

## ~pub\_interval\_2020\_2022 + pub\_interval\_2015\_2019 + pub\_interval\_2010\_2014 +   
## pub\_interval\_2005\_2009 + pub\_interval\_2000\_2004 + pub\_interval\_1995\_1999 +   
## pub\_interval\_1985\_1994 + pub\_interval\_1975\_1984 + pub\_interval\_1965\_1974 +   
## pub\_interval\_1900\_1964 + pub\_interval\_1824\_1899

Kvals <- seq(from = 5, to = 40, by = 5) # K values to try  
search\_results <- searchK(documents = out\_text$documents,   
 vocab = out\_text$vocab,   
 K = Kvals,   
 prevalence = prevalence\_formula,   
 data = out\_text$meta,   
 init.type = "Spectral",   
 verbose = FALSE)  
  
# Plot the results  
plot(search\_results)



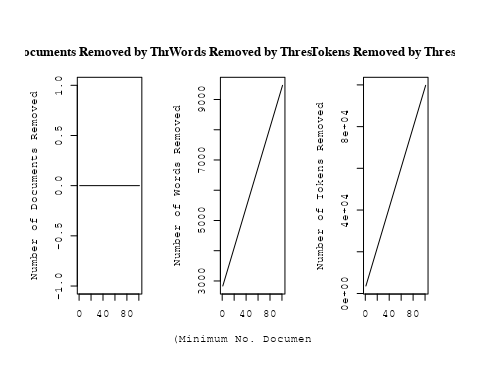
##Topic generation for Americas (in collaboration) publications  
  
data\_collab <- data[data[["Americas"]] != 0,]  
  
# Save the original title data for future use  
data\_collab$original\_concatenated\_title\_abstract <- data\_collab$concatenated\_title\_abstract  
  
#pre-processing the titles using textProcessor from the stm package  
processed\_text <- textProcessor(data\_collab$concatenated\_title\_abstract, metadata = data\_collab)

## Building corpus...   
## Converting to Lower Case...   
## Removing punctuation...   
## Removing stopwords...   
## Removing numbers...   
## Stemming...   
## Creating Output...

# Further prepare the data by removing low-frequency terms  
out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta)

## Removing 2834 of 9898 terms (2834 of 176304 tokens) due to frequency   
## Your corpus now has 2128 documents, 7064 terms and 173470 tokens.

docs\_text <- out\_text$documents  
vocab\_text <- out\_text$vocab  
meta\_text <- out\_text$meta  
  
  
#Prepare data  
plotRemoved(processed\_text$documents, lower.thresh = seq(1, 200, by = 100))



out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta, lower.thresh = 5)

## Removing 7083 of 9898 terms (14681 of 176304 tokens) due to frequency   
## Your corpus now has 2128 documents, 2815 terms and 161623 tokens.

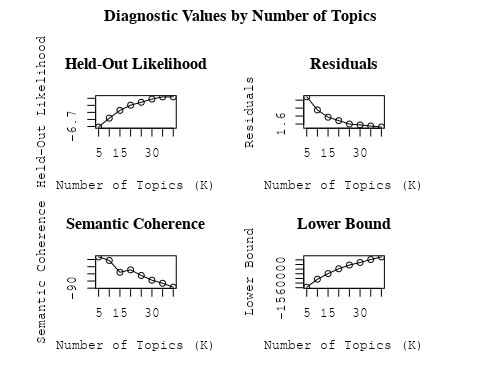
str(out\_text$meta)

## 'data.frame': 2128 obs. of 39 variables:  
## $ concept\_id : chr "https://openalex.org/C44870925" "https://openalex.org/C1276947" "https://openalex.org/C1276947" "https://openalex.org/C44870925" ...  
## $ work\_id : chr "https://openalex.org/W2982505463" "https://openalex.org/W2950791913" "https://openalex.org/W1760966995" "https://openalex.org/W3104422960" ...  
## $ publication\_year : int 2019 2010 2010 2004 1998 2006 2011 2002 2008 2020 ...  
## $ title : chr "On the possibilities of classical nova identifications among historical Far Eastern guest star observations" "A comprehensive classification of galaxies in the SDSS: How to tell true from fake AGN?" "Alternative diagnostic diagrams and the 'forgotten' population of weak line galaxies in the SDSS" "The star formation history of Seyfert 2 nuclei" ...  
## $ paperabstract : chr "More than 100 guest star observations have been obtained by Chinese, Korean, Japanese and Vietnamese astronomer"| \_\_truncated\_\_ "We use the W Hα versus [NII]/Hα (WHAN) diagram introduced by us in previous work to provide a comprehensive emi"| \_\_truncated\_\_ "A numerous population of weak line galaxies (WLGs) is often left out of statistical studies on emission-line ga"| \_\_truncated\_\_ "We present a study of the stellar populations in the central ∼200 pc of a large and homogeneous sample comprisi"| \_\_truncated\_\_ ...  
## $ country : chr "CL US" "BR BR BR BR" "BR BR BR BR" "BR FR FR BR" ...  
## $ year\_concept : chr "2019+https://openalex.org/C44870925" "2010+https://openalex.org/C1276947" "2010+https://openalex.org/C1276947" "2004+https://openalex.org/C44870925" ...  
## $ concatenated\_title\_abstract : chr "On the possibilities of classical nova identifications among historical Far Eastern guest star observations Mor"| \_\_truncated\_\_ "A comprehensive classification of galaxies in the SDSS: How to tell true from fake AGN? We use the W Hα versus "| \_\_truncated\_\_ "Alternative diagnostic diagrams and the 'forgotten' population of weak line galaxies in the SDSS A numerous pop"| \_\_truncated\_\_ "The star formation history of Seyfert 2 nuclei We present a study of the stellar populations in the central ∼20"| \_\_truncated\_\_ ...  
## $ US : num 50 0 0 0 66.7 ...  
## $ IN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DE : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CH : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ GB : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ FR : num 0 0 0 50 0 0 50 0 0 0 ...  
## $ IT : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ RU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CA : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ NL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ JP : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ES : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Americas : num 50 100 100 50 33.3 ...  
## $ Europe : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Africa : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AsiaAndOceania : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2020\_2022 : int 0 0 0 0 0 0 0 0 0 1 ...  
## $ pub\_interval\_2015\_2019 : int 1 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2010\_2014 : int 0 1 1 0 0 0 1 0 0 0 ...  
## $ pub\_interval\_2005\_2009 : int 0 0 0 0 0 1 0 0 1 0 ...  
## $ pub\_interval\_2000\_2004 : int 0 0 0 1 0 0 0 1 0 0 ...  
## $ pub\_interval\_1995\_1999 : int 0 0 0 0 1 0 0 0 0 0 ...  
## $ pub\_interval\_1985\_1994 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1975\_1984 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1965\_1974 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1900\_1964 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1824\_1899 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ original\_concatenated\_title\_abstract: chr "On the possibilities of classical nova identifications among historical Far Eastern guest star observations Mor"| \_\_truncated\_\_ "A comprehensive classification of galaxies in the SDSS: How to tell true from fake AGN? We use the W Hα versus "| \_\_truncated\_\_ "Alternative diagnostic diagrams and the 'forgotten' population of weak line galaxies in the SDSS A numerous pop"| \_\_truncated\_\_ "The star formation history of Seyfert 2 nuclei We present a study of the stellar populations in the central ∼20"| \_\_truncated\_\_ ...

# Initialize an empty formula string  
prevalence\_formula\_str <- "~"  
  
# Define the publication intervals  
pub\_intervals <- c("pub\_interval\_2020\_2022", "pub\_interval\_2015\_2019", "pub\_interval\_2010\_2014",   
 "pub\_interval\_2005\_2009", "pub\_interval\_2000\_2004", "pub\_interval\_1995\_1999",  
 "pub\_interval\_1985\_1994", "pub\_interval\_1975\_1984", "pub\_interval\_1965\_1974",  
 "pub\_interval\_1900\_1964", "pub\_interval\_1824\_1899")  
  
# Add each publication interval to the formula string  
for (interval in pub\_intervals) {  
 # add an if statement to handle the first addition (without '+')  
 if (prevalence\_formula\_str == "~") {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, interval)  
 } else {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, "+", interval)  
 }  
}  
  
  
# Convert the string to a formula  
prevalence\_formula <- as.formula(prevalence\_formula\_str)  
print(prevalence\_formula)

## ~pub\_interval\_2020\_2022 + pub\_interval\_2015\_2019 + pub\_interval\_2010\_2014 +   
## pub\_interval\_2005\_2009 + pub\_interval\_2000\_2004 + pub\_interval\_1995\_1999 +   
## pub\_interval\_1985\_1994 + pub\_interval\_1975\_1984 + pub\_interval\_1965\_1974 +   
## pub\_interval\_1900\_1964 + pub\_interval\_1824\_1899

Kvals <- seq(from = 5, to = 40, by = 5) # K values to try  
search\_results <- searchK(documents = out\_text$documents,   
 vocab = out\_text$vocab,   
 K = Kvals,   
 prevalence = prevalence\_formula,   
 data = out\_text$meta,   
 init.type = "Spectral",   
 verbose = FALSE)  
  
# Plot the results  
plot(search\_results)



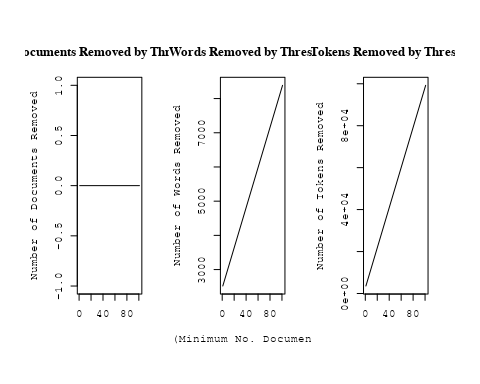
##Topic generation for CN (in collaboration) publications  
  
data\_collab <- data[data[["CN"]] != 0,]  
  
# Save the original title data for future use  
data\_collab$original\_concatenated\_title\_abstract <- data\_collab$concatenated\_title\_abstract  
  
#pre-processing the titles using textProcessor from the stm package  
processed\_text <- textProcessor(data\_collab$concatenated\_title\_abstract, metadata = data\_collab)

## Building corpus...   
## Converting to Lower Case...   
## Removing punctuation...   
## Removing stopwords...   
## Removing numbers...   
## Stemming...   
## Creating Output...

# Further prepare the data by removing low-frequency terms  
out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta)

## Removing 2519 of 8725 terms (2519 of 140803 tokens) due to frequency   
## Your corpus now has 1814 documents, 6206 terms and 138284 tokens.

docs\_text <- out\_text$documents  
vocab\_text <- out\_text$vocab  
meta\_text <- out\_text$meta  
  
  
#Prepare data  
plotRemoved(processed\_text$documents, lower.thresh = seq(1, 200, by = 100))



out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta, lower.thresh = 5)

## Removing 6196 of 8725 terms (12773 of 140803 tokens) due to frequency   
## Your corpus now has 1814 documents, 2529 terms and 128030 tokens.

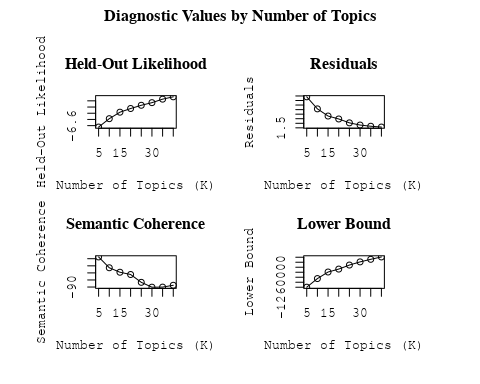
str(out\_text$meta)

## 'data.frame': 1814 obs. of 39 variables:  
## $ concept\_id : chr "https://openalex.org/C44870925" "https://openalex.org/C44870925" "https://openalex.org/C1276947" "https://openalex.org/C44870925" ...  
## $ work\_id : chr "https://openalex.org/W3105293521" "https://openalex.org/W3122804126" "https://openalex.org/W3122804126" "https://openalex.org/W2798563684" ...  
## $ publication\_year : int 2007 2010 2010 2018 2009 2003 2021 2021 2002 2016 ...  
## $ title : chr "Post-Oligarchic Evolution of Protoplanetary Embryos and the Stability of Planetary Systems" "Lepto-Hadronic Origin of gamma-rays from the G54.1+0.3 Pulsar Wind Nebula" "Lepto-Hadronic Origin of gamma-rays from the G54.1+0.3 Pulsar Wind Nebula" "Magnetic structure of solar flare regions producing hard X-ray pulsations" ...  
## $ paperabstract : chr "In the sequential accretion model, planets form through the sedimentation of dust, cohesive collisions of plane"| \_\_truncated\_\_ "G54.1+0.3 is a Crab-like pulsar wind nebula (PWN) with the highest γ -ray to X-ray luminosity ratio among all t"| \_\_truncated\_\_ "G54.1+0.3 is a Crab-like pulsar wind nebula (PWN) with the highest γ -ray to X-ray luminosity ratio among all t"| \_\_truncated\_\_ "Abstract We present analysis of the magnetic field in seven solar flare regions accompanied by the pulsations o"| \_\_truncated\_\_ ...  
## $ country : chr "CN CN CN CN" "CN" "CN" "CH RU GB RU CN" ...  
## $ year\_concept : chr "2007+https://openalex.org/C44870925" "2010+https://openalex.org/C44870925" "2010+https://openalex.org/C1276947" "2018+https://openalex.org/C44870925" ...  
## $ concatenated\_title\_abstract : chr "Post-Oligarchic Evolution of Protoplanetary Embryos and the Stability of Planetary Systems In the sequential ac"| \_\_truncated\_\_ "Lepto-Hadronic Origin of gamma-rays from the G54.1+0.3 Pulsar Wind Nebula G54.1+0.3 is a Crab-like pulsar wind "| \_\_truncated\_\_ "Lepto-Hadronic Origin of gamma-rays from the G54.1+0.3 Pulsar Wind Nebula G54.1+0.3 is a Crab-like pulsar wind "| \_\_truncated\_\_ "Magnetic structure of solar flare regions producing hard X-ray pulsations Abstract We present analysis of the m"| \_\_truncated\_\_ ...  
## $ US : num 0 0 0 0 25 0 25 25 0 0 ...  
## $ IN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DE : num 0 0 0 0 0 50 0 0 0 0 ...  
## $ CH : num 0 0 0 20 0 0 0 0 0 0 ...  
## $ GB : num 0 0 0 20 0 0 0 0 50 0 ...  
## $ CN : num 100 100 100 20 75 50 25 25 50 100 ...  
## $ FR : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IT : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ RU : num 0 0 0 40 0 0 0 0 0 0 ...  
## $ CA : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ NL : num 0 0 0 0 0 0 25 25 0 0 ...  
## $ AU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ JP : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ES : num 0 0 0 0 0 0 25 25 0 0 ...  
## $ IL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Americas : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Europe : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Africa : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AsiaAndOceania : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2020\_2022 : int 0 0 0 0 0 0 1 1 0 0 ...  
## $ pub\_interval\_2015\_2019 : int 0 0 0 1 0 0 0 0 0 1 ...  
## $ pub\_interval\_2010\_2014 : int 0 1 1 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2005\_2009 : int 1 0 0 0 1 0 0 0 0 0 ...  
## $ pub\_interval\_2000\_2004 : int 0 0 0 0 0 1 0 0 1 0 ...  
## $ pub\_interval\_1995\_1999 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1985\_1994 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1975\_1984 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1965\_1974 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1900\_1964 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1824\_1899 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ original\_concatenated\_title\_abstract: chr "Post-Oligarchic Evolution of Protoplanetary Embryos and the Stability of Planetary Systems In the sequential ac"| \_\_truncated\_\_ "Lepto-Hadronic Origin of gamma-rays from the G54.1+0.3 Pulsar Wind Nebula G54.1+0.3 is a Crab-like pulsar wind "| \_\_truncated\_\_ "Lepto-Hadronic Origin of gamma-rays from the G54.1+0.3 Pulsar Wind Nebula G54.1+0.3 is a Crab-like pulsar wind "| \_\_truncated\_\_ "Magnetic structure of solar flare regions producing hard X-ray pulsations Abstract We present analysis of the m"| \_\_truncated\_\_ ...

# Initialize an empty formula string  
prevalence\_formula\_str <- "~"  
  
# Define the publication intervals  
pub\_intervals <- c("pub\_interval\_2020\_2022", "pub\_interval\_2015\_2019", "pub\_interval\_2010\_2014",   
 "pub\_interval\_2005\_2009", "pub\_interval\_2000\_2004", "pub\_interval\_1995\_1999",  
 "pub\_interval\_1985\_1994", "pub\_interval\_1975\_1984", "pub\_interval\_1965\_1974",  
 "pub\_interval\_1900\_1964", "pub\_interval\_1824\_1899")  
  
# Add each publication interval to the formula string  
for (interval in pub\_intervals) {  
 # add an if statement to handle the first addition (without '+')  
 if (prevalence\_formula\_str == "~") {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, interval)  
 } else {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, "+", interval)  
 }  
}  
  
  
# Convert the string to a formula  
prevalence\_formula <- as.formula(prevalence\_formula\_str)  
print(prevalence\_formula)

## ~pub\_interval\_2020\_2022 + pub\_interval\_2015\_2019 + pub\_interval\_2010\_2014 +   
## pub\_interval\_2005\_2009 + pub\_interval\_2000\_2004 + pub\_interval\_1995\_1999 +   
## pub\_interval\_1985\_1994 + pub\_interval\_1975\_1984 + pub\_interval\_1965\_1974 +   
## pub\_interval\_1900\_1964 + pub\_interval\_1824\_1899

Kvals <- seq(from = 5, to = 40, by = 5) # K values to try  
search\_results <- searchK(documents = out\_text$documents,   
 vocab = out\_text$vocab,   
 K = Kvals,   
 prevalence = prevalence\_formula,   
 data = out\_text$meta,   
 init.type = "Spectral",   
 verbose = FALSE)  
  
# Plot the results  
plot(search\_results)



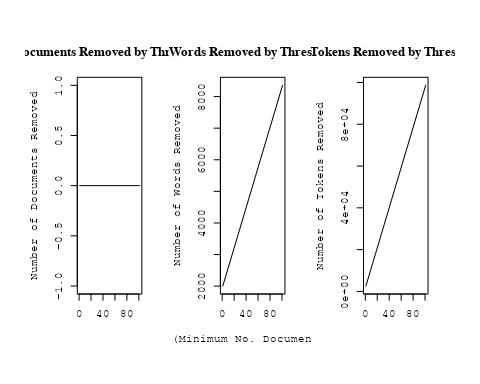
##Topic generation for NL (in collaboration) publications  
  
data\_collab <- data[data[["NL"]] != 0,]  
  
# Save the original title data for future use  
data\_collab$original\_concatenated\_title\_abstract <- data\_collab$concatenated\_title\_abstract  
  
#pre-processing the titles using textProcessor from the stm package  
processed\_text <- textProcessor(data\_collab$concatenated\_title\_abstract, metadata = data\_collab)

## Building corpus...   
## Converting to Lower Case...   
## Removing punctuation...   
## Removing stopwords...   
## Removing numbers...   
## Stemming...   
## Creating Output...

# Further prepare the data by removing low-frequency terms  
out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta)

## Removing 2004 of 8772 terms (2004 of 167908 tokens) due to frequency   
## Your corpus now has 2002 documents, 6768 terms and 165904 tokens.

docs\_text <- out\_text$documents  
vocab\_text <- out\_text$vocab  
meta\_text <- out\_text$meta  
  
  
#Prepare data  
plotRemoved(processed\_text$documents, lower.thresh = seq(1, 200, by = 100))



out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta, lower.thresh = 5)

## Removing 5974 of 8772 terms (13042 of 167908 tokens) due to frequency   
## Your corpus now has 2002 documents, 2798 terms and 154866 tokens.

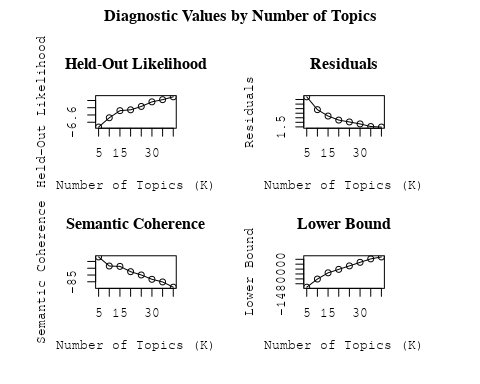
str(out\_text$meta)

## 'data.frame': 2002 obs. of 39 variables:  
## $ concept\_id : chr "https://openalex.org/C44870925" "https://openalex.org/C1276947" "https://openalex.org/C1276947" "https://openalex.org/C1276947" ...  
## $ work\_id : chr "https://openalex.org/W2030196560" "https://openalex.org/W2953328966" "https://openalex.org/W2039791607" "https://openalex.org/W2133651179" ...  
## $ publication\_year : int 2009 2009 1988 2007 2008 2007 2004 2021 2021 2009 ...  
## $ title : chr "The Effect of Stellar Rotation on Colour-Magnitude Diagrams: On the apparent presence of multiple populations i"| \_\_truncated\_\_ "Galactic chemical evolution in hierarchical formation models - I. Early-type galaxies in the local Universe" "Fate of the companion stars of ultra-rapid pulsars" "Infrared Molecular Starburst Fingerprints in Deeply Obscured (Ultra)Luminous Infrared Galaxy Nuclei" ...  
## $ paperabstract : chr "A significant number of intermediate age clusters (1 − 2 Gyr) in the Magellanic Clouds appear to have multiple "| \_\_truncated\_\_ "We study the metallicities and abundance ratios of early-type galaxies in cosmological semi-analytic models (SA"| \_\_truncated\_\_ "Abstract A millisecond pulsar that is formed by spin-up 'recycling'1,2 in a binary system will, once the mass t"| \_\_truncated\_\_ "High-resolution spectra of the Spitzer Space Telescope show vibration-rotation absorption bands of gaseous C2H2"| \_\_truncated\_\_ ...  
## $ country : chr "GB NL GB NL" "GB GB NL NL" "NL NL" "NL" ...  
## $ year\_concept : chr "2009+https://openalex.org/C44870925" "2009+https://openalex.org/C1276947" "1988+https://openalex.org/C1276947" "2007+https://openalex.org/C1276947" ...  
## $ concatenated\_title\_abstract : chr "The Effect of Stellar Rotation on Colour-Magnitude Diagrams: On the apparent presence of multiple populations i"| \_\_truncated\_\_ "Galactic chemical evolution in hierarchical formation models - I. Early-type galaxies in the local Universe We "| \_\_truncated\_\_ "Fate of the companion stars of ultra-rapid pulsars Abstract A millisecond pulsar that is formed by spin-up 'rec"| \_\_truncated\_\_ "Infrared Molecular Starburst Fingerprints in Deeply Obscured (Ultra)Luminous Infrared Galaxy Nuclei High-resolu"| \_\_truncated\_\_ ...  
## $ US : num 0 0 0 0 60 0 50 25 25 0 ...  
## $ IN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DE : num 0 0 0 0 0 0 0 0 0 25 ...  
## $ CH : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ GB : num 50 50 0 0 0 ...  
## $ CN : num 0 0 0 0 0 0 0 25 25 0 ...  
## $ FR : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IT : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ RU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CA : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ NL : num 50 50 100 100 40 ...  
## $ AU : num 0 0 0 0 0 0 0 0 0 25 ...  
## $ JP : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ES : num 0 0 0 0 0 0 0 25 25 0 ...  
## $ IL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Americas : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Europe : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Africa : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AsiaAndOceania : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2020\_2022 : int 0 0 0 0 0 0 0 1 1 0 ...  
## $ pub\_interval\_2015\_2019 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2010\_2014 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2005\_2009 : int 1 1 0 1 1 1 0 0 0 1 ...  
## $ pub\_interval\_2000\_2004 : int 0 0 0 0 0 0 1 0 0 0 ...  
## $ pub\_interval\_1995\_1999 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1985\_1994 : int 0 0 1 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1975\_1984 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1965\_1974 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1900\_1964 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1824\_1899 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ original\_concatenated\_title\_abstract: chr "The Effect of Stellar Rotation on Colour-Magnitude Diagrams: On the apparent presence of multiple populations i"| \_\_truncated\_\_ "Galactic chemical evolution in hierarchical formation models - I. Early-type galaxies in the local Universe We "| \_\_truncated\_\_ "Fate of the companion stars of ultra-rapid pulsars Abstract A millisecond pulsar that is formed by spin-up 'rec"| \_\_truncated\_\_ "Infrared Molecular Starburst Fingerprints in Deeply Obscured (Ultra)Luminous Infrared Galaxy Nuclei High-resolu"| \_\_truncated\_\_ ...

# Initialize an empty formula string  
prevalence\_formula\_str <- "~"  
  
# Define the publication intervals  
pub\_intervals <- c("pub\_interval\_2020\_2022", "pub\_interval\_2015\_2019", "pub\_interval\_2010\_2014",   
 "pub\_interval\_2005\_2009", "pub\_interval\_2000\_2004", "pub\_interval\_1995\_1999",  
 "pub\_interval\_1985\_1994", "pub\_interval\_1975\_1984", "pub\_interval\_1965\_1974",  
 "pub\_interval\_1900\_1964", "pub\_interval\_1824\_1899")  
  
# Add each publication interval to the formula string  
for (interval in pub\_intervals) {  
 # add an if statement to handle the first addition (without '+')  
 if (prevalence\_formula\_str == "~") {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, interval)  
 } else {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, "+", interval)  
 }  
}  
  
  
# Convert the string to a formula  
prevalence\_formula <- as.formula(prevalence\_formula\_str)  
print(prevalence\_formula)

## ~pub\_interval\_2020\_2022 + pub\_interval\_2015\_2019 + pub\_interval\_2010\_2014 +   
## pub\_interval\_2005\_2009 + pub\_interval\_2000\_2004 + pub\_interval\_1995\_1999 +   
## pub\_interval\_1985\_1994 + pub\_interval\_1975\_1984 + pub\_interval\_1965\_1974 +   
## pub\_interval\_1900\_1964 + pub\_interval\_1824\_1899

Kvals <- seq(from = 5, to = 40, by = 5) # K values to try  
search\_results <- searchK(documents = out\_text$documents,   
 vocab = out\_text$vocab,   
 K = Kvals,   
 prevalence = prevalence\_formula,   
 data = out\_text$meta,   
 init.type = "Spectral",   
 verbose = FALSE)  
  
# Plot the results  
plot(search\_results)



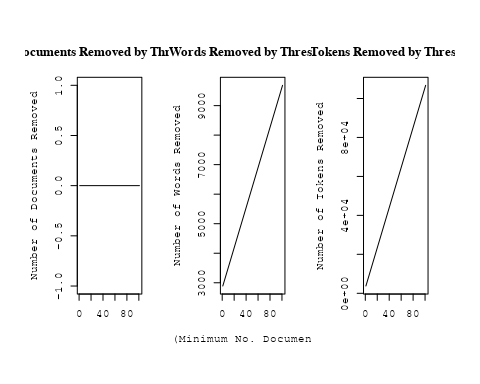
##Topic generation for FR (in collaboration) publications  
  
data\_collab <- data[data[["FR"]] != 0,]  
  
# Save the original title data for future use  
data\_collab$original\_concatenated\_title\_abstract <- data\_collab$concatenated\_title\_abstract  
  
#pre-processing the titles using textProcessor from the stm package  
processed\_text <- textProcessor(data\_collab$concatenated\_title\_abstract, metadata = data\_collab)

## Building corpus...   
## Converting to Lower Case...   
## Removing punctuation...   
## Removing stopwords...   
## Removing numbers...   
## Stemming...   
## Creating Output...

# Further prepare the data by removing low-frequency terms  
out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta)

## Removing 2890 of 10117 terms (2890 of 175894 tokens) due to frequency   
## Your corpus now has 2060 documents, 7227 terms and 173004 tokens.

docs\_text <- out\_text$documents  
vocab\_text <- out\_text$vocab  
meta\_text <- out\_text$meta  
  
  
#Prepare data  
plotRemoved(processed\_text$documents, lower.thresh = seq(1, 200, by = 100))



out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta, lower.thresh = 5)

## Removing 7210 of 10117 terms (14884 of 175894 tokens) due to frequency   
## Your corpus now has 2060 documents, 2907 terms and 161010 tokens.

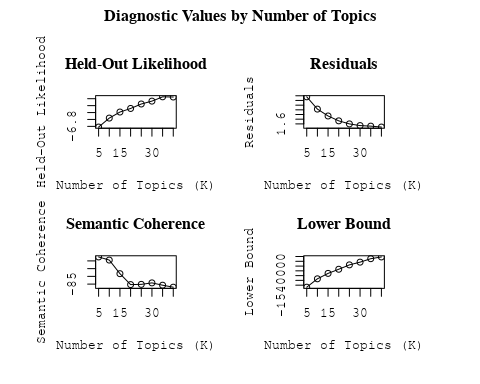
str(out\_text$meta)

## 'data.frame': 2060 obs. of 39 variables:  
## $ concept\_id : chr "https://openalex.org/C1276947" "https://openalex.org/C44870925" "https://openalex.org/C44870925" "https://openalex.org/C44870925" ...  
## $ work\_id : chr "https://openalex.org/W3099195213" "https://openalex.org/W2072757986" "https://openalex.org/W3104422960" "https://openalex.org/W2007504429" ...  
## $ publication\_year : int 1996 1977 2004 2011 2020 2008 2004 2012 2008 2020 ...  
## $ title : chr "Angular Sizes of Faint Field Disk Galaxies: Intrinsic Luminosity Evolution" "Hourly variations in O VI P Cygni profiles of hot stars" "The star formation history of Seyfert 2 nuclei" "A Pluto-like radius and a high albedo for the dwarf planet Eris from an occultation" ...  
## $ paperabstract : chr "In order to explain the small scale lengths detected in the recent deep field observations performed from large"| \_\_truncated\_\_ "Significant changes in the ultraviolet absorption profiles are reported in the spectra of Delta Ori A, Iota Ori"| \_\_truncated\_\_ "We present a study of the stellar populations in the central ∼200 pc of a large and homogeneous sample comprisi"| \_\_truncated\_\_ "The dwarf planet Eris is a trans-Neptunian object with an orbital eccentricity of 0.44, an inclination of 44 de"| \_\_truncated\_\_ ...  
## $ country : chr "FR US" "FR US" "BR FR FR BR" "MX FR MX FR" ...  
## $ year\_concept : chr "1996+https://openalex.org/C1276947" "1977+https://openalex.org/C44870925" "2004+https://openalex.org/C44870925" "2011+https://openalex.org/C44870925" ...  
## $ concatenated\_title\_abstract : chr "Angular Sizes of Faint Field Disk Galaxies: Intrinsic Luminosity Evolution In order to explain the small scale "| \_\_truncated\_\_ "Hourly variations in O VI P Cygni profiles of hot stars Significant changes in the ultraviolet absorption profi"| \_\_truncated\_\_ "The star formation history of Seyfert 2 nuclei We present a study of the stellar populations in the central ∼20"| \_\_truncated\_\_ "A Pluto-like radius and a high albedo for the dwarf planet Eris from an occultation The dwarf planet Eris is a "| \_\_truncated\_\_ ...  
## $ US : num 50 50 0 0 0 50 0 50 0 0 ...  
## $ IN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DE : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CH : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ GB : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ FR : num 50 50 50 50 100 50 50 50 50 100 ...  
## $ IT : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ RU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CA : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ NL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AU : num 0 0 0 0 0 0 50 0 0 0 ...  
## $ JP : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ES : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Americas : num 0 0 50 50 0 0 0 0 0 0 ...  
## $ Europe : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Africa : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AsiaAndOceania : num 0 0 0 0 0 0 0 0 50 0 ...  
## $ pub\_interval\_2020\_2022 : int 0 0 0 0 1 0 0 0 0 1 ...  
## $ pub\_interval\_2015\_2019 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2010\_2014 : int 0 0 0 1 0 0 0 1 0 0 ...  
## $ pub\_interval\_2005\_2009 : int 0 0 0 0 0 1 0 0 1 0 ...  
## $ pub\_interval\_2000\_2004 : int 0 0 1 0 0 0 1 0 0 0 ...  
## $ pub\_interval\_1995\_1999 : int 1 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1985\_1994 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1975\_1984 : int 0 1 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1965\_1974 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1900\_1964 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1824\_1899 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ original\_concatenated\_title\_abstract: chr "Angular Sizes of Faint Field Disk Galaxies: Intrinsic Luminosity Evolution In order to explain the small scale "| \_\_truncated\_\_ "Hourly variations in O VI P Cygni profiles of hot stars Significant changes in the ultraviolet absorption profi"| \_\_truncated\_\_ "The star formation history of Seyfert 2 nuclei We present a study of the stellar populations in the central ∼20"| \_\_truncated\_\_ "A Pluto-like radius and a high albedo for the dwarf planet Eris from an occultation The dwarf planet Eris is a "| \_\_truncated\_\_ ...

# Initialize an empty formula string  
prevalence\_formula\_str <- "~"  
  
# Define the publication intervals  
pub\_intervals <- c("pub\_interval\_2020\_2022", "pub\_interval\_2015\_2019", "pub\_interval\_2010\_2014",   
 "pub\_interval\_2005\_2009", "pub\_interval\_2000\_2004", "pub\_interval\_1995\_1999",  
 "pub\_interval\_1985\_1994", "pub\_interval\_1975\_1984", "pub\_interval\_1965\_1974",  
 "pub\_interval\_1900\_1964", "pub\_interval\_1824\_1899")  
  
# Add each publication interval to the formula string  
for (interval in pub\_intervals) {  
 # add an if statement to handle the first addition (without '+')  
 if (prevalence\_formula\_str == "~") {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, interval)  
 } else {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, "+", interval)  
 }  
}  
  
  
# Convert the string to a formula  
prevalence\_formula <- as.formula(prevalence\_formula\_str)  
print(prevalence\_formula)

## ~pub\_interval\_2020\_2022 + pub\_interval\_2015\_2019 + pub\_interval\_2010\_2014 +   
## pub\_interval\_2005\_2009 + pub\_interval\_2000\_2004 + pub\_interval\_1995\_1999 +   
## pub\_interval\_1985\_1994 + pub\_interval\_1975\_1984 + pub\_interval\_1965\_1974 +   
## pub\_interval\_1900\_1964 + pub\_interval\_1824\_1899

Kvals <- seq(from = 5, to = 40, by = 5) # K values to try  
search\_results <- searchK(documents = out\_text$documents,   
 vocab = out\_text$vocab,   
 K = Kvals,   
 prevalence = prevalence\_formula,   
 data = out\_text$meta,   
 init.type = "Spectral",   
 verbose = FALSE)  
  
# Plot the results  
plot(search\_results)



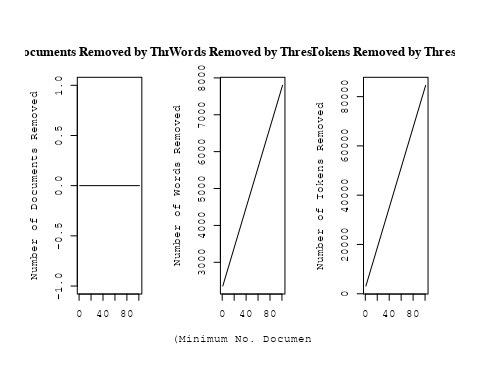
##Topic generation for RU (in collaboration) publications  
  
data\_collab <- data[data[["RU"]] != 0,]  
  
# Save the original title data for future use  
data\_collab$original\_concatenated\_title\_abstract <- data\_collab$concatenated\_title\_abstract  
  
#pre-processing the titles using textProcessor from the stm package  
processed\_text <- textProcessor(data\_collab$concatenated\_title\_abstract, metadata = data\_collab)

## Building corpus...   
## Converting to Lower Case...   
## Removing punctuation...   
## Removing stopwords...   
## Removing numbers...   
## Stemming...   
## Creating Output...

# Further prepare the data by removing low-frequency terms  
out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta)

## Removing 2357 of 8083 terms (2357 of 118251 tokens) due to frequency   
## Your corpus now has 1629 documents, 5726 terms and 115894 tokens.

docs\_text <- out\_text$documents  
vocab\_text <- out\_text$vocab  
meta\_text <- out\_text$meta  
  
  
#Prepare data  
plotRemoved(processed\_text$documents, lower.thresh = seq(1, 200, by = 100))



out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta, lower.thresh = 5)

## Removing 5765 of 8083 terms (11687 of 118251 tokens) due to frequency   
## Your corpus now has 1629 documents, 2318 terms and 106564 tokens.

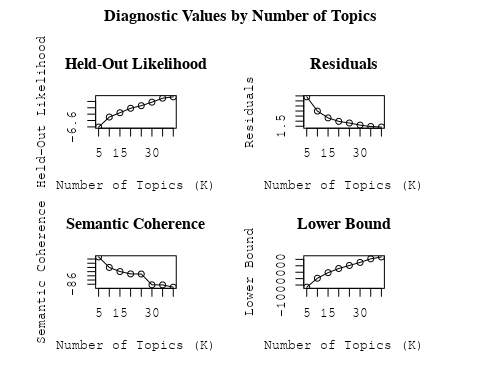
str(out\_text$meta)

## 'data.frame': 1629 obs. of 39 variables:  
## $ concept\_id : chr "https://openalex.org/C44870925" "https://openalex.org/C44870925" "https://openalex.org/C44870925" "https://openalex.org/C44870925" ...  
## $ work\_id : chr "https://openalex.org/W2091953534" "https://openalex.org/W2950971381" "https://openalex.org/W1974527970" "https://openalex.org/W2798563684" ...  
## $ publication\_year : int 2004 2010 1989 2018 2002 2014 2006 2006 1984 2014 ...  
## $ title : chr "Inner Polar Rings in Regular Lenticular Galaxies" "Gravitational lensing in a non-uniform plasma" "A survey of the radio continuum emission of RS Canum Venaticorum and related active binary systems" "Magnetic structure of solar flare regions producing hard X-ray pulsations" ...  
## $ paperabstract : chr "We have investigated a sample of S0 galaxies, mostly with circumnuclear dust lanes orthogonal to their major ax"| \_\_truncated\_\_ "We develop a model of gravitational lensing in a non-uniform plasma. When a gravitating body is surrounded by a"| \_\_truncated\_\_ "Observations of 77 RS CVn and related active binary systems were made with VLA at a continuum frequency of 4.86"| \_\_truncated\_\_ "Abstract We present analysis of the magnetic field in seven solar flare regions accompanied by the pulsations o"| \_\_truncated\_\_ ...  
## $ country : chr "RU RU GB" "RU RU RU RU RU RU RU RU" "RU RU RU RU" "CH RU GB RU CN" ...  
## $ year\_concept : chr "2004+https://openalex.org/C44870925" "2010+https://openalex.org/C44870925" "1989+https://openalex.org/C44870925" "2018+https://openalex.org/C44870925" ...  
## $ concatenated\_title\_abstract : chr "Inner Polar Rings in Regular Lenticular Galaxies We have investigated a sample of S0 galaxies, mostly with circ"| \_\_truncated\_\_ "Gravitational lensing in a non-uniform plasma We develop a model of gravitational lensing in a non-uniform plas"| \_\_truncated\_\_ "A survey of the radio continuum emission of RS Canum Venaticorum and related active binary systems Observations"| \_\_truncated\_\_ "Magnetic structure of solar flare regions producing hard X-ray pulsations Abstract We present analysis of the m"| \_\_truncated\_\_ ...  
## $ US : num 0 0 0 0 0 0 0 0 50 0 ...  
## $ IN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DE : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CH : num 0 0 0 20 0 0 0 0 0 0 ...  
## $ GB : num 33.3 0 0 20 0 ...  
## $ CN : num 0 0 0 20 0 0 0 0 0 0 ...  
## $ FR : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IT : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ RU : num 66.7 100 100 40 100 ...  
## $ CA : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ NL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ JP : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ES : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Americas : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Europe : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Africa : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AsiaAndOceania : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2020\_2022 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2015\_2019 : int 0 0 0 1 0 0 0 0 0 0 ...  
## $ pub\_interval\_2010\_2014 : int 0 1 0 0 0 1 0 0 0 1 ...  
## $ pub\_interval\_2005\_2009 : int 0 0 0 0 0 0 1 1 0 0 ...  
## $ pub\_interval\_2000\_2004 : int 1 0 0 0 1 0 0 0 0 0 ...  
## $ pub\_interval\_1995\_1999 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1985\_1994 : int 0 0 1 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1975\_1984 : int 0 0 0 0 0 0 0 0 1 0 ...  
## $ pub\_interval\_1965\_1974 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1900\_1964 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1824\_1899 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ original\_concatenated\_title\_abstract: chr "Inner Polar Rings in Regular Lenticular Galaxies We have investigated a sample of S0 galaxies, mostly with circ"| \_\_truncated\_\_ "Gravitational lensing in a non-uniform plasma We develop a model of gravitational lensing in a non-uniform plas"| \_\_truncated\_\_ "A survey of the radio continuum emission of RS Canum Venaticorum and related active binary systems Observations"| \_\_truncated\_\_ "Magnetic structure of solar flare regions producing hard X-ray pulsations Abstract We present analysis of the m"| \_\_truncated\_\_ ...

# Initialize an empty formula string  
prevalence\_formula\_str <- "~"  
  
# Define the publication intervals  
pub\_intervals <- c("pub\_interval\_2020\_2022", "pub\_interval\_2015\_2019", "pub\_interval\_2010\_2014",   
 "pub\_interval\_2005\_2009", "pub\_interval\_2000\_2004", "pub\_interval\_1995\_1999",  
 "pub\_interval\_1985\_1994", "pub\_interval\_1975\_1984", "pub\_interval\_1965\_1974",  
 "pub\_interval\_1900\_1964", "pub\_interval\_1824\_1899")  
  
# Add each publication interval to the formula string  
for (interval in pub\_intervals) {  
 # add an if statement to handle the first addition (without '+')  
 if (prevalence\_formula\_str == "~") {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, interval)  
 } else {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, "+", interval)  
 }  
}  
  
  
# Convert the string to a formula  
prevalence\_formula <- as.formula(prevalence\_formula\_str)  
print(prevalence\_formula)

## ~pub\_interval\_2020\_2022 + pub\_interval\_2015\_2019 + pub\_interval\_2010\_2014 +   
## pub\_interval\_2005\_2009 + pub\_interval\_2000\_2004 + pub\_interval\_1995\_1999 +   
## pub\_interval\_1985\_1994 + pub\_interval\_1975\_1984 + pub\_interval\_1965\_1974 +   
## pub\_interval\_1900\_1964 + pub\_interval\_1824\_1899

Kvals <- seq(from = 5, to = 40, by = 5) # K values to try  
search\_results <- searchK(documents = out\_text$documents,   
 vocab = out\_text$vocab,   
 K = Kvals,   
 prevalence = prevalence\_formula,   
 data = out\_text$meta,   
 init.type = "Spectral",   
 verbose = FALSE)  
  
# Plot the results  
plot(search\_results)



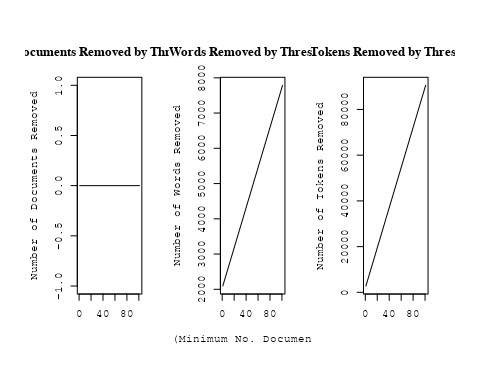
##Topic generation for ES (in collaboration) publications  
  
data\_collab <- data[data[["ES"]] != 0,]  
  
# Save the original title data for future use  
data\_collab$original\_concatenated\_title\_abstract <- data\_collab$concatenated\_title\_abstract  
  
#pre-processing the titles using textProcessor from the stm package  
processed\_text <- textProcessor(data\_collab$concatenated\_title\_abstract, metadata = data\_collab)

## Building corpus...   
## Converting to Lower Case...   
## Removing punctuation...   
## Removing stopwords...   
## Removing numbers...   
## Stemming...   
## Creating Output...

# Further prepare the data by removing low-frequency terms  
out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta)

## Removing 2094 of 8129 terms (2094 of 137739 tokens) due to frequency   
## Your corpus now has 1669 documents, 6035 terms and 135645 tokens.

docs\_text <- out\_text$documents  
vocab\_text <- out\_text$vocab  
meta\_text <- out\_text$meta  
  
  
#Prepare data  
plotRemoved(processed\_text$documents, lower.thresh = seq(1, 200, by = 100))



out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta, lower.thresh = 5)

## Removing 5581 of 8129 terms (12134 of 137739 tokens) due to frequency   
## Your corpus now has 1669 documents, 2548 terms and 125605 tokens.

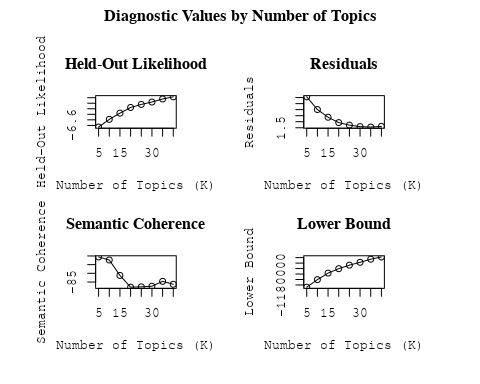
str(out\_text$meta)

## 'data.frame': 1669 obs. of 39 variables:  
## $ concept\_id : chr "https://openalex.org/C44870925" "https://openalex.org/C44870925" "https://openalex.org/C44870925" "https://openalex.org/C1276947" ...  
## $ work\_id : chr "https://openalex.org/W2122555548" "https://openalex.org/W3123041789" "https://openalex.org/W4206317057" "https://openalex.org/W4206317057" ...  
## $ publication\_year : int 2007 2003 2021 2021 2012 2007 2019 1998 1998 2011 ...  
## $ title : chr "Medium-resolution Isaac Newton Telescope library of empirical spectra - II. The stellar atmospheric parameters" "Detection of the ISW and SZ effects from the CMB-Galaxy correlation" "The effect of mission duration on LISA science objectives" "The effect of mission duration on LISA science objectives" ...  
## $ paperabstract : chr "We present a homogeneous set of stellar atmospheric parameters (T-eff, log g, [Fe/H]) for MILES, a new spectral"| \_\_truncated\_\_ "We present a cross-correlation analysis of the W ilkinson Microwave Anisotropy Probe cosmic microwave backgroun"| \_\_truncated\_\_ "The science objectives of the LISA mission have been defined under the implicit assumption of a 4-years continu"| \_\_truncated\_\_ "The science objectives of the LISA mission have been defined under the implicit assumption of a 4-years continu"| \_\_truncated\_\_ ...  
## $ country : chr "ES ES" "ES ES" "ES CN ES NL CN NL US US" "ES CN ES NL CN NL US US" ...  
## $ year\_concept : chr "2007+https://openalex.org/C44870925" "2003+https://openalex.org/C44870925" "2021+https://openalex.org/C44870925" "2021+https://openalex.org/C1276947" ...  
## $ concatenated\_title\_abstract : chr "Medium-resolution Isaac Newton Telescope library of empirical spectra - II. The stellar atmospheric parameters "| \_\_truncated\_\_ "Detection of the ISW and SZ effects from the CMB-Galaxy correlation We present a cross-correlation analysis of "| \_\_truncated\_\_ "The effect of mission duration on LISA science objectives The science objectives of the LISA mission have been "| \_\_truncated\_\_ "The effect of mission duration on LISA science objectives The science objectives of the LISA mission have been "| \_\_truncated\_\_ ...  
## $ US : num 0 0 25 25 0 ...  
## $ IN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DE : num 0 0 0 0 50 50 0 0 0 0 ...  
## $ CH : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ GB : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CN : num 0 0 25 25 0 0 0 0 0 0 ...  
## $ FR : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IT : num 0 0 0 0 0 0 0 0 0 50 ...  
## $ RU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CA : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ NL : num 0 0 25 25 0 0 0 0 0 0 ...  
## $ AU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ JP : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ES : num 100 100 25 25 50 ...  
## $ IL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Americas : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Europe : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Africa : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AsiaAndOceania : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2020\_2022 : int 0 0 1 1 0 0 0 0 0 0 ...  
## $ pub\_interval\_2015\_2019 : int 0 0 0 0 0 0 1 0 0 0 ...  
## $ pub\_interval\_2010\_2014 : int 0 0 0 0 1 0 0 0 0 1 ...  
## $ pub\_interval\_2005\_2009 : int 1 0 0 0 0 1 0 0 0 0 ...  
## $ pub\_interval\_2000\_2004 : int 0 1 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1995\_1999 : int 0 0 0 0 0 0 0 1 1 0 ...  
## $ pub\_interval\_1985\_1994 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1975\_1984 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1965\_1974 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1900\_1964 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1824\_1899 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ original\_concatenated\_title\_abstract: chr "Medium-resolution Isaac Newton Telescope library of empirical spectra - II. The stellar atmospheric parameters "| \_\_truncated\_\_ "Detection of the ISW and SZ effects from the CMB-Galaxy correlation We present a cross-correlation analysis of "| \_\_truncated\_\_ "The effect of mission duration on LISA science objectives The science objectives of the LISA mission have been "| \_\_truncated\_\_ "The effect of mission duration on LISA science objectives The science objectives of the LISA mission have been "| \_\_truncated\_\_ ...

# Initialize an empty formula string  
prevalence\_formula\_str <- "~"  
  
# Define the publication intervals  
pub\_intervals <- c("pub\_interval\_2020\_2022", "pub\_interval\_2015\_2019", "pub\_interval\_2010\_2014",   
 "pub\_interval\_2005\_2009", "pub\_interval\_2000\_2004", "pub\_interval\_1995\_1999",  
 "pub\_interval\_1985\_1994", "pub\_interval\_1975\_1984", "pub\_interval\_1965\_1974",  
 "pub\_interval\_1900\_1964", "pub\_interval\_1824\_1899")  
  
# Add each publication interval to the formula string  
for (interval in pub\_intervals) {  
 # add an if statement to handle the first addition (without '+')  
 if (prevalence\_formula\_str == "~") {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, interval)  
 } else {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, "+", interval)  
 }  
}  
  
  
# Convert the string to a formula  
prevalence\_formula <- as.formula(prevalence\_formula\_str)  
print(prevalence\_formula)

## ~pub\_interval\_2020\_2022 + pub\_interval\_2015\_2019 + pub\_interval\_2010\_2014 +   
## pub\_interval\_2005\_2009 + pub\_interval\_2000\_2004 + pub\_interval\_1995\_1999 +   
## pub\_interval\_1985\_1994 + pub\_interval\_1975\_1984 + pub\_interval\_1965\_1974 +   
## pub\_interval\_1900\_1964 + pub\_interval\_1824\_1899

Kvals <- seq(from = 5, to = 40, by = 5) # K values to try  
search\_results <- searchK(documents = out\_text$documents,   
 vocab = out\_text$vocab,   
 K = Kvals,   
 prevalence = prevalence\_formula,   
 data = out\_text$meta,   
 init.type = "Spectral",   
 verbose = FALSE)  
  
# Plot the results  
plot(search\_results)



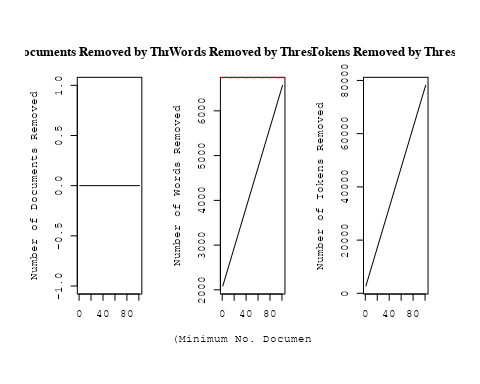
##Topic generation for IN (in collaboration) publications  
  
data\_collab <- data[data[["IN"]] != 0,]  
  
# Save the original title data for future use  
data\_collab$original\_concatenated\_title\_abstract <- data\_collab$concatenated\_title\_abstract  
  
#pre-processing the titles using textProcessor from the stm package  
processed\_text <- textProcessor(data\_collab$concatenated\_title\_abstract, metadata = data\_collab)

## Building corpus...   
## Converting to Lower Case...   
## Removing punctuation...   
## Removing stopwords...   
## Removing numbers...   
## Stemming...   
## Creating Output...

# Further prepare the data by removing low-frequency terms  
out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta)

## Removing 2086 of 6741 terms (2086 of 85067 tokens) due to frequency   
## Your corpus now has 1126 documents, 4655 terms and 82981 tokens.

docs\_text <- out\_text$documents  
vocab\_text <- out\_text$vocab  
meta\_text <- out\_text$meta  
  
  
#Prepare data  
plotRemoved(processed\_text$documents, lower.thresh = seq(1, 200, by = 100))



out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta, lower.thresh = 5)

## Removing 4702 of 6741 terms (9462 of 85067 tokens) due to frequency   
## Your corpus now has 1126 documents, 2039 terms and 75605 tokens.

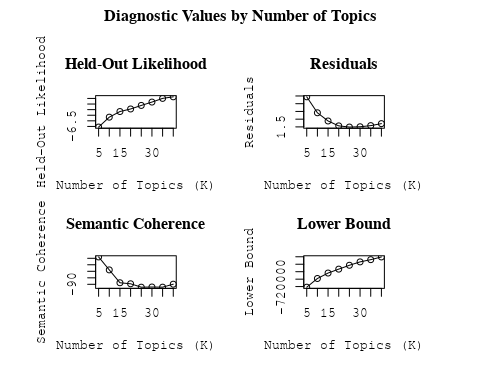
str(out\_text$meta)

## 'data.frame': 1126 obs. of 39 variables:  
## $ concept\_id : chr "https://openalex.org/C1276947" "https://openalex.org/C44870925" "https://openalex.org/C44870925" "https://openalex.org/C44870925" ...  
## $ work\_id : chr "https://openalex.org/W3104346433" "https://openalex.org/W2948771331" "https://openalex.org/W3101308139" "https://openalex.org/W1586945094" ...  
## $ publication\_year : int 2003 2019 2006 1998 2004 2007 2006 1997 2010 2010 ...  
## $ title : chr "The Origin of X-shaped Radio Galaxies: Clues from the Z-symmetric Secondary Lobes" "Observational constraints on viscous Ricci dark energy model" "Photometric and spectroscopic evolution of the type IIP supernova SN 2004et" "A stellar origin for the short-lived nuclides in the early Solar System" ...  
## $ paperabstract : chr "Existing radio images of a few X-shaped radio galaxies reveal Z-symmetric morphologies in their weaker secondar"| \_\_truncated\_\_ "In this paper, Ricci dark energy (RDE) model with bulk viscosity is studied to observe the cosmic accelerating "| \_\_truncated\_\_ "We present optical photometry and spectroscopy of the normal type IIP supernova SN2004A, which was discovered i"| \_\_truncated\_\_ "Primitive meteorites contain isotopes that are the decay products of short-lived nuclides in the early Solar Sy"| \_\_truncated\_\_ ...  
## $ country : chr "IN US" "IN IN" "IN IN" "US IN" ...  
## $ year\_concept : chr "2003+https://openalex.org/C1276947" "2019+https://openalex.org/C44870925" "2006+https://openalex.org/C44870925" "1998+https://openalex.org/C44870925" ...  
## $ concatenated\_title\_abstract : chr "The Origin of X-shaped Radio Galaxies: Clues from the Z-symmetric Secondary Lobes Existing radio images of a fe"| \_\_truncated\_\_ "Observational constraints on viscous Ricci dark energy model In this paper, Ricci dark energy (RDE) model with "| \_\_truncated\_\_ "Photometric and spectroscopic evolution of the type IIP supernova SN 2004et We present optical photometry and s"| \_\_truncated\_\_ "A stellar origin for the short-lived nuclides in the early Solar System Primitive meteorites contain isotopes t"| \_\_truncated\_\_ ...  
## $ US : num 50 0 0 50 50 0 0 0 0 0 ...  
## $ IN : num 50 100 100 50 50 100 100 100 100 100 ...  
## $ DE : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CH : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ GB : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ FR : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IT : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ RU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CA : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ NL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ JP : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ES : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Americas : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Europe : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Africa : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AsiaAndOceania : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2020\_2022 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2015\_2019 : int 0 1 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2010\_2014 : int 0 0 0 0 0 0 0 0 1 1 ...  
## $ pub\_interval\_2005\_2009 : int 0 0 1 0 0 1 1 0 0 0 ...  
## $ pub\_interval\_2000\_2004 : int 1 0 0 0 1 0 0 0 0 0 ...  
## $ pub\_interval\_1995\_1999 : int 0 0 0 1 0 0 0 1 0 0 ...  
## $ pub\_interval\_1985\_1994 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1975\_1984 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1965\_1974 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1900\_1964 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1824\_1899 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ original\_concatenated\_title\_abstract: chr "The Origin of X-shaped Radio Galaxies: Clues from the Z-symmetric Secondary Lobes Existing radio images of a fe"| \_\_truncated\_\_ "Observational constraints on viscous Ricci dark energy model In this paper, Ricci dark energy (RDE) model with "| \_\_truncated\_\_ "Photometric and spectroscopic evolution of the type IIP supernova SN 2004et We present optical photometry and s"| \_\_truncated\_\_ "A stellar origin for the short-lived nuclides in the early Solar System Primitive meteorites contain isotopes t"| \_\_truncated\_\_ ...

# Initialize an empty formula string  
prevalence\_formula\_str <- "~"  
  
# Define the publication intervals  
pub\_intervals <- c("pub\_interval\_2020\_2022", "pub\_interval\_2015\_2019", "pub\_interval\_2010\_2014",   
 "pub\_interval\_2005\_2009", "pub\_interval\_2000\_2004", "pub\_interval\_1995\_1999",  
 "pub\_interval\_1985\_1994", "pub\_interval\_1975\_1984", "pub\_interval\_1965\_1974",  
 "pub\_interval\_1900\_1964", "pub\_interval\_1824\_1899")  
  
# Add each publication interval to the formula string  
for (interval in pub\_intervals) {  
 # add an if statement to handle the first addition (without '+')  
 if (prevalence\_formula\_str == "~") {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, interval)  
 } else {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, "+", interval)  
 }  
}  
  
  
# Convert the string to a formula  
prevalence\_formula <- as.formula(prevalence\_formula\_str)  
print(prevalence\_formula)

## ~pub\_interval\_2020\_2022 + pub\_interval\_2015\_2019 + pub\_interval\_2010\_2014 +   
## pub\_interval\_2005\_2009 + pub\_interval\_2000\_2004 + pub\_interval\_1995\_1999 +   
## pub\_interval\_1985\_1994 + pub\_interval\_1975\_1984 + pub\_interval\_1965\_1974 +   
## pub\_interval\_1900\_1964 + pub\_interval\_1824\_1899

Kvals <- seq(from = 5, to = 40, by = 5) # K values to try  
search\_results <- searchK(documents = out\_text$documents,   
 vocab = out\_text$vocab,   
 K = Kvals,   
 prevalence = prevalence\_formula,   
 data = out\_text$meta,   
 init.type = "Spectral",   
 verbose = FALSE)  
  
# Plot the results  
plot(search\_results)



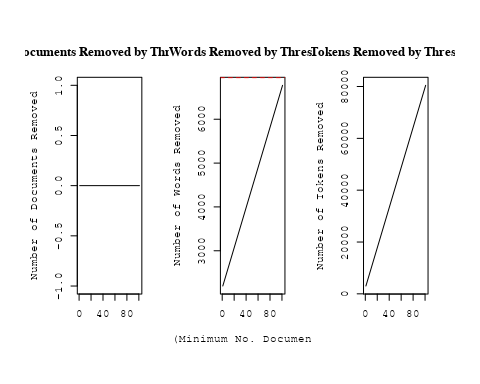
##Topic generation for AsiaAndOceania (in collaboration) publications  
  
data\_collab <- data[data[["AsiaAndOceania"]] != 0,]  
  
# Save the original title data for future use  
data\_collab$original\_concatenated\_title\_abstract <- data\_collab$concatenated\_title\_abstract  
  
#pre-processing the titles using textProcessor from the stm package  
processed\_text <- textProcessor(data\_collab$concatenated\_title\_abstract, metadata = data\_collab)

## Building corpus...   
## Converting to Lower Case...   
## Removing punctuation...   
## Removing stopwords...   
## Removing numbers...   
## Stemming...   
## Creating Output...

# Further prepare the data by removing low-frequency terms  
out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta)

## Removing 2196 of 6950 terms (2196 of 88052 tokens) due to frequency   
## Your corpus now has 1136 documents, 4754 terms and 85856 tokens.

docs\_text <- out\_text$documents  
vocab\_text <- out\_text$vocab  
meta\_text <- out\_text$meta  
  
  
#Prepare data  
plotRemoved(processed\_text$documents, lower.thresh = seq(1, 200, by = 100))



out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta, lower.thresh = 5)

## Removing 4918 of 6950 terms (9924 of 88052 tokens) due to frequency   
## Your corpus now has 1136 documents, 2032 terms and 78128 tokens.

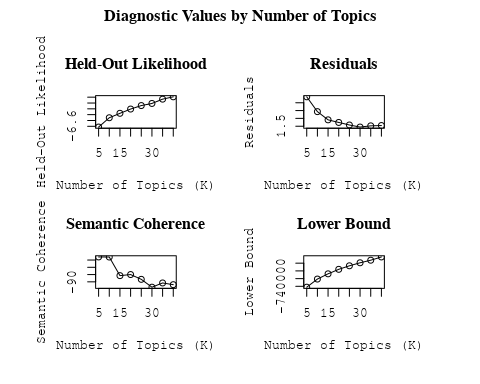
str(out\_text$meta)

## 'data.frame': 1136 obs. of 39 variables:  
## $ concept\_id : chr "https://openalex.org/C44870925" "https://openalex.org/C44870925" "https://openalex.org/C1276947" "https://openalex.org/C44870925" ...  
## $ work\_id : chr "https://openalex.org/W1971228182" "https://openalex.org/W3121271699" "https://openalex.org/W2312475649" "https://openalex.org/W2020890029" ...  
## $ publication\_year : int 2003 2008 2013 2007 2007 1999 2020 2003 2018 2000 ...  
## $ title : chr "Limits on the evolution of galaxies from the statistics of gravitational lenses" "Lyman alpha emitting galaxies at 0.2 < z < 0.35 from GALEX spectroscopy" "Death of Massive Stars (IAU S279): Supernovae and Gamma-Ray Bursts, Proceedings of the International Astronomic"| \_\_truncated\_\_ "The NIRSPEC Brown Dwarf Spectroscopic Survey. II. High-Resolution J-Band Spectra of M, L, and T Dwarfs" ...  
## $ paperabstract : chr "We use gravitational lenses from the Cosmic Lens All-Sky Survey to constrain the evolution of galaxies since re"| \_\_truncated\_\_ "The GALEX (Galaxy Evolution Explorer) spectroscopic survey mode, with a resolution of about 8 A in the FUV (135"| \_\_truncated\_\_ "students and, sometimes, even in scholars. While motion and change are the nature of the world – to quote the b"| \_\_truncated\_\_ "We present a sequence of high-resolution (R ~ 20,000, or 15 km s^(-1)) infrared spectra of stars and brown dwar"| \_\_truncated\_\_ ...  
## $ country : chr "KR GB GB KR" "KR FR KR FR" "TW" "US KR" ...  
## $ year\_concept : chr "2003+https://openalex.org/C44870925" "2008+https://openalex.org/C44870925" "2013+https://openalex.org/C1276947" "2007+https://openalex.org/C44870925" ...  
## $ concatenated\_title\_abstract : chr "Limits on the evolution of galaxies from the statistics of gravitational lenses We use gravitational lenses fro"| \_\_truncated\_\_ "Lyman alpha emitting galaxies at 0.2 < z < 0.35 from GALEX spectroscopy The GALEX (Galaxy Evolution Explorer) s"| \_\_truncated\_\_ "Death of Massive Stars (IAU S279): Supernovae and Gamma-Ray Bursts, Proceedings of the International Astronomic"| \_\_truncated\_\_ "The NIRSPEC Brown Dwarf Spectroscopic Survey. II. High-Resolution J-Band Spectra of M, L, and T Dwarfs We prese"| \_\_truncated\_\_ ...  
## $ US : num 0 0 0 50 50 0 0 0 25 0 ...  
## $ IN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DE : num 0 0 0 0 0 0 0 0 50 0 ...  
## $ CH : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ GB : num 50 0 0 0 0 0 0 0 0 0 ...  
## $ CN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ FR : num 0 50 0 0 0 0 0 0 0 0 ...  
## $ IT : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ RU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CA : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ NL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ JP : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ES : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Americas : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Europe : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Africa : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AsiaAndOceania : num 50 50 100 50 50 100 100 100 25 100 ...  
## $ pub\_interval\_2020\_2022 : int 0 0 0 0 0 0 1 0 0 0 ...  
## $ pub\_interval\_2015\_2019 : int 0 0 0 0 0 0 0 0 1 0 ...  
## $ pub\_interval\_2010\_2014 : int 0 0 1 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2005\_2009 : int 0 1 0 1 1 0 0 0 0 0 ...  
## $ pub\_interval\_2000\_2004 : int 1 0 0 0 0 0 0 1 0 1 ...  
## $ pub\_interval\_1995\_1999 : int 0 0 0 0 0 1 0 0 0 0 ...  
## $ pub\_interval\_1985\_1994 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1975\_1984 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1965\_1974 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1900\_1964 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1824\_1899 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ original\_concatenated\_title\_abstract: chr "Limits on the evolution of galaxies from the statistics of gravitational lenses We use gravitational lenses fro"| \_\_truncated\_\_ "Lyman alpha emitting galaxies at 0.2 < z < 0.35 from GALEX spectroscopy The GALEX (Galaxy Evolution Explorer) s"| \_\_truncated\_\_ "Death of Massive Stars (IAU S279): Supernovae and Gamma-Ray Bursts, Proceedings of the International Astronomic"| \_\_truncated\_\_ "The NIRSPEC Brown Dwarf Spectroscopic Survey. II. High-Resolution J-Band Spectra of M, L, and T Dwarfs We prese"| \_\_truncated\_\_ ...

# Initialize an empty formula string  
prevalence\_formula\_str <- "~"  
  
# Define the publication intervals  
pub\_intervals <- c("pub\_interval\_2020\_2022", "pub\_interval\_2015\_2019", "pub\_interval\_2010\_2014",   
 "pub\_interval\_2005\_2009", "pub\_interval\_2000\_2004", "pub\_interval\_1995\_1999",  
 "pub\_interval\_1985\_1994", "pub\_interval\_1975\_1984", "pub\_interval\_1965\_1974",  
 "pub\_interval\_1900\_1964", "pub\_interval\_1824\_1899")  
  
# Add each publication interval to the formula string  
for (interval in pub\_intervals) {  
 # add an if statement to handle the first addition (without '+')  
 if (prevalence\_formula\_str == "~") {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, interval)  
 } else {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, "+", interval)  
 }  
}  
  
  
# Convert the string to a formula  
prevalence\_formula <- as.formula(prevalence\_formula\_str)  
print(prevalence\_formula)

## ~pub\_interval\_2020\_2022 + pub\_interval\_2015\_2019 + pub\_interval\_2010\_2014 +   
## pub\_interval\_2005\_2009 + pub\_interval\_2000\_2004 + pub\_interval\_1995\_1999 +   
## pub\_interval\_1985\_1994 + pub\_interval\_1975\_1984 + pub\_interval\_1965\_1974 +   
## pub\_interval\_1900\_1964 + pub\_interval\_1824\_1899

Kvals <- seq(from = 5, to = 40, by = 5) # K values to try  
search\_results <- searchK(documents = out\_text$documents,   
 vocab = out\_text$vocab,   
 K = Kvals,   
 prevalence = prevalence\_formula,   
 data = out\_text$meta,   
 init.type = "Spectral",   
 verbose = FALSE)  
  
# Plot the results  
plot(search\_results)



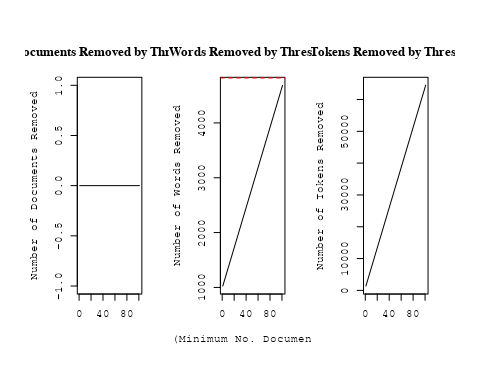
##Topic generation for IL (in collaboration) publications  
  
data\_collab <- data[data[["IL"]] != 0,]  
  
# Save the original title data for future use  
data\_collab$original\_concatenated\_title\_abstract <- data\_collab$concatenated\_title\_abstract  
  
#pre-processing the titles using textProcessor from the stm package  
processed\_text <- textProcessor(data\_collab$concatenated\_title\_abstract, metadata = data\_collab)

## Building corpus...   
## Converting to Lower Case...   
## Removing punctuation...   
## Removing stopwords...   
## Removing numbers...   
## Stemming...   
## Creating Output...

# Further prepare the data by removing low-frequency terms  
out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta)

## Removing 1026 of 4819 terms (1026 of 67316 tokens) due to frequency   
## Your corpus now has 868 documents, 3793 terms and 66290 tokens.

docs\_text <- out\_text$documents  
vocab\_text <- out\_text$vocab  
meta\_text <- out\_text$meta  
  
  
#Prepare data  
plotRemoved(processed\_text$documents, lower.thresh = seq(1, 200, by = 100))



out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta, lower.thresh = 5)

## Removing 3115 of 4819 terms (7020 of 67316 tokens) due to frequency   
## Your corpus now has 868 documents, 1704 terms and 60296 tokens.

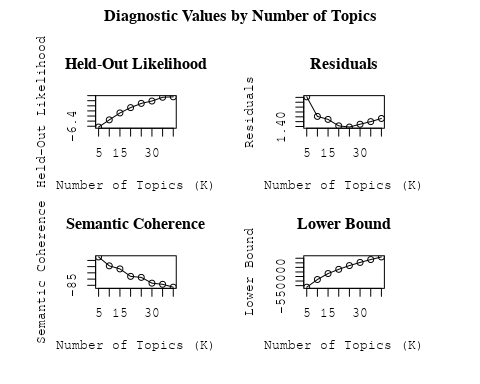
str(out\_text$meta)

## 'data.frame': 868 obs. of 39 variables:  
## $ concept\_id : chr "https://openalex.org/C1276947" "https://openalex.org/C44870925" "https://openalex.org/C44870925" "https://openalex.org/C44870925" ...  
## $ work\_id : chr "https://openalex.org/W2001391456" "https://openalex.org/W3105164062" "https://openalex.org/W1967579928" "https://openalex.org/W2057332312" ...  
## $ publication\_year : int 2011 2002 2005 1989 2009 2001 2004 2016 2006 2006 ...  
## $ title : chr "Supernova SN 2011fe from an exploding carbon–oxygen white dwarf star" "Quantification of Uncertainty in the Measurement of Magnetic Fields in Clusters of Galaxies" "An intriguing correlation between the masses and periods of the transiting planets" "Recovering the full velocity and density fields from large-scale redshift-distance samples" ...  
## $ paperabstract : chr "Type Ia supernovae have been used empirically as ‘standard candles’ to demonstrate the acceleration of the expa"| \_\_truncated\_\_ "We assess the principal statistical and physical uncertainties associated with the determination of magnetic fi"| \_\_truncated\_\_ "We point out an intriguing relation between the masses of the transiting planets and their orbital periods. For"| \_\_truncated\_\_ "A new method for extracting the large-scale three-dimensional velocity and mass density fields from measurement"| \_\_truncated\_\_ ...  
## $ country : chr "IL US US US US US US US IL US" "US IL IL US" "IL" "IL US" ...  
## $ year\_concept : chr "2011+https://openalex.org/C1276947" "2002+https://openalex.org/C44870925" "2005+https://openalex.org/C44870925" "1989+https://openalex.org/C44870925" ...  
## $ concatenated\_title\_abstract : chr "Supernova SN 2011fe from an exploding carbon–oxygen white dwarf star Type Ia supernovae have been used empirica"| \_\_truncated\_\_ "Quantification of Uncertainty in the Measurement of Magnetic Fields in Clusters of Galaxies We assess the princ"| \_\_truncated\_\_ "An intriguing correlation between the masses and periods of the transiting planets We point out an intriguing r"| \_\_truncated\_\_ "Recovering the full velocity and density fields from large-scale redshift-distance samples A new method for ext"| \_\_truncated\_\_ ...  
## $ US : num 80 50 0 50 50 0 0 0 0 0 ...  
## $ IN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DE : num 0 0 0 0 0 ...  
## $ CH : num 0 0 0 0 0 ...  
## $ GB : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ FR : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IT : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ RU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CA : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ NL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ JP : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ES : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IL : num 20 50 100 50 50 ...  
## $ Americas : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Europe : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Africa : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AsiaAndOceania : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2020\_2022 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2015\_2019 : int 0 0 0 0 0 0 0 1 0 0 ...  
## $ pub\_interval\_2010\_2014 : int 1 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2005\_2009 : int 0 0 1 0 1 0 0 0 1 1 ...  
## $ pub\_interval\_2000\_2004 : int 0 1 0 0 0 1 1 0 0 0 ...  
## $ pub\_interval\_1995\_1999 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1985\_1994 : int 0 0 0 1 0 0 0 0 0 0 ...  
## $ pub\_interval\_1975\_1984 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1965\_1974 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1900\_1964 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1824\_1899 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ original\_concatenated\_title\_abstract: chr "Supernova SN 2011fe from an exploding carbon–oxygen white dwarf star Type Ia supernovae have been used empirica"| \_\_truncated\_\_ "Quantification of Uncertainty in the Measurement of Magnetic Fields in Clusters of Galaxies We assess the princ"| \_\_truncated\_\_ "An intriguing correlation between the masses and periods of the transiting planets We point out an intriguing r"| \_\_truncated\_\_ "Recovering the full velocity and density fields from large-scale redshift-distance samples A new method for ext"| \_\_truncated\_\_ ...

# Initialize an empty formula string  
prevalence\_formula\_str <- "~"  
  
# Define the publication intervals  
pub\_intervals <- c("pub\_interval\_2020\_2022", "pub\_interval\_2015\_2019", "pub\_interval\_2010\_2014",   
 "pub\_interval\_2005\_2009", "pub\_interval\_2000\_2004", "pub\_interval\_1995\_1999",  
 "pub\_interval\_1985\_1994", "pub\_interval\_1975\_1984", "pub\_interval\_1965\_1974",  
 "pub\_interval\_1900\_1964", "pub\_interval\_1824\_1899")  
  
# Add each publication interval to the formula string  
for (interval in pub\_intervals) {  
 # add an if statement to handle the first addition (without '+')  
 if (prevalence\_formula\_str == "~") {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, interval)  
 } else {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, "+", interval)  
 }  
}  
  
  
# Convert the string to a formula  
prevalence\_formula <- as.formula(prevalence\_formula\_str)  
print(prevalence\_formula)

## ~pub\_interval\_2020\_2022 + pub\_interval\_2015\_2019 + pub\_interval\_2010\_2014 +   
## pub\_interval\_2005\_2009 + pub\_interval\_2000\_2004 + pub\_interval\_1995\_1999 +   
## pub\_interval\_1985\_1994 + pub\_interval\_1975\_1984 + pub\_interval\_1965\_1974 +   
## pub\_interval\_1900\_1964 + pub\_interval\_1824\_1899

Kvals <- seq(from = 5, to = 40, by = 5) # K values to try  
search\_results <- searchK(documents = out\_text$documents,   
 vocab = out\_text$vocab,   
 K = Kvals,   
 prevalence = prevalence\_formula,   
 data = out\_text$meta,   
 init.type = "Spectral",   
 verbose = FALSE)  
  
# Plot the results  
plot(search\_results)



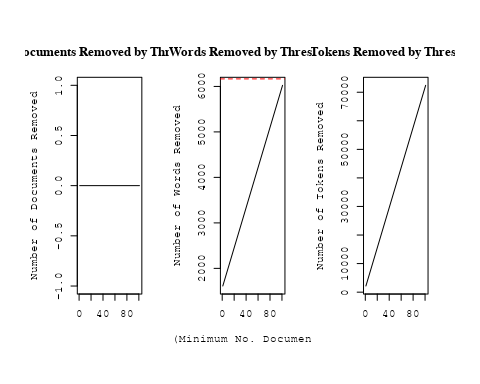
##Topic generation for CH (in collaboration) publications  
  
data\_collab <- data[data[["CH"]] != 0,]  
  
# Save the original title data for future use  
data\_collab$original\_concatenated\_title\_abstract <- data\_collab$concatenated\_title\_abstract  
  
#pre-processing the titles using textProcessor from the stm package  
processed\_text <- textProcessor(data\_collab$concatenated\_title\_abstract, metadata = data\_collab)

## Building corpus...   
## Converting to Lower Case...   
## Removing punctuation...   
## Removing stopwords...   
## Removing numbers...   
## Stemming...   
## Creating Output...

# Further prepare the data by removing low-frequency terms  
out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta)

## Removing 1612 of 6166 terms (1612 of 76116 tokens) due to frequency   
## Your corpus now has 899 documents, 4554 terms and 74504 tokens.

docs\_text <- out\_text$documents  
vocab\_text <- out\_text$vocab  
meta\_text <- out\_text$meta  
  
  
#Prepare data  
plotRemoved(processed\_text$documents, lower.thresh = seq(1, 200, by = 100))



out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta, lower.thresh = 5)

## Removing 4264 of 6166 terms (8961 of 76116 tokens) due to frequency   
## Your corpus now has 899 documents, 1902 terms and 67155 tokens.

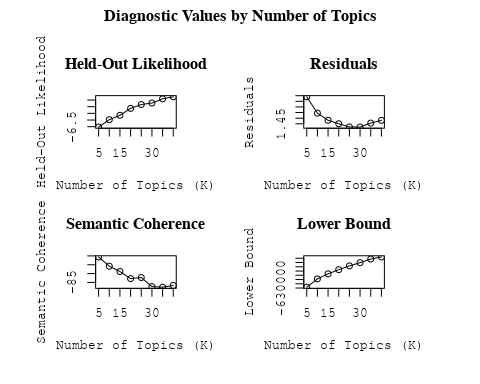
str(out\_text$meta)

## 'data.frame': 899 obs. of 39 variables:  
## $ concept\_id : chr "https://openalex.org/C44870925" "https://openalex.org/C44870925" "https://openalex.org/C44870925" "https://openalex.org/C44870925" ...  
## $ work\_id : chr "https://openalex.org/W2055827975" "https://openalex.org/W2798563684" "https://openalex.org/W2167125520" "https://openalex.org/W2174274577" ...  
## $ publication\_year : int 2007 2018 2003 2010 2012 2009 2010 1997 2009 2011 ...  
## $ title : chr "Cosmic magnetic fields and the CMB" "Magnetic structure of solar flare regions producing hard X-ray pulsations" "Dark matter and dark energy: summary and future directions" "WASP-32b: A Transiting Hot Jupiter Planet Orbiting a Lithium-Poor, Solar-Type Star" ...  
## $ paperabstract : chr "Abstract I describe the imprint which primordial magnetic fields can leave on the cosmic microwave background ("| \_\_truncated\_\_ "Abstract We present analysis of the magnetic field in seven solar flare regions accompanied by the pulsations o"| \_\_truncated\_\_ "This paper reviews the progress reported at the Discussion Meeting and advertises some possible future directio"| \_\_truncated\_\_ "ABSTRACT. We report the discovery of a transiting planet orbiting the star TYC 2-1155-1. The star, WASP-32, is "| \_\_truncated\_\_ ...  
## $ country : chr "CH CH" "CH RU GB RU CN" "CH" "GB CH" ...  
## $ year\_concept : chr "2007+https://openalex.org/C44870925" "2018+https://openalex.org/C44870925" "2003+https://openalex.org/C44870925" "2010+https://openalex.org/C44870925" ...  
## $ concatenated\_title\_abstract : chr "Cosmic magnetic fields and the CMB Abstract I describe the imprint which primordial magnetic fields can leave o"| \_\_truncated\_\_ "Magnetic structure of solar flare regions producing hard X-ray pulsations Abstract We present analysis of the m"| \_\_truncated\_\_ "Dark matter and dark energy: summary and future directions This paper reviews the progress reported at the Disc"| \_\_truncated\_\_ "WASP-32b: A Transiting Hot Jupiter Planet Orbiting a Lithium-Poor, Solar-Type Star ABSTRACT. We report the disc"| \_\_truncated\_\_ ...  
## $ US : num 0 0 0 0 50 0 0 50 0 20 ...  
## $ IN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DE : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CH : num 100 20 100 50 25 50 50 50 50 20 ...  
## $ GB : num 0 20 0 50 25 0 0 0 0 0 ...  
## $ CN : num 0 20 0 0 0 0 0 0 0 0 ...  
## $ FR : num 0 0 0 0 0 0 0 0 0 20 ...  
## $ IT : num 0 0 0 0 0 50 50 0 50 0 ...  
## $ RU : num 0 40 0 0 0 0 0 0 0 0 ...  
## $ CA : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ NL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ JP : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ES : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Americas : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Europe : num 0 0 0 0 0 0 0 0 0 40 ...  
## $ Africa : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AsiaAndOceania : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2020\_2022 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2015\_2019 : int 0 1 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_2010\_2014 : int 0 0 0 1 1 0 1 0 0 1 ...  
## $ pub\_interval\_2005\_2009 : int 1 0 0 0 0 1 0 0 1 0 ...  
## $ pub\_interval\_2000\_2004 : int 0 0 1 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1995\_1999 : int 0 0 0 0 0 0 0 1 0 0 ...  
## $ pub\_interval\_1985\_1994 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1975\_1984 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1965\_1974 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1900\_1964 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1824\_1899 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ original\_concatenated\_title\_abstract: chr "Cosmic magnetic fields and the CMB Abstract I describe the imprint which primordial magnetic fields can leave o"| \_\_truncated\_\_ "Magnetic structure of solar flare regions producing hard X-ray pulsations Abstract We present analysis of the m"| \_\_truncated\_\_ "Dark matter and dark energy: summary and future directions This paper reviews the progress reported at the Disc"| \_\_truncated\_\_ "WASP-32b: A Transiting Hot Jupiter Planet Orbiting a Lithium-Poor, Solar-Type Star ABSTRACT. We report the disc"| \_\_truncated\_\_ ...

# Initialize an empty formula string  
prevalence\_formula\_str <- "~"  
  
# Define the publication intervals  
pub\_intervals <- c("pub\_interval\_2020\_2022", "pub\_interval\_2015\_2019", "pub\_interval\_2010\_2014",   
 "pub\_interval\_2005\_2009", "pub\_interval\_2000\_2004", "pub\_interval\_1995\_1999",  
 "pub\_interval\_1985\_1994", "pub\_interval\_1975\_1984", "pub\_interval\_1965\_1974",  
 "pub\_interval\_1900\_1964", "pub\_interval\_1824\_1899")  
  
# Add each publication interval to the formula string  
for (interval in pub\_intervals) {  
 # add an if statement to handle the first addition (without '+')  
 if (prevalence\_formula\_str == "~") {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, interval)  
 } else {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, "+", interval)  
 }  
}  
  
  
# Convert the string to a formula  
prevalence\_formula <- as.formula(prevalence\_formula\_str)  
print(prevalence\_formula)

## ~pub\_interval\_2020\_2022 + pub\_interval\_2015\_2019 + pub\_interval\_2010\_2014 +   
## pub\_interval\_2005\_2009 + pub\_interval\_2000\_2004 + pub\_interval\_1995\_1999 +   
## pub\_interval\_1985\_1994 + pub\_interval\_1975\_1984 + pub\_interval\_1965\_1974 +   
## pub\_interval\_1900\_1964 + pub\_interval\_1824\_1899

Kvals <- seq(from = 5, to = 40, by = 5) # K values to try  
search\_results <- searchK(documents = out\_text$documents,   
 vocab = out\_text$vocab,   
 K = Kvals,   
 prevalence = prevalence\_formula,   
 data = out\_text$meta,   
 init.type = "Spectral",   
 verbose = FALSE)  
  
# Plot the results  
plot(search\_results)



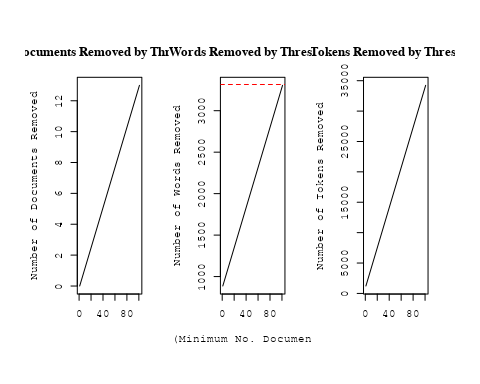
##Topic generation for Africa (in collaboration) publications  
  
data\_collab <- data[data[["Africa"]] != 0,]  
  
# Save the original title data for future use  
data\_collab$original\_concatenated\_title\_abstract <- data\_collab$concatenated\_title\_abstract  
  
#pre-processing the titles using textProcessor from the stm package  
processed\_text <- textProcessor(data\_collab$concatenated\_title\_abstract, metadata = data\_collab)

## Building corpus...   
## Converting to Lower Case...   
## Removing punctuation...   
## Removing stopwords...   
## Removing numbers...   
## Stemming...   
## Creating Output...

# Further prepare the data by removing low-frequency terms  
out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta)

## Removing 886 of 3316 terms (886 of 24261 tokens) due to frequency   
## Your corpus now has 340 documents, 2430 terms and 23375 tokens.

docs\_text <- out\_text$documents  
vocab\_text <- out\_text$vocab  
meta\_text <- out\_text$meta  
  
  
#Prepare data  
plotRemoved(processed\_text$documents, lower.thresh = seq(1, 200, by = 100))



out\_text <- prepDocuments(processed\_text$documents, processed\_text$vocab, processed\_text$meta, lower.thresh = 5)

## Removing 2352 of 3316 terms (5220 of 24261 tokens) due to frequency   
## Your corpus now has 340 documents, 964 terms and 19041 tokens.

str(out\_text$meta)

## 'data.frame': 340 obs. of 39 variables:  
## $ concept\_id : chr "https://openalex.org/C44870925" "https://openalex.org/C1276947" "https://openalex.org/C44870925" "https://openalex.org/C44870925" ...  
## $ work\_id : chr "https://openalex.org/W2977830513" "https://openalex.org/W2027124269" "https://openalex.org/W4214743726" "https://openalex.org/W2481259536" ...  
## $ publication\_year : int 2019 2003 2022 2016 2010 1998 1998 2022 2007 2003 ...  
## $ title : chr "Relativistic Beaming Effects and Structural Asymmetries in Highly Asymmetric Double Radio Sources" "The interplay of astronomy and philosophy" "Discovery of pulsating components in eclipsing binary systems through the TESS light curves: the cases of CPD-3"| \_\_truncated\_\_ "Global Gradients for Cosmic-Ray Protons in the Heliosphere During the Solar Minimum of Cycle 23/24" ...  
## $ paperabstract : chr "We have studied the comparative importance of the relativistic beaming model (RBM) and the density variation mo"| \_\_truncated\_\_ "Astronomical & Astrophysical Transactions The Journal of the Eurasian Astronomical Society Publication details,"| \_\_truncated\_\_ "We present the first evidence for the pulsations of primary components of four eclipsing binary systems. The TE"| \_\_truncated\_\_ "Global gradients for cosmic-ray (CR) protons in the heliosphere are computed with a comprehensive modulation mo"| \_\_truncated\_\_ ...  
## $ country : chr "NG NG" "MZ" "TR BW BW TR" "ZA ZA" ...  
## $ year\_concept : chr "2019+https://openalex.org/C44870925" "2003+https://openalex.org/C1276947" "2022+https://openalex.org/C44870925" "2016+https://openalex.org/C44870925" ...  
## $ concatenated\_title\_abstract : chr "Relativistic Beaming Effects and Structural Asymmetries in Highly Asymmetric Double Radio Sources We have studi"| \_\_truncated\_\_ "The interplay of astronomy and philosophy Astronomical & Astrophysical Transactions The Journal of the Eurasian"| \_\_truncated\_\_ "Discovery of pulsating components in eclipsing binary systems through the TESS light curves: the cases of CPD-3"| \_\_truncated\_\_ "Global Gradients for Cosmic-Ray Protons in the Heliosphere During the Solar Minimum of Cycle 23/24 Global gradi"| \_\_truncated\_\_ ...  
## $ US : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ DE : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CH : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ GB : num 0 0 0 0 50 0 0 0 0 0 ...  
## $ CN : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ FR : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IT : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ RU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ CA : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ NL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ AU : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ JP : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ES : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ IL : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Americas : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Europe : num 0 0 50 0 0 0 0 0 0 0 ...  
## $ Africa : num 100 100 50 100 50 100 100 100 50 100 ...  
## $ AsiaAndOceania : num 0 0 0 0 0 0 0 0 50 0 ...  
## $ pub\_interval\_2020\_2022 : int 0 0 1 0 0 0 0 1 0 0 ...  
## $ pub\_interval\_2015\_2019 : int 1 0 0 1 0 0 0 0 0 0 ...  
## $ pub\_interval\_2010\_2014 : int 0 0 0 0 1 0 0 0 0 0 ...  
## $ pub\_interval\_2005\_2009 : int 0 0 0 0 0 0 0 0 1 0 ...  
## $ pub\_interval\_2000\_2004 : int 0 1 0 0 0 0 0 0 0 1 ...  
## $ pub\_interval\_1995\_1999 : int 0 0 0 0 0 1 1 0 0 0 ...  
## $ pub\_interval\_1985\_1994 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1975\_1984 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1965\_1974 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1900\_1964 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ pub\_interval\_1824\_1899 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ original\_concatenated\_title\_abstract: chr "Relativistic Beaming Effects and Structural Asymmetries in Highly Asymmetric Double Radio Sources We have studi"| \_\_truncated\_\_ "The interplay of astronomy and philosophy Astronomical & Astrophysical Transactions The Journal of the Eurasian"| \_\_truncated\_\_ "Discovery of pulsating components in eclipsing binary systems through the TESS light curves: the cases of CPD-3"| \_\_truncated\_\_ "Global Gradients for Cosmic-Ray Protons in the Heliosphere During the Solar Minimum of Cycle 23/24 Global gradi"| \_\_truncated\_\_ ...

# Initialize an empty formula string  
prevalence\_formula\_str <- "~"  
  
# Define the publication intervals  
pub\_intervals <- c("pub\_interval\_2020\_2022", "pub\_interval\_2015\_2019", "pub\_interval\_2010\_2014",   
 "pub\_interval\_2005\_2009", "pub\_interval\_2000\_2004", "pub\_interval\_1995\_1999",  
 "pub\_interval\_1985\_1994", "pub\_interval\_1975\_1984", "pub\_interval\_1965\_1974",  
 "pub\_interval\_1900\_1964", "pub\_interval\_1824\_1899")  
  
# Add each publication interval to the formula string  
for (interval in pub\_intervals) {  
 # add an if statement to handle the first addition (without '+')  
 if (prevalence\_formula\_str == "~") {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, interval)  
 } else {  
 prevalence\_formula\_str <- paste(prevalence\_formula\_str, "+", interval)  
 }  
}  
  
  
# Convert the string to a formula  
prevalence\_formula <- as.formula(prevalence\_formula\_str)  
print(prevalence\_formula)

## ~pub\_interval\_2020\_2022 + pub\_interval\_2015\_2019 + pub\_interval\_2010\_2014 +   
## pub\_interval\_2005\_2009 + pub\_interval\_2000\_2004 + pub\_interval\_1995\_1999 +   
## pub\_interval\_1985\_1994 + pub\_interval\_1975\_1984 + pub\_interval\_1965\_1974 +   
## pub\_interval\_1900\_1964 + pub\_interval\_1824\_1899

Kvals <- seq(from = 5, to = 40, by = 5) # K values to try  
search\_results <- searchK(documents = out\_text$documents,   
 vocab = out\_text$vocab,   
 K = Kvals,   
 prevalence = prevalence\_formula,   
 data = out\_text$meta,   
 init.type = "Spectral",   
 verbose = FALSE)  
  
# Plot the results  
plot(search\_results)

