MATH6183 — Data Mining Coursework

Textmining: Analyzing Abstracts

Contents

Introduce

This report aims to conduct text analysis on 4385 article abstracts from Journal of the Operational Research Society, Health Systems, and Simulation Journal from 2000 to 2022. It will be analyzed in the following steps:
Abstract text preprocessing, bag-of-words model and TF-IDF analysis, LDA topic modeling, multiple regression analysis, classification model, association rule analysis, clustering, and PCA dimensionality reduction visualization.

Data preprocessing

Before the analysis begins, the summary data needs to be standardized and preprocessed, including the following steps:

Text cleaning: remove punctuation, special characters, and numbers in the summary.

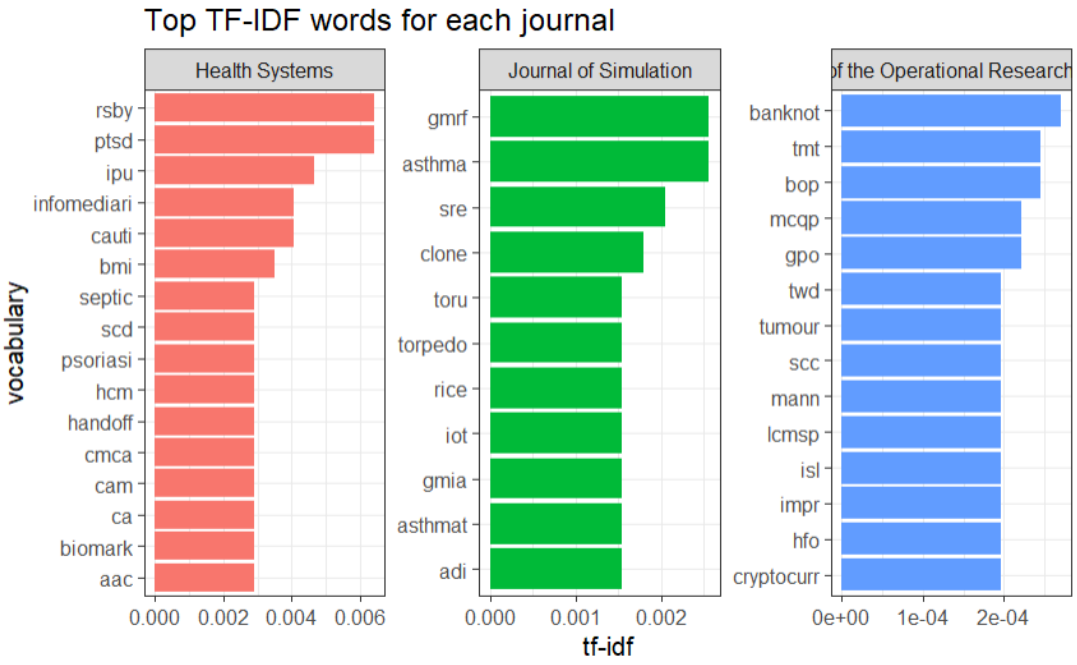
Case conversion: convert all text to lowercase to unify the vocabulary format.

Stop word removal: remove high-frequency stop words (such as "the", "is", "and", etc.) that have no practical significance for semantic analysis.

Stem extraction and lemmatization: simplify words to their root form to reduce the redundancy of different forms of the same word.

Analysis

Bag of Words and TF-IDF



From the figure Top TF-IDF words for each journal, we can see:
The high-frequency words of **Health Systems** are: rsby, ptsd, ipu, infomediari, cauti, etc.
Theme direction: biology, medical and health fields, such as medical policy (RSBY), mental health (PTSD), disease control (cautious), biomarkers (biomark), etc.

The high-frequency words in **Journal of Simulation** are: gmrf, asthma, clone, iot, etc.
Theme direction: simulation and modeling technology in different scenarios, such as research related to the Internet of Things (IOT) and disease modeling (such as asthma).

The high-frequency words in **Journal of the Operational Research Society** are: banknot, tmt, bop, tumour, cryptocurrency, etc.
The subject direction is: different operations and decision optimization problems based on different industries, such as: TMT (Technology, Media, and Telecommunications), banknote (Finance), BOP (Finance), MCQP (OR), GPO (OR), TWD (Taiwan Dollar), etc.


The most important words of each year
2000
relearn
rct
tsh
sdss
ploss
mentor
deadlock
cmb
virus
smp
reel
mpidp
mch
gspn
fork
0.0e+00 5.0e-06 1.0e-05 1.5e-05 2.0e-05
2001
tbc
pgi
lcc
atc
ruin
plotter
vod
stockroom
pgp
oversatur
multicriterion
mot
extran
bvar
builder
am
aem
0.0e+00 5.0e-06 1.0e-05 1.5e-05 2.0e-05
2002
kosovo
vrpsdp
vrpbm
tukei
securit
qdea
nurseri
mdd
lpp
ink
ganjavi
crt
aoi
0e+00 1e-05 2e-05
2003
isl
ksn
ceo
sack
premiership
pcuflp
lotteri
igp
foul
dcd
bvp
bp
0e+00 1e-05 2e-05 3e-05
2004
compass
cbrn
bundeswehr
nimrod
acii
wtp
mod
cro
pclp
miap
humanist
expeditionari
0e+00 1e-05 2e-05 3e-05
2005
mann
yoke
poka
sstsp
rdc
jail
hdi
flp
whitnei
refractori
pop
omf
inmat
faco
cor
artist
0e+00 1e-05 2e-05 3e-05
2006
jo
dotnetsim
dilig
biscai
wwii
nlpc
seminar
wt
sentinel
gpm
0e+00 2e-05 4e-05 6e-05
2007
sre
congruenti
bre
bna
wnt
wet
underwai
tout
ith
gillespi
fulsim
enunci
biologist
0e+00 1e-04 2e-04
2008
spsm
sbatch
amus
spacer
gsp
wassp
titan
worship
underus
trifl
thrill
talk
pegasu
methodolatri
jehovah
iv
interbatch
generalist
ferranti
expressli
evil
connection
andinsight
aero
0.00000 0.00005 0.00010 0.00015
2009
ergonom
carpool
cfe
uml
osaka
heater
tsunami
workplac
undifferenti
undevelop
sfax
receipt
provok
pier
landfal
harbour
habib
cum
cornish
bourguiba
0.00000 0.00005 0.00010 0.00015
2010
thrombophilia
mcl
vte
virtualep
mrsa
colonis
warfarin
avo
nitrogen
ncchta
lln
haemorrhag
ept
eggplant
eea
0.00000 0.00005 0.00010 0.00015
2011
coi
tsont
cdm
referenti
mascom
foup
fab
werent
unequivoc
sysml
subacut
resuscit
presid
omg
inbc
flavour
dorm
antivir
airwai
0.00000 0.00005 0.00010 0.00015

After summarizing the most frequent words from 2002 to 2022, all words are divided into 7 categories: 1. Science and Medicine, 2. Technology and Programming, 3. Geography and International Affairs, 4. Language and Culture, 5. Economics and Business, 6. Military and Government, 7. Society and Education.

Science and Medicine: thrombophilia, warfarin, haemorrhage, nitrogen, ncctha, psoriatic, biomark, mcd, asthma, caregiver, mammograph, virtualme, VTE (Venous Thromboembolism), hyperox, medic, DDM (Diabetes Disease Management), decicardem, antimalar, unambiguous, pseudomembranous, mastoc

Computer Technology: dotnetism, SRE (Site Reliability Engineering), multiclearner, builder, atic, PGC (Procedural Generated Content), AEM (Adobe Experience Manager), vrvsp (Virtual Reality Video Signal Processing), vrpdm (Virtual Reality Product Data Management), SMP (Symmetric Multiprocessing), sentinel, gpm, ipu, platform, CMS, TCM, TOC, electrogram, anonymize, dashboard, netlog, sarima, IPsec.

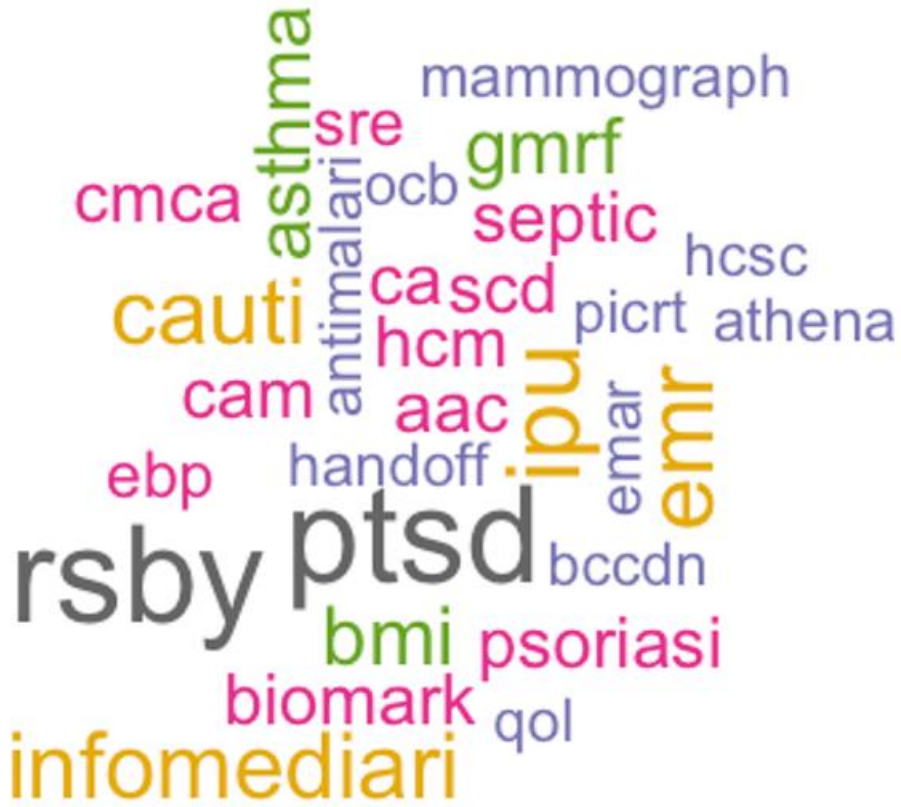
Geography and International Affairs: Kosovo, Bundeswehr, compass, nimrod, Whitehall, CRO (Clinical Research Organization), Toronto, Georgia, Tanzania.

Learning and Food: relearn, plots, masoor (Lentil), imam, biscuit, Robby, ennui, fulsome, gillespi, congruente:

Business: BVP (Business Value Proposition), stock over, port lip, info media, extra, underwriter.
Military and Government: nimrod, sack, premiership, Bundeswehr, tacco, expeditionary, sentinel, bureau, underdog

Society and Education: seminar, teacher, caretaker

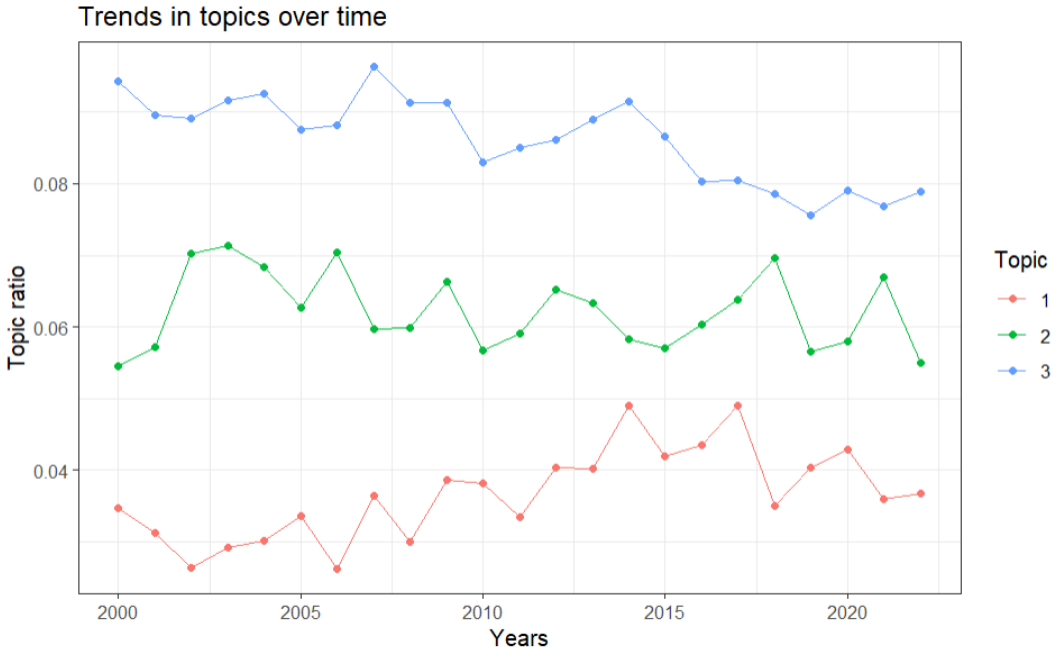
The 30 most frequent words



Topic modelling(LDA)

LDA

	topic	terms
1	1	simul, patient, health, care, system, servic, data, provid, hosp...
2	2	effici, propos, cost, optim, decis, set, solut, time, algorithm, ...
3	3	model, base, paper, system, develop, process, studi, approac...



LDA (latent Dirichlet allocation) was used to find three main topics in the data:

Topic division:

Topic 1: simul, patient, health, care, system, servic, data, provide, hospit, improv

Topic 2: efficiency, propos, cost, optim, decis, set, solut, time, algorithm, method

Topic 3: model, base, paper, system, develop, process, studi, approach, time, result

Time trend:

The proportion of Topic 1 shows a fluctuating trend, and the overall trend is an increasing trend, indicating that the attention to medical and health-related fields continues to increase.

The proportion of Topic 2 shows a fluctuating trend, and the overall trend remains unchanged, indicating that the overall research proportion of topics related to OR research is relatively stable.

Topic 3 has the highest proportion, but the overall proportion remains unchanged. It shows that technology development and system research have the largest research proportion among all topics.

Multiple Regression Model to Predict Number of Citations

```
The specific form of the regression model:
Citations = 5.366 + -0.054 * societi + 0.000 * repli + 0.000 * job + 0.000 * research +
0.000 * price + 0.000 * oper + 0.000 * risk + 0.000 * comment + 0.000 * dea + 0.000 *
commentari + 0.000 * project + 0.000 * rout + 0.000 * product + 0.000 * patient + 0.000
* machin + 0.000 * knowledg + 0.000 * schedul + 0.000 * dmu + 0.000 * supplier + 0.000
* facil + 0.000 * vehicl + 0.000 * simul + 0.000 * balkhi + 0.000 * servic + 0.000 *
agent + 0.000 * retail + 0.000 * heurist + 0.000 * respons + 0.000 * forecast + 0.000 *
attack + 0.000 * health + 0.000 * mainten + 0.000 * algorithm + 0.000 * network + 0.000
* inventori + 0.000 * effici + 0.000 * credit + 0.000 * dye + 0.108 * educ + 0.000 *
psm + 0.000 * bank + 0.000 * firm + 0.000 * ganjavi + 0.000 * rank + 0.000 * game +
0.000 * custom + 0.000 * ship + 0.000 * contract + 0.000 * fuzzi + 0.000 * care + 0.000
* chain + 0.000 * advertis + 0.000 * polici + 0.000 * repair + 0.000 * system + 0.000 *
bonnei + 0.000 * jaber + 0.000 * portfolio + 0.000 * hub + 0.000 * professor + 0.000 *
school + 0.000 * weight + 0.000 * demand + 0.000 * sleeper + 0.000 * suppli + 0.000 *
sampl + 0.000 * cost + 0.000 * contain + 0.000 * bound + 0.000 * search + 0.000 * learn
+ 0.000 * stage + 0.000 * prefer + 0.000 * ganeshan + 0.000 * sarker + 0.000 * ogryczak
+ 0.000 * strategi + 0.000 * hospit + 0.000 * plan + 0.000 * channel + 0.000 * organ +
0.000 * market + 0.000 * manag + 0.000 * locat + 0.000 * item + 0.000 * scorecard +
0.000 * score + 0.000 * erratum + 0.000 * rt + 0.000 * invest + 0.000 * time + 0.000 *
resourc + 0.000 * output + 0.000 * centuri + 0.000 * ssm + 0.000 * method + 0.000 *
select + 0.000 * nurs + 0.000 * target + 0.000 * diet + 0.000 * year + 0.000 * pages +
2.427 * views
```

Multiple Regression Model:

$$\text{Citations} = 5.366 - 0.054 \cdot \text{societi} + 0.108 \cdot \text{educ} + 2.427 \cdot \text{views}$$

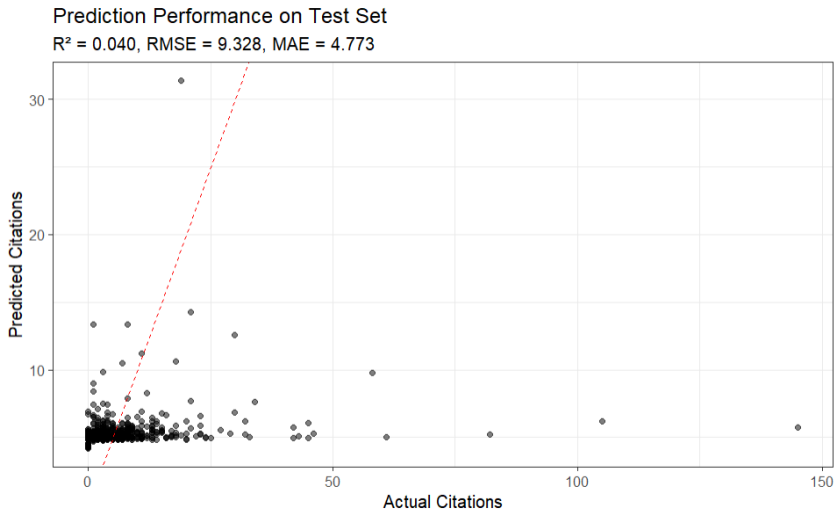
Explanation of the multivariate linear regression model:

societi: The coefficient of this variable is -0.054, which means that for every increase of one unit of "societi", the predicted number of citations decreases by 0.054.

educ: The coefficient of this variable is 0.108, which means that for every increase of one unit of "educ", the predicted number of citations increases by 0.108.

views: The coefficient of this variable is 2.427, which means that for every increase of one unit of "views", the predicted number of citations increases by 2.427.

It can be seen that the variable views has the greatest impact on the number of citations.



From the regression scatter graph: a large number of points are concentrated in the lower left corner, but there are fewer points distributed on the diagonal, indicating that the model

has a large error when predicting samples with high citation times. R square = 0.040 indicates a poor fit, RMSE = 9.328 indicates a large error between the predicted value and the actual value, and MAE of 4.773 also indicates a large error.

Classification and Association

Association Rules

The following figure shows the relevant rules. Each rule consists of LHS and RHS, reflecting the relationship between words that appear together in the summary.

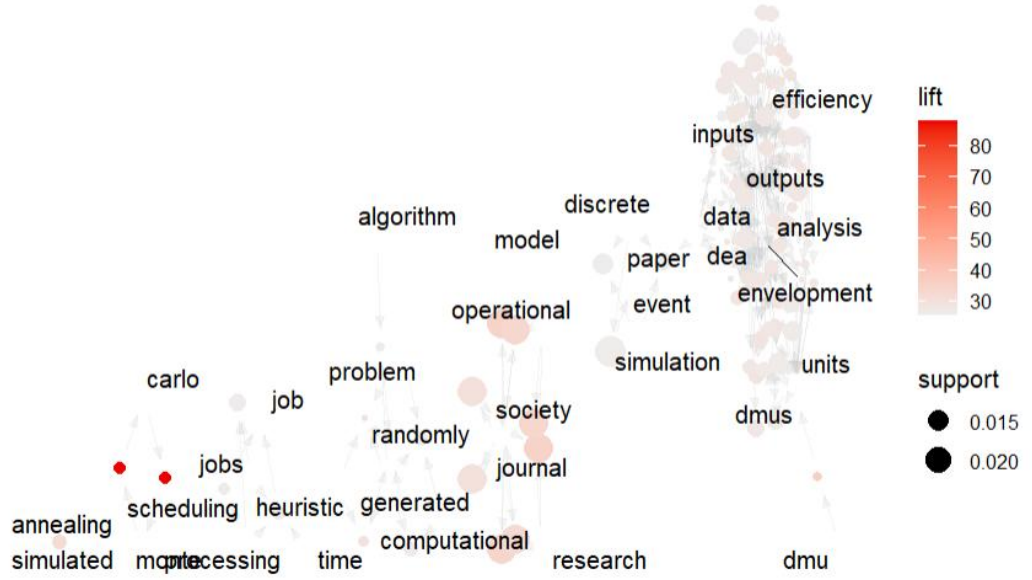
For example: {monte} => {carlo}:

Support: 0.0111, indicating that "monte" and "carlo" appear in 1.11% of transactions at the same time.

Confidence: 1.0000, indicating that whenever "monte" appears, "carlo" appears 100%, indicating that this is an inevitable association.

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>
[1]	{monte}	=>	{carlo}	0.01110843	1.0000000
[2]	{carlo}	=>	{monte}	0.01110843	0.9787234
[3]	{dmu}	=>	{dmus}	0.01038397	0.8113208
[4]	{society, operational}	=>	{journal}	0.02318281	0.9600000
[5]	{research, society, operational}	=>	{journal}	0.02318281	0.9600000
[6]	{research, society}	=>	{journal}	0.02318281	0.9320388
[7]	{research, journal, operational}	=>	{society}	0.02318281	0.9504950
[8]	{journal, operational}	=>	{society}	0.02318281	0.9411765
[9]	{research, journal}	=>	{society}	0.02318281	0.9056604
[10]	{annealing}	=>	{simulated}	0.01183289	0.9607843

Association rule graph



Each point in the figure represents a term, and the size reflects the support of the term. (The higher the support, the larger the node) The edge connects the antecedent (lhs) and the consequent (rhs) of the association rule to indicate the correlation, and the color depth indicates the lift of the rule.

Classification

Confusion Matrix

```
Confusion Matrix and Statistics

                                          Reference
Prediction                                 Health Systems Journal of Simulation
  Health Systems                                       19                     3
  Journal of Simulation                                 1                    34
  Journal of the Operational Research Society          21                    43
                                          Reference
Prediction                                 Journal of the Operational Research Society
  Health Systems                                                                    10
  Journal of Simulation                                                             24
  Journal of the Operational Research Society                                      671
```

The rows of the confusion matrix represent the prediction results of the model, and the columns represent the actual categories.

```
Overall Statistics

               Accuracy : 0.8765
                 95% CI : (0.8521, 0.8982)
    No Information Rate : 0.8535
    P-Value [Acc > NIR] : 0.03218

                  Kappa : 0.467

 Mcnemar's Test P-Value : 0.01625
```

The overall accuracy of the model is 87.65%. The true classification accuracy is likely to be between 85.21% and 89.82%. The P value indicates that the model performs significantly better than random guessing. The Kappa value measures the quality of the classification results. Kappa=0.467 indicates that the classification results are moderate. The Mcnemar's Test P-Value is less than 0.05, indicating that the classification model has systematic bias.

```
Statistics by Class:

                     Class: Health Systems Class: Journal of Simulation
Sensitivity                        0.46341                       0.42500
Specificity                        0.98344                       0.96649
Pos Pred Value                     0.59375                       0.57627
Neg Pred Value                     0.97229                       0.94003
Prevalence                         0.04964                       0.09685
Detection Rate                     0.02300                       0.04116
Detection Prevalence               0.03874                       0.07143
Balanced Accuracy                  0.72343                       0.69574
                     Class: Journal of the Operational Research Society
Sensitivity                                                       0.9518
Specificity                                                       0.4711
Pos Pred Value                                                    0.9129
Neg Pred Value                                                    0.6264
Prevalence                                                        0.8535
Detection Rate                                                    0.8123
Detection Prevalence                                              0.8898
Balanced Accuracy                                                 0.7114
```

Sensitivity: The proportion of samples that are actually in this category that are correctly predicted to be in this category: TP/(TP+FN)
Specificity: The proportion of samples that are actually not in this category that are correctly predicted to be in other categories TN/(TN+FP)

Pos Pred Value: The proportion of samples that are actually in this category among all samples predicted to be in this category TP/(TP+FP)
Neg Pred Value: The proportion of samples that are actually not in this category among all samples predicted to be not in this category TN/(TN+FN)
Prevalence: The proportion of samples that actually belong to this category to the total samples (TP+FN)/Total Sample
Detection Rate: The proportion of samples that are correctly predicted to be in this category to the total samples TP/Total Sample
Detection Prevalence: The proportion of samples that are predicted to be in this category to the total samples (TP+FP)/Total Sample
Balanced Accuracy: a half of Sensitivity+Specificity.

Health Systems
Better performance:
1. Specificity (0.98344): The specificity is high, indicating that the model performs well in excluding non-"Health Systems" samples.
2. Neg Pred Value (0.97229): The negative prediction value is high, indicating that the results predicted as non-"Health Systems" are very reliable

Poor performance:
Detection Rate (0.02300): The detection rate is very low, indicating that the proportion of samples predicted as "Health Systems" is small.

Journal of Simulation
Better performance: Neg Pred Value (0.94003): The negative prediction value is high, indicating that the results predicted as non-this category are more reliable.

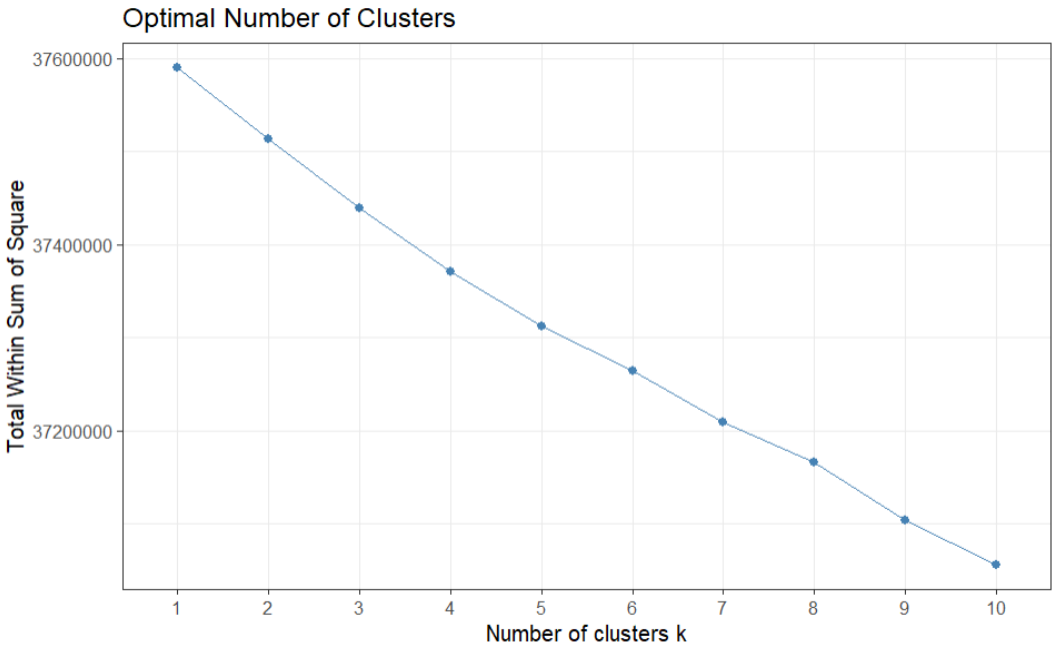
Poor performance: Detection Rate (0.04116): The detection rate is low, indicating that the proportion of samples identified as this category is small.

Journal of the Operational Research Society
Better performance: Sensitivity (0.9518): The sensitivity is very high, indicating that the model works well in identifying samples of this category.

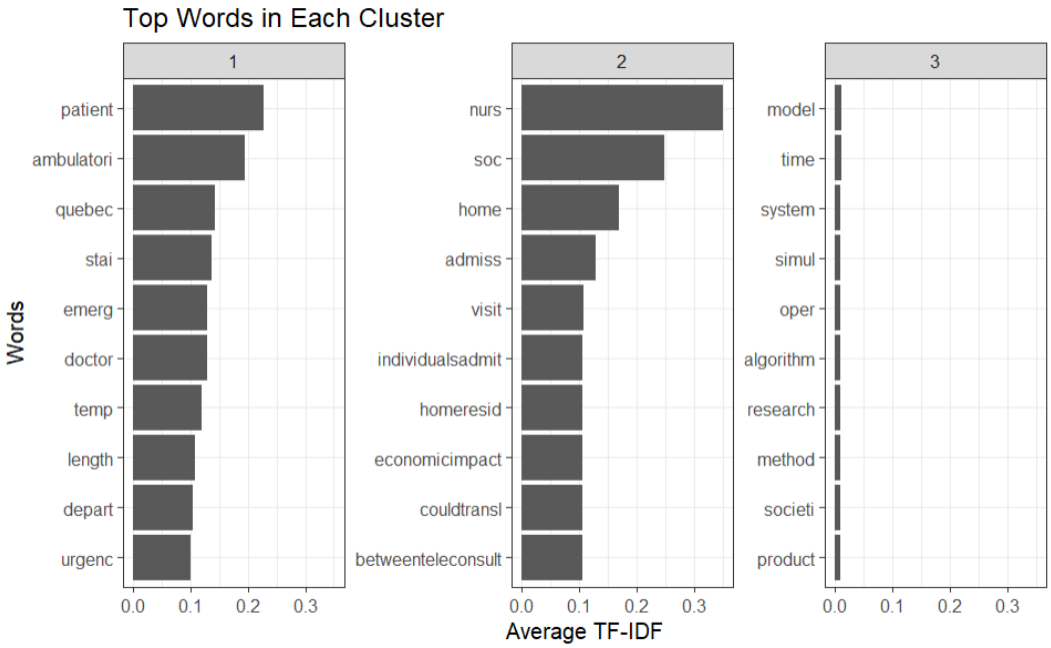
Poor performance: Specificity (0.4711): The specificity is low, indicating that the model performs averagely in excluding samples that are not of this category.

Clustering

First, the elbow rule is used as the initial judgment basis to determine the initial k. The elbow rule determines the appropriate k by assuming that as the number of clusters k increases, TWSS usually decreases, but the decrease gradually slows down. However, as can be seen from the figure, TWSS shows a univariate linear decrease trend as the number of clusters increases. Therefore, this study uses the number of magazines as 3 and takes the value of K as 3 for cluster analysis.



Through the conclusion above , all summaries can be divided into three main categories (Cluster 1, 2, 3), and the words corresponding to the top 10 Average TF-IDF score stems in each category are analyzed. The following is a summary of the characteristics of each category:



Cluster1: patient, ambulatory, Quebec, STAI(State-Trait Anxiety Inventory), emerge, doctor, temp, length, depart, urgency

This part is most likely from Health Systems. Words such as patient, ambulatory, STAI, emergency, doctor, etc. can be related to topics such as hospital health.

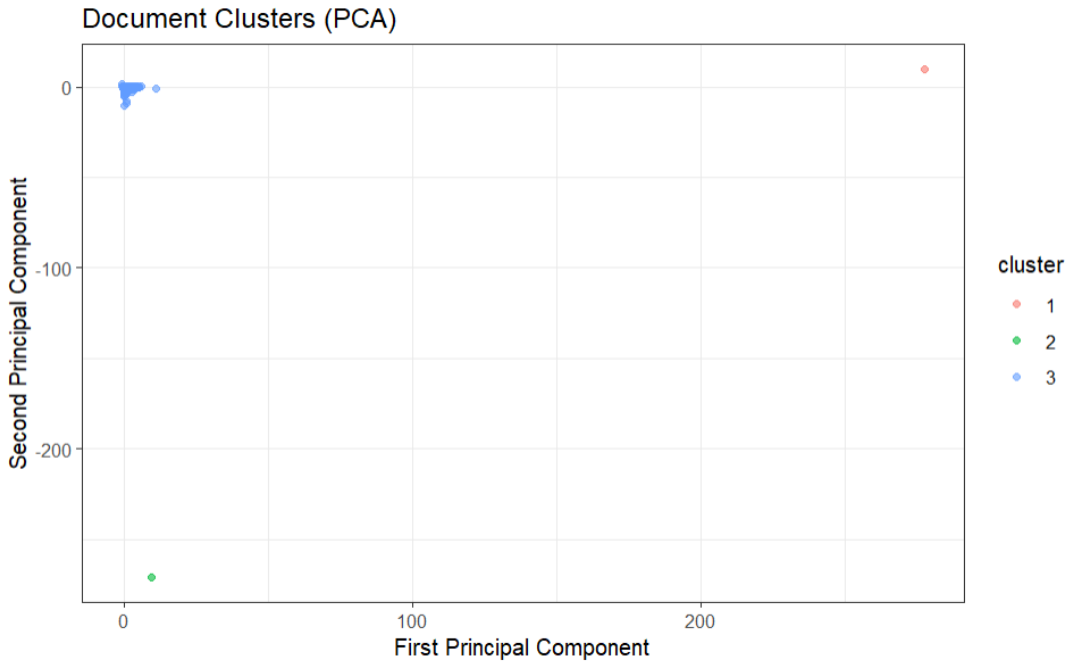
Cluster2: nurse, social, home, admission, visit, individuals admitted, home residents, economic impact, could translate, between teleconsultations

This section is similar to the first section and still covers hospital and medical related topics such as: nurse, admission, individuals admitted, and between teleconsultations.

Cluster3: model, time, system, simulation, operation, algorithm, research, method, society, product

This section is significantly different from the medical topics and is likely to come from the Journal of Simulation or the Journal of the Operational Research Society.

PCA




```
1    2    3
1    1 4138
```

This section uses principal component analysis (PCA) to reduce high-dimensional data to two-dimensional space, and uses clustering results to classify data points.

The horizontal axis is the first principal component (First Principal Component), and the vertical axis is the second principal component (Second Principal Component). As can be seen from the figure, the data points are divided into three clusters, represented by red, green, and blue respectively.

Number of samples: Cluster 1: 1 sample. Cluster 2: 1 sample. Cluster 3: 4138 samples. In this sample, most data points are divided into Cluster 3, resulting in an extremely unbalanced distribution of categories. This imbalance may indicate that the model fails to effectively distinguish other categories.