CAPITOLO 3: Analisi dei contenuti di un dataset di articoli di giornale riguardanti l'IA ed il suo impatto sul mondo del Lavoro

Analisi del contenuto di un gruppo di 70 articoli trattanti il tema dell'IA e del suo impatto sul mondo del lavoro. Tutti gli articoli considerati sono in lingua inglese e provengono da autorevoli testate giornalistiche. Le informazioni su ciascuno degli articoli sono poste all'interno di un file Excel, che viene subito convertito in un dataframe utilizzabile in R. Segue una breve spiegazione dell'organizzazione delle informazioni all'interno del dataframe, con il nome delle sue variabili ed il relativo significato.

Data: gli anni a cui risalgono gli articoli vanno dal 2019 al 2025.

Giornale: sigla del giornale su cui è comparso l'articolo. Legenda: FT Financial Times; WSJ Wall Street Journal; TGU The Guardian; FBS Forbes; NYT New York Times; ECO The Economist; SSIR Stanford Social Innovation Review; AP American Progress.

Titolo: titolo dell'articolo.

Testo: testo dell'articolo.

Il corpus oggetto di analisi è costituito dai testi dei diversi articoli, ognuno dei quali rappresenterà un documento.

3.1 Analisi esplorativa

Nella fase iniziale, dopo aver eseguito le operazioni di preprocessamento, per formare un'idea sulle cartteristiche principali degli articoli in esame, si disegna la wordcloud con i principali termini che compaiono nei testi.

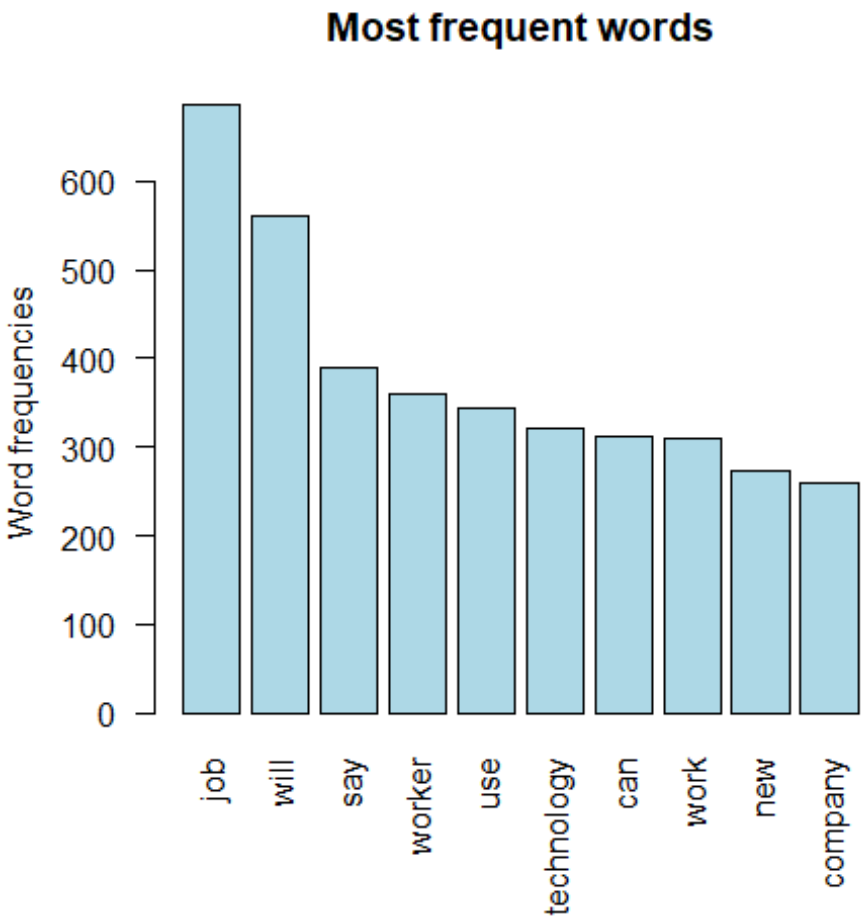
Per tenere conto principalmente dei concetti intrinseci alle parole usate, viene effettuata una lemmatizzazione del corpus.

```
library(tm)

## Caricamento del pacchetto richiesto: NLP

library(wordcloud)

## Caricamento del pacchetto richiesto: RColorBrewer

library(RColorBrewer)
library(readxl)
library(textstem)

## Caricamento del pacchetto richiesto: koRpus.lang.en

## Caricamento del pacchetto richiesto: koRpus

## Caricamento del pacchetto richiesto: sylly
```

```
## For information on available language packages for 'koRpus', run
##
##   available.koRpus.lang()
##
## and see ?install.koRpus.lang()

##
## Caricamento pacchetto: 'koRpus'

## Il seguente oggetto è mascherato da 'package:tm':
##
##     readTagged

gruppo_articoli<-
read_excel("C:/Users/1jaco/OneDrive/Desktop/SMA/Progetto_SMA/Dataset_finale/DatasetEsteso.xlsx")

str(gruppo_articoli)

## tibble [70 × 5] (S3: tbl_df/tbl/data.frame)
##  $ data    : POSIXct[1:70], format: "2019-03-13" "2019-04-01" ...
##  $ giornale: chr [1:70] "FT" "WSJ" "WSJ" "FT" ...
##  $ titolo  : chr [1:70] "AI academics under pressure to do commercial research"
"What AI Will Do to Corporate Hierarchies" "Will AI Destroy More Jobs Than It
Creates Over the Next Decade?" "Rich nations urged to prepare workers for age of
automation" ...
##  $ testo   : chr [1:70] "AI academics under pressure to do commercial
research\r\nTech giants ramp up recruitment of university scientis"| __truncated__
"What AI Will Do to Corporate Hierarchies\r\nThe conventional wisdom says we can
expect a more centralized struc"| __truncated__ "Will AI Destroy More Jobs Than It
Creates Over the Next Decade?\r\nDecide the answer for yourself, as two exper"|
__truncated__ "Rich nations urged to prepare workers for age of
automation\r\nTechnological change driving ‘creative destructi"| __truncated__ ...
##  $ anno    : num [1:70] 2019 2019 2019 2019 2019 ...

gruppo_articoli$anno <- factor(gruppo_articoli$anno)

mycorpus<-Corpus(VectorSource(gruppo_articoli$testo))

mycorpus <- tm_map(mycorpus, content_transformer(tolower))

## Warning in tm_map.SimpleCorpus(mycorpus, content_transformer(tolower)):
## transformation drops documents

mycorpus <- tm_map(mycorpus, removeNumbers)

## Warning in tm_map.SimpleCorpus(mycorpus, removeNumbers): transformation drops
## documents

mycorpus <- tm_map(mycorpus, removePunctuation)

## Warning in tm_map.SimpleCorpus(mycorpus, removePunctuation): transformation
## drops documents
```

```r
remove_custom_punctuation <- function(text) {gsub("[[:punct:]]", " ", text) }
# Rimuove tutti i segni di punteggiatura comprese le virgolette (trasformandoli in
spazi)

# Applicare la funzione al corpus
mycorpus <- tm_map(mycorpus, content_transformer(remove_custom_punctuation))

## Warning in tm_map.SimpleCorpus(mycorpus,
## content_transformer(remove_custom_punctuation)): transformation drops documents

mycorpus <- tm_map(mycorpus, stripWhitespace)

## Warning in tm_map.SimpleCorpus(mycorpus, stripWhitespace): transformation drops
## documents

mycorpus<-tm_map(mycorpus, removeWords, stopwords("en"))

## Warning in tm_map.SimpleCorpus(mycorpus, removeWords, stopwords("en")):
## transformation drops documents

#rimozione delle parole "tema" dal corpus
theme_words <- c( "ai", "artificial", "intelligence")

mycorpus <- tm_map(mycorpus, removeWords, theme_words)

mycorpus_lem <- tm_map(mycorpus, lemmatize_strings)

#ottenimento della term-document matrix
tdm<-TermDocumentMatrix(mycorpus_lem)

#Sparsity 94%

# transformazione tdm in un oggetto matrix
Matrixtd <- as.matrix(tdm)

#frequenze assolute di ogni termine all'interno del corpus
v <- sort(rowSums(Matrixtd),decreasing=TRUE)

#aggiunta di una colonna con i nomi delle parole
d <- data.frame(word = names(v),freq=v)

library(wordcloud)
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words=200,random.order=FALSE,rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```


```
barplot(d[1:10,]$freq, las = 2, names.arg = d[1:10,]$word,
                col ="lightblue", main ="Most frequent words",
                 ylab = "Word frequencies")
```

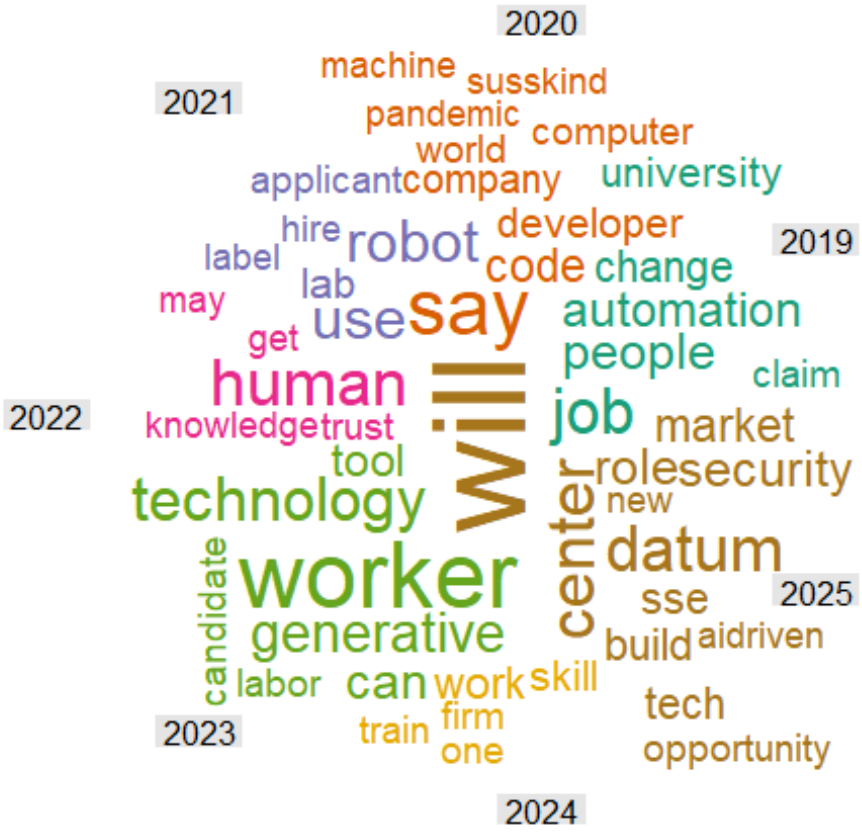


Le precedento rappresentazioni riflettono i temi di cui ci si aspettava si parlasse: gli impieghi, i lavoratori e la tecnologia. Degno di nota è il fatto che uno dei termini complessivamente più utilizzati sia “will”, ciò fa capire come il contesto temporale in cui vengono collocati i temi negli articoli sia il futuro, ciò deriva dal carattere innovativo dell’IA. La presenza dei termini “work” e “worker” testimoniano come, oltre ad avere un impatto sugli impieghi “job”, l’intelligenza artificiale influenza il modo in cui questi vengono svolti, ossia il lavoro e coloro che lo svolgono.

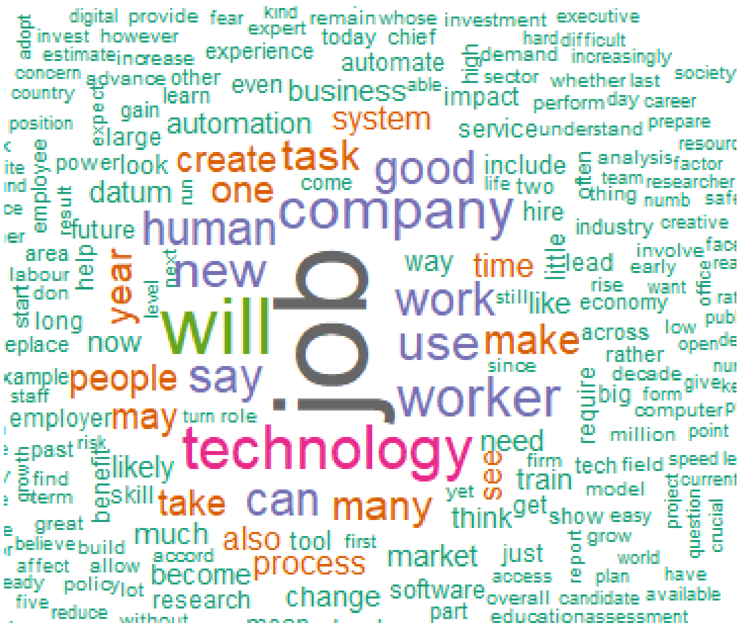
Si vogliono ora valutare le eventuali differenze nei vocaboli usati tra i diversi anni a cui gli articoli oggetto di analisi risalgono e, dunque, nei temi trattati dagli articoli a seconda dei diversi anni in cui sono stati scritti. Oltre a ciò, per contrasto, si vogliono far emergere i termini che sono ricorrenti in tutti gli anni considerati.

```
#costruzione di una tabella lessicale contenente i termini presenti nel corpus
nelle righe e gli anni in cui sono stati scritti gli articoli nelle colonne.
#Ogni elemento della tabella rappresenta la frequenza del termine riga all'interno
di articoli ascrivibili all'anno corrispondente alla colonna.

dim(Matrixtd)

## [1] 5751   70
```

```
row_sums <- rowSums(Matrixtd)

# Ordina la matrice in base alla somma delle righe
Matrixtd <- Matrixtd[order(row_sums, decreasing = TRUE), ]


TL<-matrix(nrow=5751,ncol=7) #inizialemnte è una matrice vuota, con tante righe quanti sono i vocaboli diversi e tante colonne quanti sono gli anni in cui gli articoli sono stati scritti.
for (j in 1:5751)
{TL[j,]<-tapply(Matrixtd[j,],gruppo_articoli$anno, sum)
TL<-rbind(TL)}
colnames(TL)<-c("2019","2020", "2021", "2022", "2023", "2024", "2025")
row.names(TL)<-row.names(Matrixtd)


comparison.cloud(TL, max.words = 50,
                 title.size = 1)
```




```
commonality.cloud(TL,  random.order=FALSE,
                  colors = brewer.pal(8, "Dark2"),
                  title.size=1.5)
```



Dalla comparison cloud si nota l'evoluzione dell'opinione riguardo al campo di influenza dell'Intelligenza artificiale sul lavoro. In particolare all'inizio, nel 2019, i compiti che potevano essere assegnati alla tecnologia non erano ancora chiari, ma il termine "automation" suggerisce che già era presente l'idea di utilizzare la tecnologia per automatizzare alcuni compiti; tra le dimensioni di sviluppo di questa tecnologia vi era quella accademica, come testimonia la presenza del termine "university". Nel 2020, viene sottolineato l'utilizzo del termine "robot", il che suggerisce la possibilità dell'utilizzo di automi nel mondo del lavoro. I termini "applicant" ed "hire" indica l'utilizzo dell'IA all'inerno dei processi di assunzione. Nel 2021 il focus passa agli esseri umani "human", probabilmente poichè si tiene conto del rapporto con l'IA; nel contesto di questo rapporto, appare rilevante sottolineare l'emergere del termine "trust", ossia la fiducia che in questa tecnologia riporre. Nel 2023 spicca l'utilizzo del termine "generative" (probabilmente usato accanto ad "ai", che è stata rimossa in quanto theme word), a testimoniare il risalto che questa declinazione dell'IA ha avuto. Inoltre, nel 2023, la presenza dei termini "worker", "technology" e "can" suggerisce le nuove possibilità per i lavoratori date dall'utilizzo della tecnologia. Quest'ultimo tema è collegabile ai termini "skill" e "train", che sono specifici del 2024 e indicano che le possibilità precedenti possono essere sfruttate a condizione che i lavoratori acquisicano le capacità richieste per l'utilizzo dei nuovi strumenti disponibili. Infine, nel 2025 emerge la centralità del dato ("datum"), considerabile come

elemento fondante dell'IA; oltre a ciò, da sottolineare la rilevanza del tema della sicurezza "security", ""sse" ossia Security Server Edge.

Dalla commonality cloud emergono i termini caratterizzanti i temi dell'indagine in generale. In particolare la presenza del termine "will" testimonia come nella maggior parte dei casi gli autori degli articoli effettuano previsioni sull'utilizzo futuro dell'IA, essendo questo un tema in continua evoluzione e dai confini non completamente delineabili, neanche negli anni più recenti.

3.2 Studio delle associazioni tra termini.

Studio delle associazioni per i termini maggiormente rilevanti per il corpus, così da comprenderne le relazioni semantiche.

Ci si concentra sui termini più rilevanti all'interno del corpus in analisi decidendo di diminuire il livello di "sparsity" della term document matrix eliminando i termini che occorrono complessivamente meno di 5 volte.

```
freq_words <- findFreqTerms(tdm,5)
tdm_freqwords <- tdm[freq_words,]

associations <- findAssocs(tdm_freqwords,terms= c("worker","job", "technology",
"time", "datum"),corlimit= 0.50) #impostazione della soglia minima delle
correlazioni riportate a 0.5

worker_associations<-as.data.frame(associations$worker)

print(worker_associations)
```

```
##                associations$worker
## retrain                       0.73
## include                       0.69
## displace                      0.67
## emphasis                      0.66
## congress                      0.64
## biden                         0.63
## net                           0.61
## policy                        0.60
## steer                         0.60
## technological                 0.58
## policymaker                   0.57
## complement                    0.56
## administration                0.56
## adopt                         0.55
## central                       0.55
## likely                        0.54
## safety                        0.54
## degree                        0.54
## apprenticeship                0.54
## unemployment                  0.53
## occupation                    0.52
```

```
## must                          0.52
## assistance                    0.52
## goldman                       0.52
## pay                           0.51
## sachs                         0.51
```

```
job_associations<-as.data.frame(associations$job)
```

```
print(job_associations)
```

```
##          associations$job
## create               0.59
## will                 0.59
## million              0.57
## role                 0.54
## offset               0.54
## few                  0.53
## replace              0.52
## creation             0.51
## lose                 0.51
```

```
technology_associations<-as.data.frame(associations$technology)
```

```
print(technology_associations)
```

```
##            associations$technology
## prosperity                    0.57
## labor                         0.52
## mit                           0.52
## encourage                     0.51
## direction                     0.50
```

```
datum_associations <- as.data.frame(associations$datum)
```

```
print(datum_associations)
```

```
##              associations$datum
## space                      0.65
## stargate                   0.63
## center                     0.62
## build                      0.61
## trump                      0.53
## hundred                    0.53
## plant                      0.53
## boom                       0.52
## construction               0.52
```

Di seguito riassumiamo i risultati dell'applicazione della funzione findAssocs sulla termdocument matrix ottenuta dalla term-document matrix (tdm_lem_freqwords):

Per il termine "worker", le principali associazioni coinvolgono "retrain", ossia la sopramenzionata necessità di riaggiornamento dei lavoratori rispetto all'utilizzo delle nuove

tecnologie; “include”, cioè viene discusso il modo in cui l’area di influenza dell’IA include i lavoratori.Il termine “displace” indica come l’utilizzo dell’IA da parte delle aziende può significare l’automatizzazione di determinate mansioni ed il conseguente licenziamento dei lavoratori che ad esse erano adibiti. Infine, legato all’ultimo tema vi è quello della politica (“policy” e “biden”) che suggerisce come si debbano regolare i rapporti tra lavoratori e IA nelle aziende.

Le associazioni del termine “job” indicano la grande portata dell’impatto dell’IA sugli impieghi, come si vede dalla presenza del termine “milion”. Quest’impatto ha due effetti, tra loro opposti: quello di creare (“create”) nuovi impieghi, caratterizzati dalla capacità di utilizzo di questa nuova tecnologia, e quello di rimpiazzare (“replace”) gli esseri umani in impieghi che possono essere svolti dall’IA.

Connesso alla sostituzione dei lavoratori da parte dell’IA, vi è l’associazione del termine “technology” con “labor”, che potrebbe suggerire proprio l’automatizzazione di alcunne mansioni precedentemente svolte da esseri umani. Infine, da sottolineare l’associazione di “technology” con il termine “prospeity”, che può indicare come l’utilzzo corretto della tecnologia può avere un impatto positivo sul futuro.

Infine, per il termine “datum”, l’associazione con “space” potrebbe riferirsi allo spazio dei dati e al costante aumento della necessità di spazio (in termini di memoria, ma di spazio fisicamente richiesto per ospitare i data center) per immagazzinare quantità crescenti di dati, su cui l’AI si basa. La presenza dei termini “Stargate” e “boom” può essere considerata una metafora fantascientifica legata all’accesso a grandi e crescenti quantità di dati. Il termine “Center” si riferisce ai data centers, ovvero infrastrutture per l’archiviazione e l’elaborazione di dati.

3.3 test di Lafon: valutare specificità termini rispetto agli anni in cui sono scritti gli articoli

Si vuole valutare la specificità di determinati termini rispetto a sottoparti del corpus analizzato. Per fare ciò si utilizzerà la funzione specificities, basata sul test di Lafon. In particolare, le sottoparti del corpus qui considerate sono quelle ottenute dividendo gli articoli per gli anni a cui risalgono.

A questo scopo, si parte dalla tabella lessicale completa, in cui le frequenze vengono divise per i diversi anni, ottenuta dopo la “lemmatization”.

```
library(textometry)
spec<-specificities(TL)



spec["worker",]

cat("worker specificities", "\n", spec["worker",])

## worker specificities

##   2019    2020    2021    2022    2023    2024    2025
##  -0.9009 -0.9285 -0.8836 -0.5546 11.8674 -0.7589 -3.7405
```

```r
spec["say",]

cat("say specificities", "\n", spec["say",])

## say specificities

##  2019   2020   2021   2022    2023   2024    2025
##  0.6567 6.4075 0.2971 0.3417 -2.2385 0.3618 -5.8185

spec["generative",]

cat("generative specificities", "\n", spec["generative",])

## generative specificities

##    2019   2020    2021    2022    2023   2024   2025
##  -4.9029 -6.0083 -6.646 -0.2665 10.2971 5.066 -1.3107

spec["will",]


cat("will specificities", "\n", spec["will",])

## will specificities

##  2019    2020    2021    2022    2023    2024    2025
##  1.6747 -0.2851 1.1558 -5.5462 -2.5282 -3.3797 10.8639
```

Nell'analisi dei risultati forniti dall'ottenimento delle specificità dei singoli termini relativamente agli anni in cui sono stati scritti gli articoli, ci si sofferma in primis sul termine "worker", che risulta altamente specifico per l'anno 2023; questo può essere dato dal fatto che in tale anno diverse funzioni dell'IA erano state sviluppate e ne era stata verificata la efficacia, dunque si era aperto il dibattito sulla possibilità di introdurre questa tecnologia coem supporto o come sostituto dei lavoratori. L'affermazione dell'IA può aver portato, inoltre, al delinearsi di un nuovo tipo di lavoratori.

La specificità del termine "say" è molto alta nel 2020, uno dei primi anni oggetto di analisi, a testimonianza del fatto che inizialmente non si aveva una conoscenza precisa delle possibilità dell'IA, perciò si facevano previsioni e si riportava il parere di esperti.

Al contrario, il termine "generative" presenta una specificità molto alta nel 2024 e nel 2023. Questo suggerisce che vi è stato un più preciso delineamento delle funzioni dell'IA, come la tecnologia generativa, che è stata un tema chiave negli anni recenti nell'ambito del dibattito IA-lavoro.

Per il termine "will", è interessante sottolineare come questo è altamente specifico per l'anno 2025. Questo indica come il tema sia in continua evoluzione e le possibilità di sviluppo dell'IA siano tutt'altro che già completamente espresse.

3.4 Identificazione dei Segmenti di parole maggiormente usati

Ci si avvale della funzione TextData per analizzare i segmenti di parole maggiormente rilevanti all'interno del corpus esaminato, nello scopo di comprendere i contesti in cui i termini vengono utilizzati.

In questo caso, poichè si vogliono analizzare i segmenti, non si specifica una lista di stopwords da rimuovere.

```
#modifica del corpus nel dataframe per poter applicare textdata
gruppo_articoli$testo <-  gsub("[[:punct:]]", " ", gruppo_articoli$testo) #evitare che tra le parole vengano inclusi _ e "

library(Xplortext)

## Caricamento del pacchetto richiesto: FactoMineR

## Caricamento del pacchetto richiesto: ggplot2

##
## Caricamento pacchetto: 'ggplot2'

## Il seguente oggetto è mascherato da 'package:NLP':
##
##     annotate

## Registered S3 method overwritten by 'vegan':
##   method     from
##   rev.hclust dendextend

#minima frequenza del segmento nel corpus per essere incluso: 4.
res.TD_1<-TextData(gruppo_articoli,var.text= 4, idiom="en",
segment=TRUE,seg.nfreq=4, seg.nfreq2=1000, seg.nfreq3=1000)


 print(res.TD_1$indexS$segOrderFreq)

##                                     segment frequency long
## 49                        the future of work        13    4
## 44                        the adoption of ai        12    4
## 25                                in the u s        10    4
## 28                           is likely to be         9    4
## 63                          when it comes to         9    4
## 6                           are likely to be         8    4
## 10                          at the same time         8    4
## 42                      over the next decade         8    4
## 55                  the world economic forum         8    4
## 3                            ai s impact on         7    4
## 8                            as a result of         7    4
## 27                           is going to be         7    4
## 33                         more jobs than it         7    4
## 37                              of ai in the         7    4
## 53                             the use of ai         7    4
## 7                            are using ai to         6    4
```

```
## 14                          by generative a i         6    4
## 19                              in a way that         6    4
## 21                            in the long run         6    4
## 22                           in the number of         6    4
## 41                            one of the most         6    4
## 43                            per cent of the         6    4
## 46                       the company s chief          6    4
## 54                          the use of ai in          5    5
## 1                            a few years ago          5    4
## 2                     ai and machine learning         5    4
## 5                          ai will impact the         5    4
## 12                              be able to do         5    4
## 20                       in the hiring process         5    4
## 23                            in the past year         5    4
## 24                           in the short term         5    4
## 26                              in the us and          5    4
## 30                       likely to be affected         5    4
## 34                           more likely to be         5    4
## 35                           most likely to be         5    4
## 45                 the burning glass institute         5    4
## 50                           the impact of the         5    4
## 56                           the world of work         5    4
## 57                              to be able to          5    4
## 58                           to be affected by         5    4
## 59            to be affected by generative a i         4    7
## 15           destroy more jobs than it creates         4    6
## 31                    likely to be affected by         4    5
## 36                  most likely to be affected         4    5
## 47               the company s chief executive         4    5
## 51 the massachusetts institute of technology          4    5
## 4                         ai tools and systems        4    4
## 9                                      at m i t       4    4
## 11                        at the university of        4    4
## 13                              by as much as         4    4
## 16                  for the foreseeable future        4    4
## 17                       future of jobs report        4    4
## 18                        has the potential to        4    4
## 29                             it will lead to        4    4
## 32                      million full time jobs        4    4
## 38                          of the economy and        4    4
## 39                           on the job market        4    4
## 40                           on the other hand        4    4
## 48                         the creation of new        4    4
## 52                            the number of ai        4    4
## 60              use of artificial intelligence        4    4
## 61                            want to make sure       4    4
## 62                          what this means for       4    4
```

Segmenti di parole maggiormente rilevanti: “the future of work”, “the adoption of ai”. Queste espressioni suggeriscono come il legame tra il fututo del lavoro, inteso in generale, e l’intelligenza artificiale. Il segmento: “destroy more jobs than it creates”, benchè non abbia una

frequenza tra le più rilevanti, fa trasparire la presenza di preoccupazione circa il possibile aumento della disoccupazione a causa dell'utilizzo dell'IA.

3.5 Analisi delle corrispondenze lessicali e clustering gerarchico

Applicazione della funzione TextData sul testo degli articoli, facendo sì che vengano presi in considerazione unicamente i termini maggiormente ricorrenti e aggregando le frequenze di tali termini rispetto agli anni in cui sono stati scritti gli articoli. A partire dall'oggetto ottenuto, applicazione della funzione LexCa per effettuare una correspondance analysis, e, dunque, proiettare i risultati nello spazio fattoriale. Quindi, sulla base delle coordinate di termini e documenti nello spazio fattoriale ottenuto, si applica un algoritmo di clustering gerarchico per individuare raggruppamenti rilevanti.

```
library(textstem)
library(Xplortext)

gruppo_articoli$testo <-  gsub("[[:punct:]]", " ", gruppo_articoli$testo) #evitare che tra le parole vengano inclusi _ e "

#prima di utilizzare la funzione textdata, poichè questa volta l'obiettivo è produrre una mappa semantica, è opportuno effettuare una lemmatizzazione sui testi degli articoli.

gruppo_articoli$testo <- lemmatize_strings(gruppo_articoli$testo)


#funzione utilizzata per rimuovere i termini con meno di due caratteri
remove_short_words <- function(text, min_length = 2) {
  words <- unlist(strsplit(text, "\\s+"))  # Split in parole
  words <- words[nchar(words) >= min_length]  # fa rimanere unicamente parole con lunghezza maggiore di min_length
  paste(words, collapse = " ")  # Recombina in una frase
}

gruppo_articoli$testo <- sapply(gruppo_articoli$testo, remove_short_words)


#nuovo (secondo) utilizzo della funzione TextData, si aggregano i risultati secondo la variabile anno
res.TD_2<-TextData(gruppo_articoli,var.text= 4, idiom="en", var.agg="anno", Fmin=120, Dmin=5,  stop.word.user=theme_words,stop.word.tm=TRUE)

#Fmin = 120: si includono nell'analisi solo i termini che occorrono un minimo di 120 volte nel corpus: questo permette di eliminare il "rumore" generato dai termini meno frequenti e rende i risultati dell'analisi maggiormente leggibili.

#rispetto alla stessa funzione applicata in precedenza, non vengono analizzati i segmenti di parole ripetuti, concentrandosi sui singoli termini e sui documenti.
```

```
res.LexCA<-LexCA(res.TD_2, graph=FALSE)
summary(res.LexCA,metaWords=FALSE)
```

```
## Correspondence analysis summary
##
## Eigenvalues
##       Variance % of var. Cumulative % of var.
## dim 1    0.032    30.940               30.940
## dim 2    0.027    25.867               56.807
## dim 3    0.019    18.164               74.971
## dim 4    0.012    11.277               86.248
## dim 5    0.008     7.371               93.619
##
## Cramer's V  0.132    Inertia  0.104
##
##
## DOCUMENTS
## All documents are aggregate documents
##
##  Coordinates
##       Dim 1  Dim 2  Dim 3  Dim 4  Dim 5
## 2019  0.108  0.150 -0.256 -0.106 -0.009
## 2020  0.086  0.091  0.202 -0.139  0.091
## 2021  0.124  0.076  0.113  0.058 -0.177
## 2022 -0.145  0.205  0.067  0.112  0.086
## 2023 -0.246 -0.191  0.004 -0.091 -0.039
## 2024 -0.111 -0.013 -0.086  0.148  0.032
## 2025  0.329 -0.303 -0.012  0.071  0.074
##
##  Contributions (by column total=100)
##       Dim 1  Dim 2  Dim 3  Dim 4  Dim 5
## 2019  5.451 12.570 51.895 14.289  0.170
## 2020  3.071  4.174 29.206 22.355 14.458
## 2021  7.047  3.146 10.050  4.274 60.627
## 2022  7.721 18.311  2.786 12.460 11.442
## 2023 35.464 25.770  0.016 13.448  3.776
## 2024  5.905  0.091  5.972 28.677  2.040
## 2025 35.342 35.939  0.074  4.497  7.488
##
##  Square cosinus (by row total=1)
##      Dim 1 Dim 2 Dim 3 Dim 4 Dim 5
## 2019 0.105 0.203 0.588 0.101 0.001
## 2020 0.081 0.092 0.451 0.214 0.091
## 2021 0.223 0.083 0.187 0.049 0.457
## 2022 0.197 0.390 0.042 0.116 0.069
## 2023 0.560 0.340 0.000 0.077 0.014
## 2024 0.203 0.003 0.121 0.360 0.017
## 2025 0.504 0.428 0.001 0.023 0.025
##
##      Inertia
```

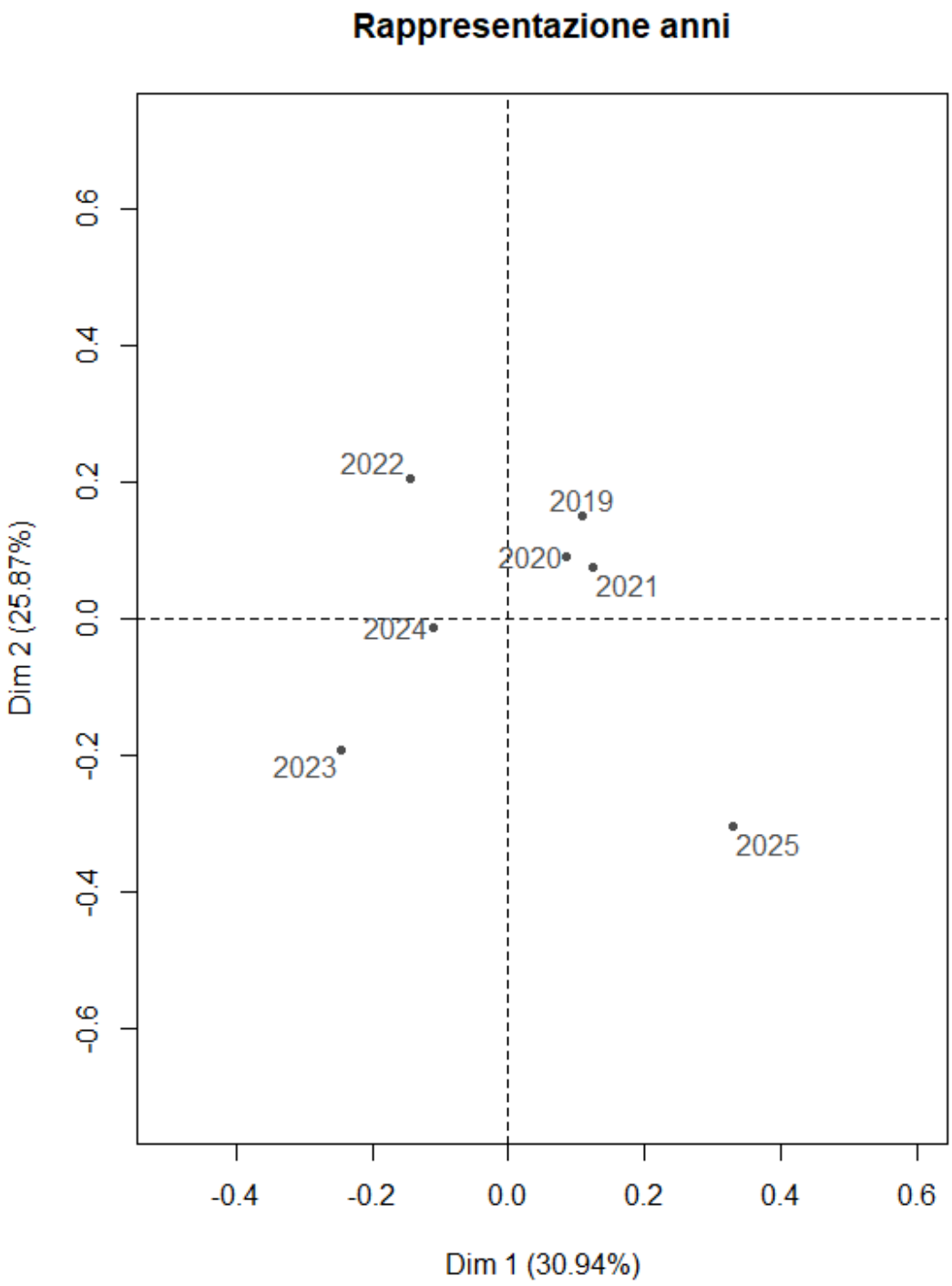
```
## 2019   0.017
## 2020   0.012
## 2021   0.010
## 2022   0.013
## 2023   0.020
## 2024   0.009
## 2025   0.023
##
##
## WORDS
##
##  Coordinates
## Only the first  10 elements are shown
##             Dim 1  Dim 2  Dim 3  Dim 4  Dim 5
## also         0.006 -0.153  0.141  0.026 -0.012
## automation   0.127  0.295 -0.127 -0.169 -0.131
## business     0.068 -0.147  0.188  0.154  0.048
## can         -0.319 -0.090 -0.045  0.060 -0.070
## company      0.083  0.046  0.175  0.029  0.006
## create       0.126 -0.035 -0.160 -0.106  0.090
## datum        0.524 -0.212  0.158  0.159 -0.157
## good        -0.153  0.044  0.090  0.089 -0.020
## human       -0.161  0.208  0.102  0.198  0.147
## job          0.036 -0.068 -0.109 -0.057  0.018
##
##  Contributions (by-column total=100)
## Only the first  10 elements are shown
##             Dim 1 Dim 2 Dim 3  Dim 4 Dim 5
## also         0.002 1.355 1.651  0.087 0.031
## automation   0.937 6.065 1.608  4.579 4.211
## business     0.242 1.328 3.127  3.378 0.508
## can         18.534 1.772 0.621  1.790 3.703
## company      0.688 0.250 5.176  0.226 0.015
## create       0.852 0.077 2.354  1.666 1.813
## datum       20.085 3.908 3.088  5.063 7.543
## good         1.814 0.178 1.061  1.692 0.133
## human        2.509 5.033 1.721 10.525 8.808
## job          0.333 1.413 5.208  2.283 0.369
##
##  Square cosinus (by-row total=1)
## Only the first  10 elements are shown
##             Dim 1 Dim 2 Dim 3 Dim 4 Dim 5
## also        0.001 0.490 0.419 0.014 0.003
## automation 0.082 0.447 0.083 0.147 0.088
## business   0.040 0.184 0.304 0.204 0.020
## can        0.846 0.068 0.017 0.030 0.040
## company    0.116 0.035 0.513 0.014 0.001
## create     0.255 0.019 0.414 0.182 0.129
## datum      0.696 0.113 0.063 0.064 0.062
## good       0.477 0.039 0.164 0.162 0.008
## human      0.167 0.280 0.067 0.255 0.140
```

```
## job         0.061 0.215 0.557 0.151 0.016
##
##             Inertia
## also          0.001
## automation    0.004
## business      0.002
## can           0.007
## company       0.002
## create        0.001
## datum         0.009
## good          0.001
## human         0.005
## job           0.002
```

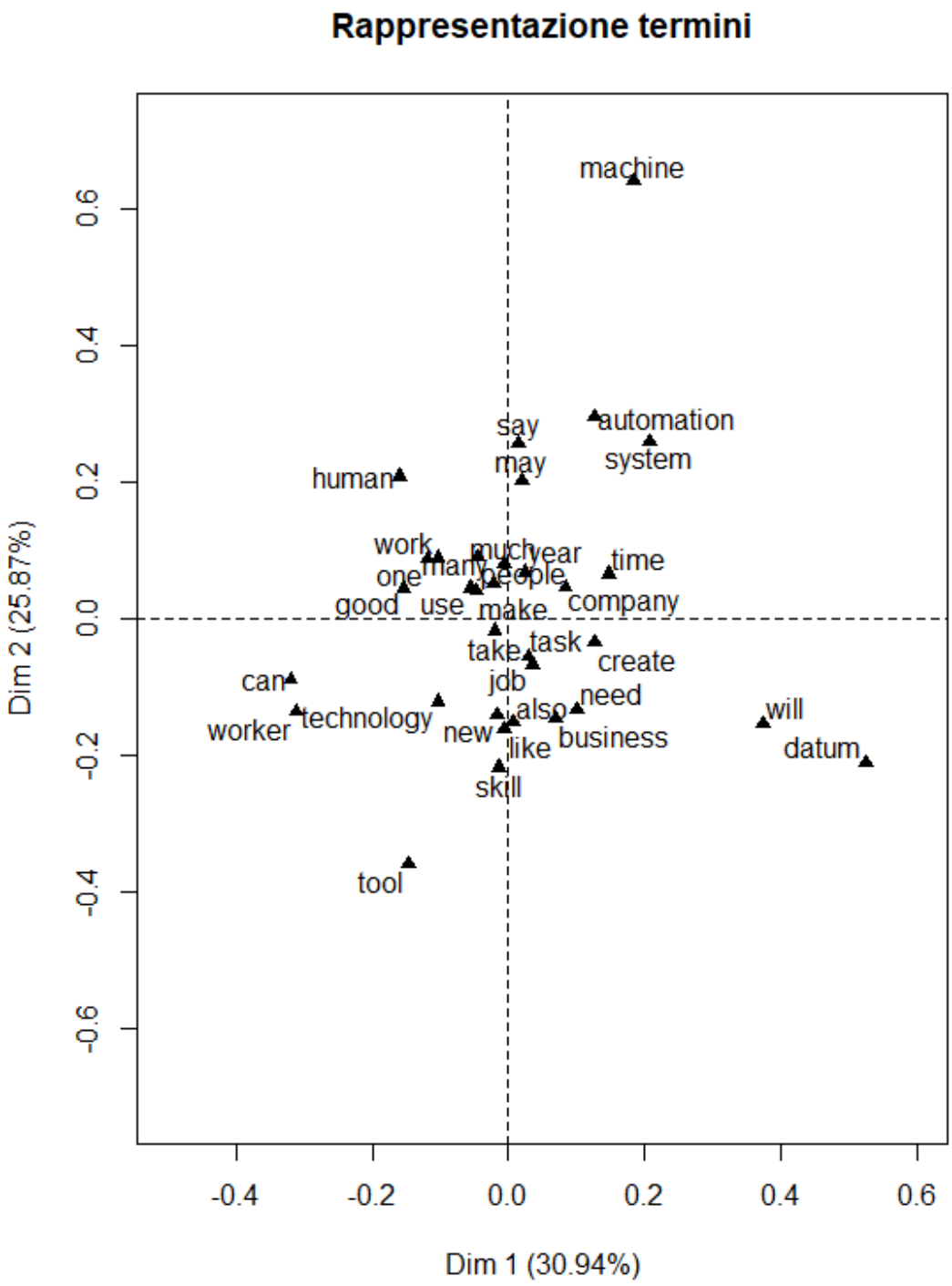
```
#con lemmatizzazione: prime 2 dimensioni spiegano circa il 57% varianza

plot(res.LexCA,eigen=TRUE,selDoc=NULL,selWord=NULL, col="blue") # barplot;
dimensioni e percentuale della varianza spiegata da ciascuna di esse
```

Il primo barplot rappresentato evidenzia come le prime due dimensioni riescano a spiegare circa il 57% della variabilità dei dati, questo permette di considerare le successive rappresentazioni su piani bidimensionali come rappresentative.

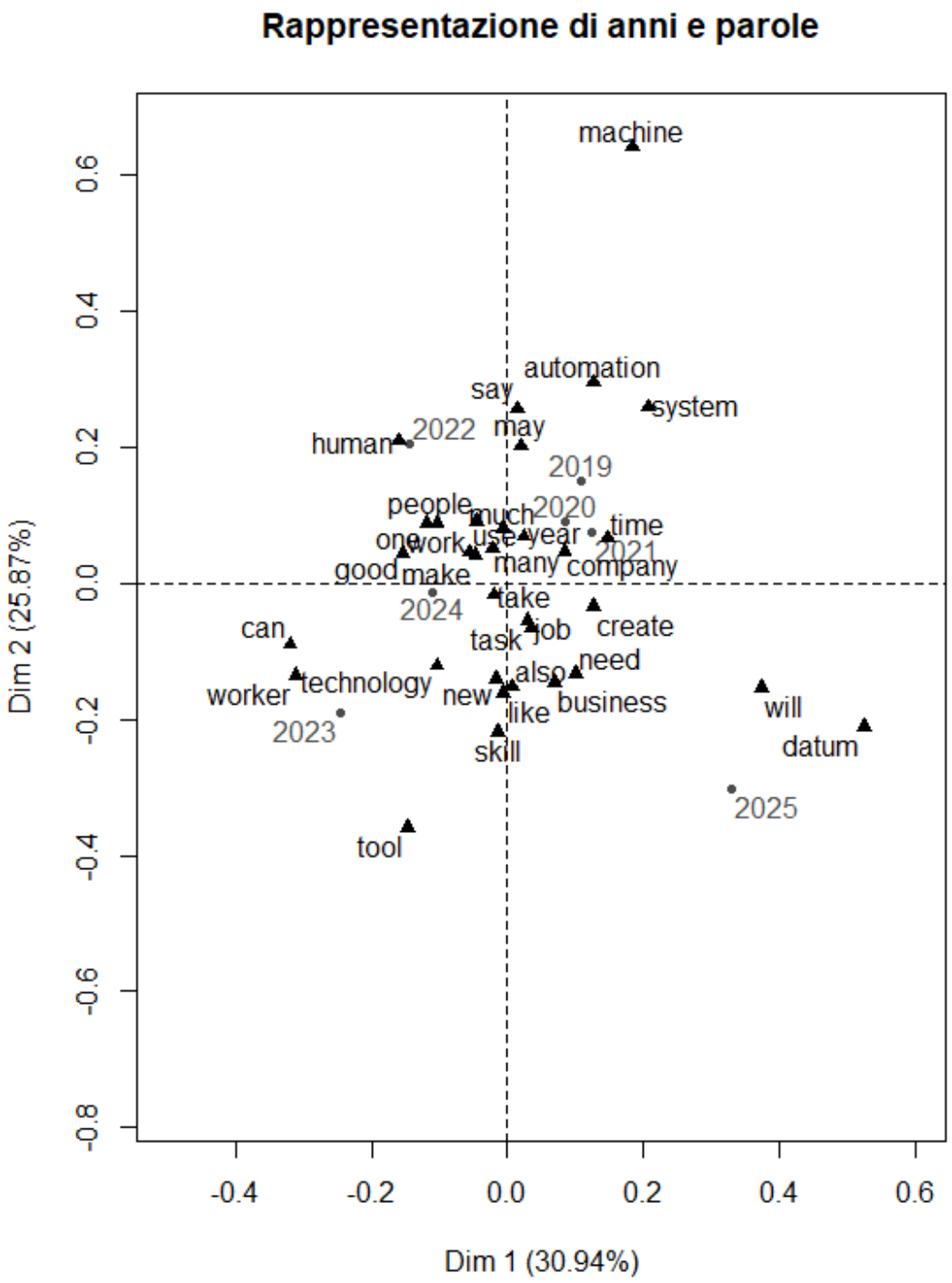
```
#varianza spiegata dalle singole dimensioni

plot(res.LexCA,selWord=NULL,xlim=c(-0.5,0.6),ylim=c(-
0.5,0.5),cex=1,col.doc="grey30",
     title="Rappresentazione anni") #rappresentazioni delle categorie (documenti),
"selword = NULL"
```




```
plot(res.LexCA,selDoc=NULL,xlim=c(-0.5,0.6),ylim=c(-
0.5,0.5),col.word="black",cex=1,
     title="Rappresentazione termini")
```



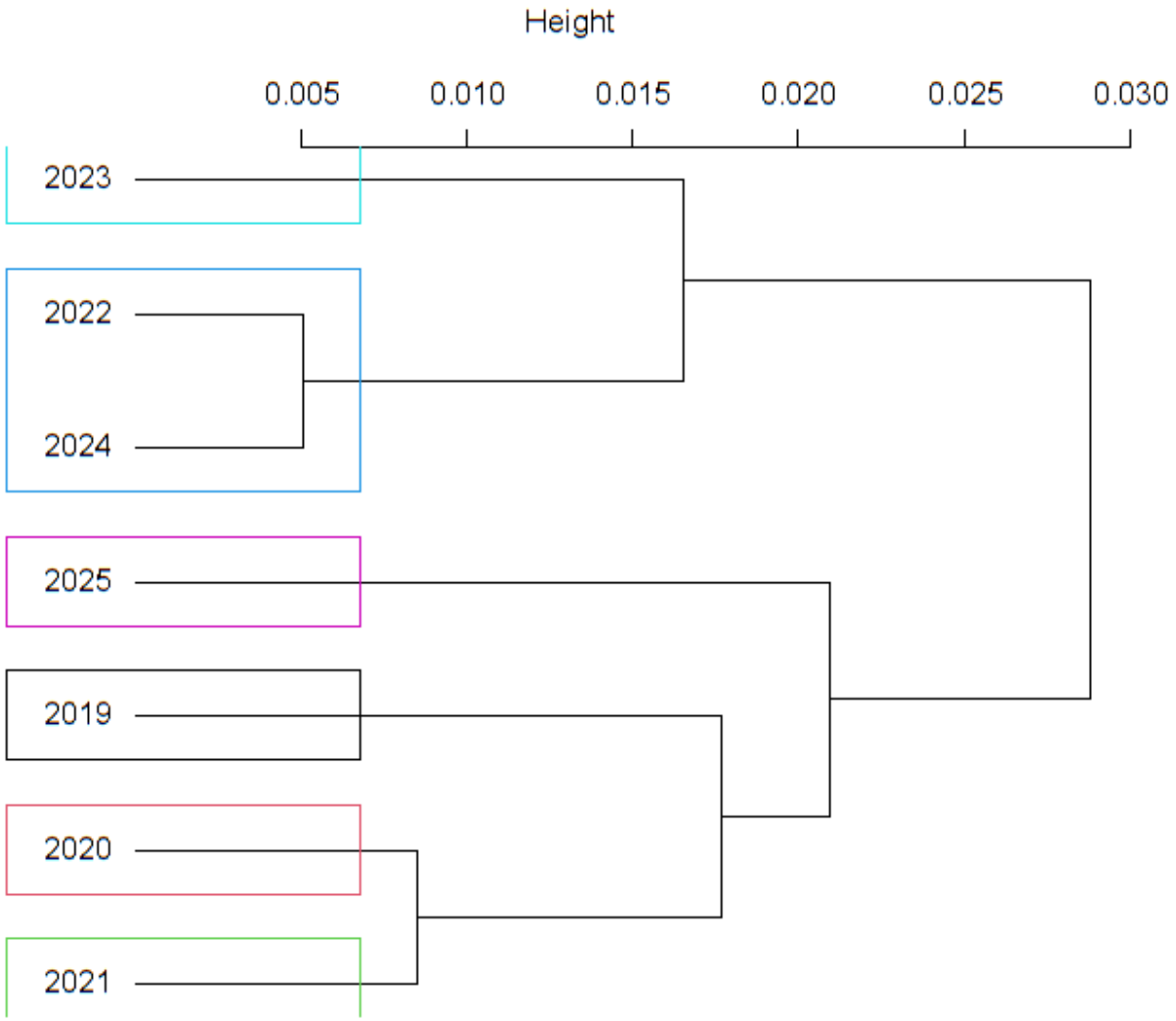

```
#rappresentazioni delle parole ,"seldoc = NULL"


plot(res.LexCA,xlim=c(-0.5,0.6),ylim=c(-
0.5,0.4),col.doc="grey30",col.word="black",cex=1,
     title="Rappresentazione di anni e parole") #Parole e anni rappresentati nello
stesso piano fattoriale.
```

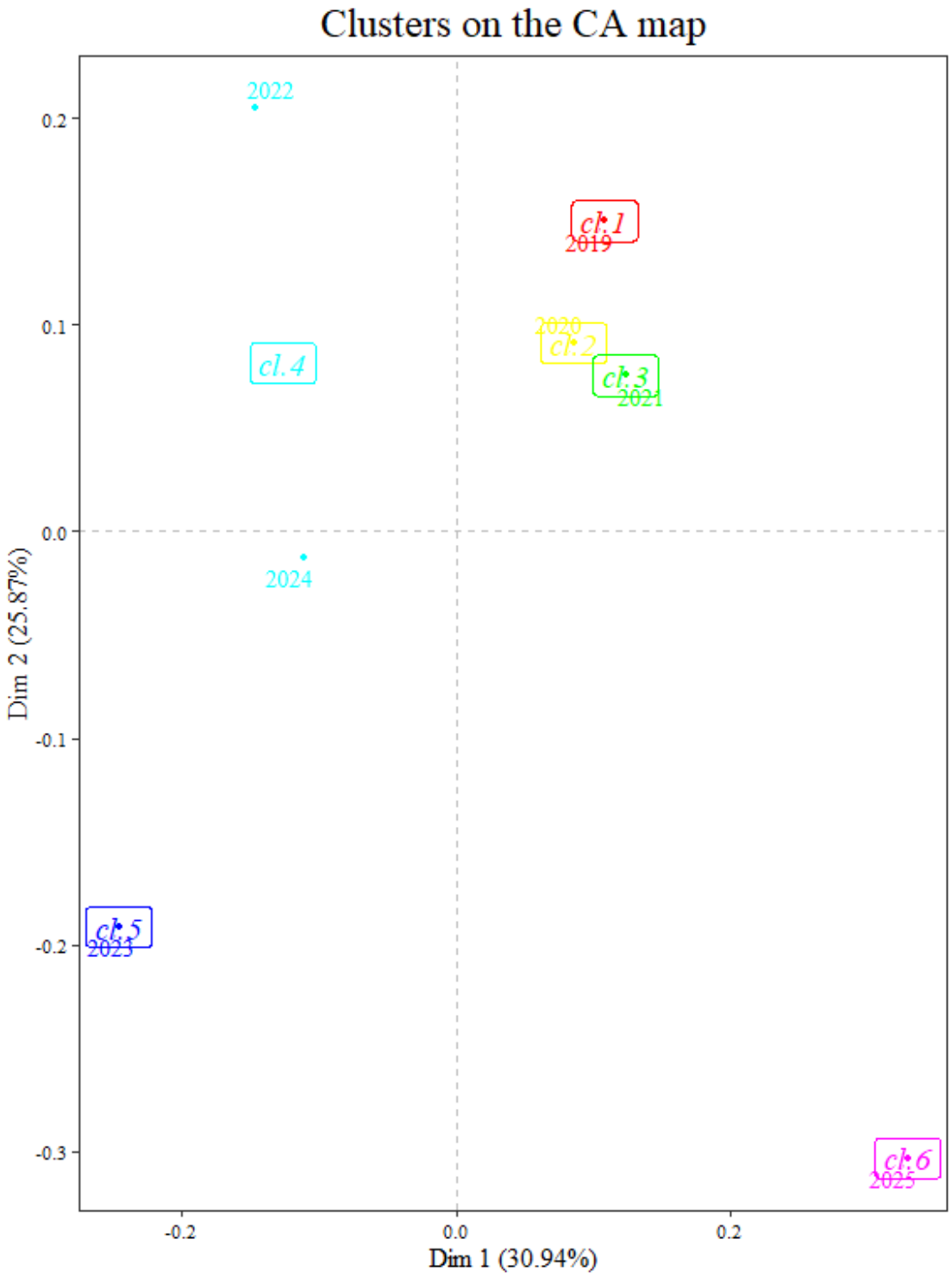


Dalla rappresentazione dei diversi anni nel piano, risulta che gli articoli risalenti agli anni dal 2019 al 2021 sono simili dal punto di vista di entrambe le dimensioni, nel 2022 e nel 2023 si ha un distacco rispetto ad entrambe le dimensioni, salvo poi ritornare su posizioni simili agli anni iniziali nel 2024, che essendo vicino all'origine denota la presenza di argomenti trattati in ogni articolo. Infine, nel 2025, si ha un nuovo distacco dalla medietà, in direzioni diverse rispetto a quelle del 2022 e 2023.

Aggiungendo i termini alla rappresentazione, si nota come tra i termini che, essendo collocati in prossimità dell'origine sono di comune utilizzo tra gli articoli vi sono gli impieghi ,"job", le aziende, "company", ed il termine "use" riferito probabilmente all'utilizzo delle nuove tecnologie. Inoltre, alcune delle vicinanze tra termini ed anni erano già state trovate nella comparison cloud e sono state qui confermate, tra di esse si menziona il forte focus degli articoli del 2025 sui dati ("datum"); gli esseri umani "human" ed il loro rapporto con l'IA come tema centrale degli articoli del 2022 ed il presentarsi di una concezione di lavoratore "worker" strettamente legata alla tecnologia "technology".

```
###clustering con Xploretext sulle 5 dimensioni

res.hcca_year<-LexHCca(res.LexCA,cluster.CA="docs", nb.clust=-1, max = 10,
order=TRUE,nb.par=10,graph=TRUE)
```

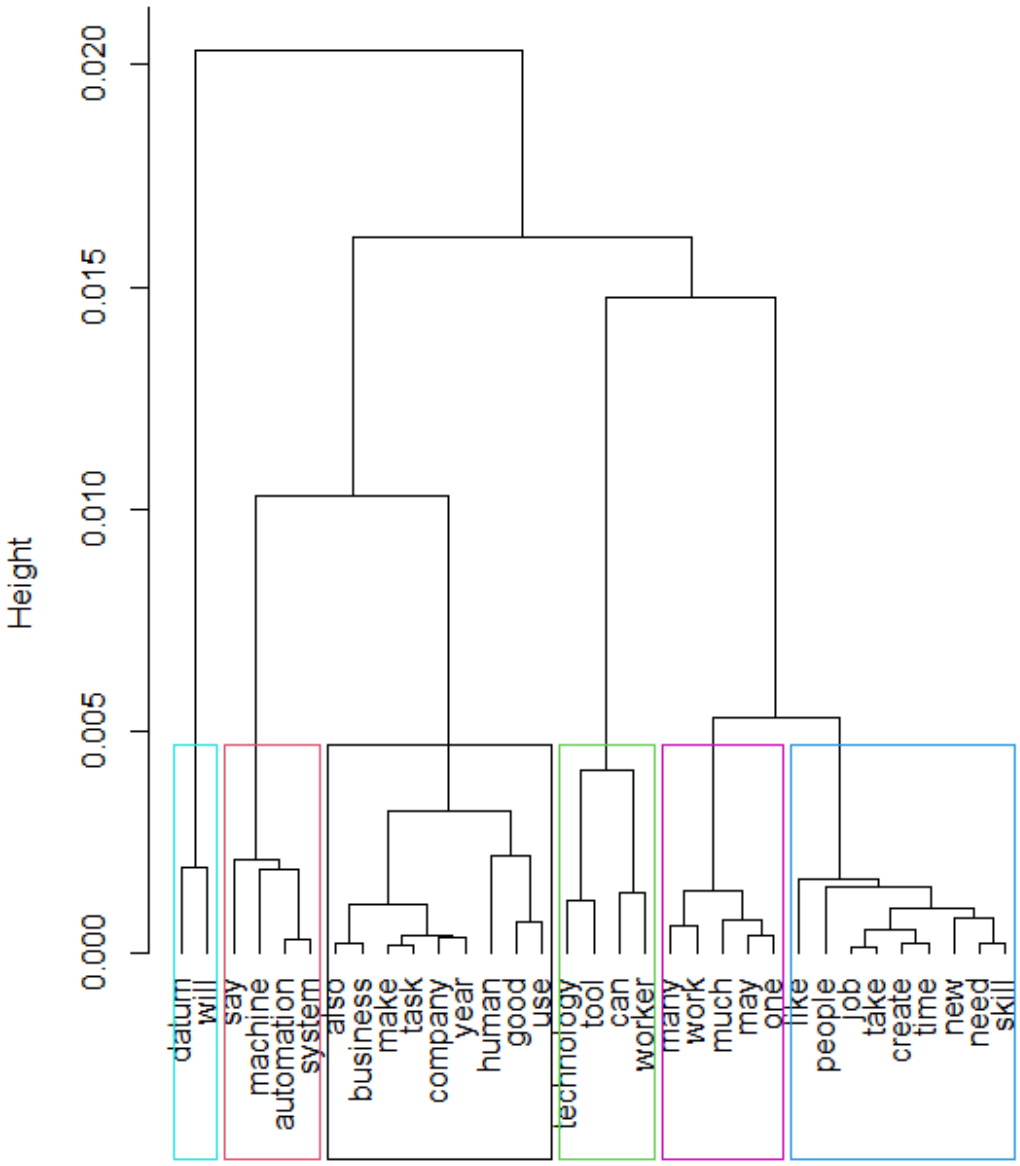

Height
0.005
0.010
0.015
0.020
0.025
0.030
2023
2022
2024
2025
2019
2020
2021



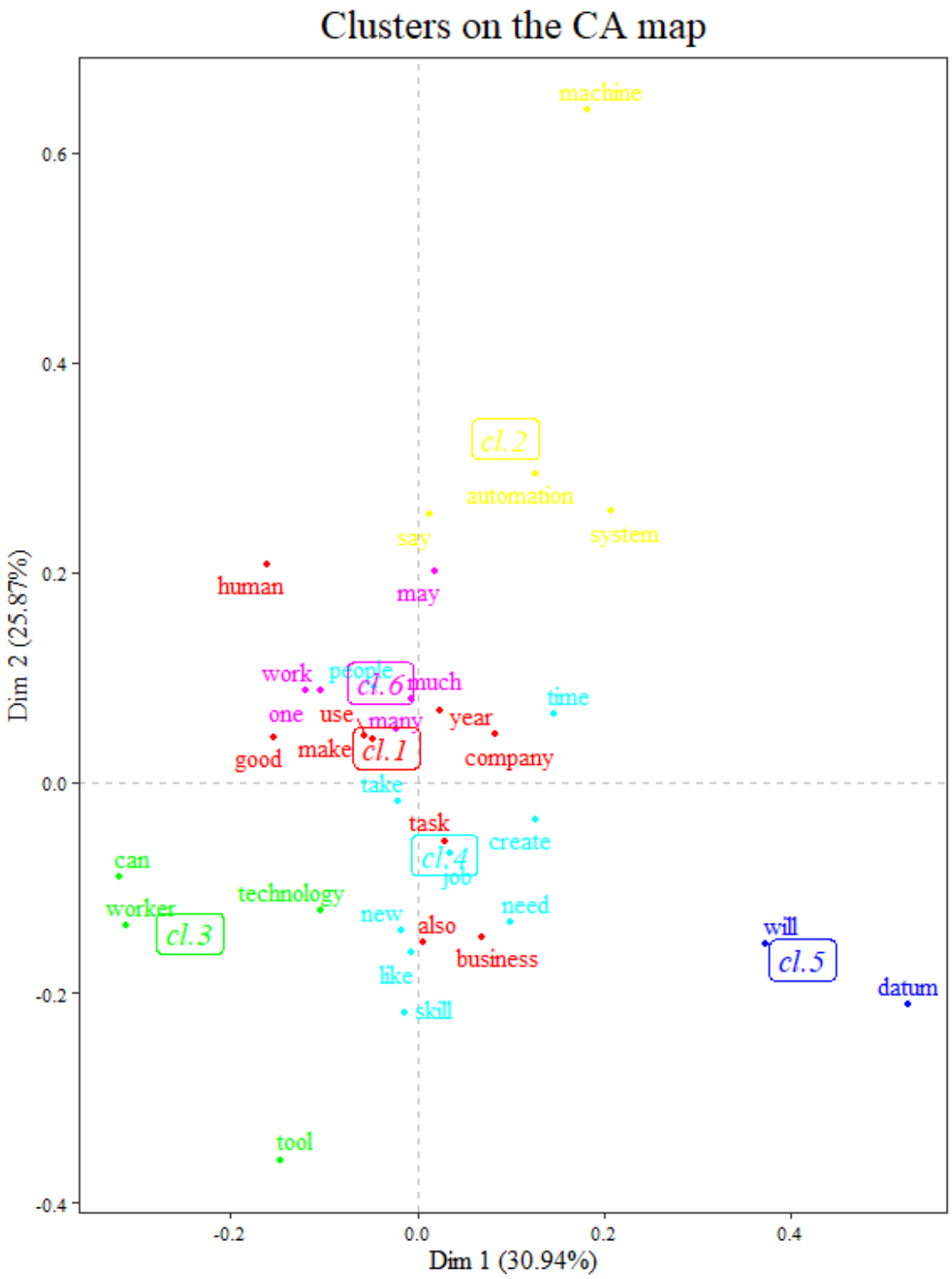

```
plot(res.hcca_year, words = T,  cex = 1.2, max.overlaps = Inf)  #benchè venga specificato di mostrare le parole, risultano presenti unicamente gli anni
```

```
#poichè molti dei cluster contengono un unico elemento, non vengono calcolati in questo caso gli indici di prestazioni
```

```r
#clustering gerarchico sui termini e le loro posizioni

res.hcca_words<-LexHCca(res.LexCA,cluster.CA="words", nb.clust=-1,min = 6, max = 10, order=TRUE,nb.par=10,graph=TRUE) ##tentativi da fare riguardo il minimo ed il massimo numero di clusters: alto livello di separazione tra i cluster
```

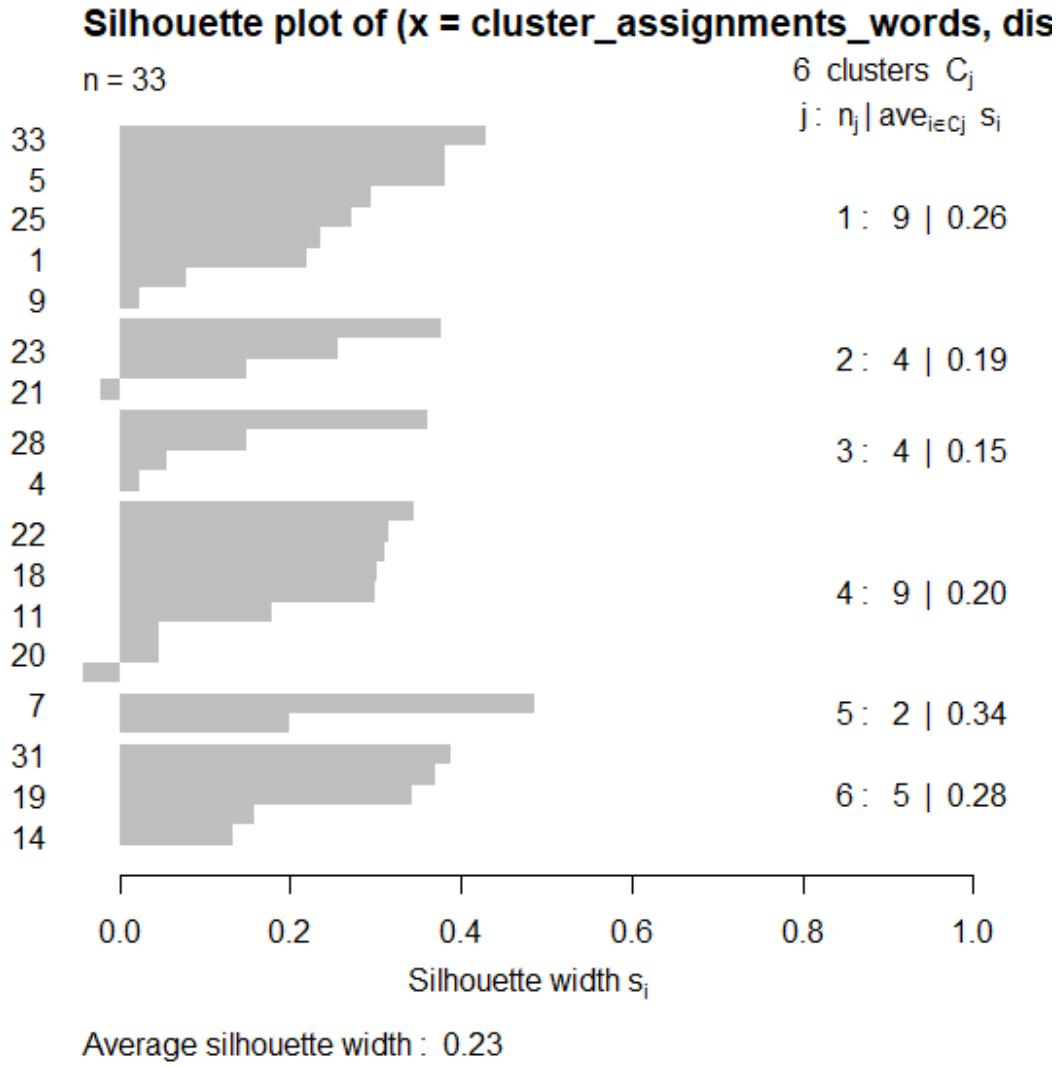


```r
plot(res.hcca_words, words = T,  cex = 1.2, max.overlaps = Inf)
```




```r
cluster_assignments_words <- res.hcca_words$data.clust$Clust_  # Cluster labels

# Compute the distance matrix (use appropriate method)
dist_matrix_word <- dist(res.LexCA$col$coord, method = "euclidean") #matrice delle
distanze tra le parole considerando le loro coordinate nelle 5 dimensioni ed una
```

```r
misura di distanza euclidea
library(cluster)
# Compute silhouette scores
silhouette_scores_words <- silhouette(cluster_assignments_words, dist_matrix_word)
plot(silhouette_scores_words, border = NA)  ##prestazioni
```



Dal momento che la funzione LexHCca effettua un clustering gerarchico sui documenti o sui termini in uno spazio fattoriale, si è deciso di effettuare prima il clustering sui documenti (in questo caso gli anni, essendo gli articoli qui considerati in modo aggregato) e poi sui termini. In questa fase, la rappresentazione nel piiano bidimensionale e l'assegnazione ai clusters possono essere fuorvianti in quanto per definire i raggruppamenti sono utilizzate le 5 dimensioni dello spazio fattoriale ottenuto, di cui ne sono rappresentate unicamente 2.

Il clustering sui termini appare più interpretabile ed i vari cluster possono essere così presentati:

- cl.1: L'utilizzo della tecnologia basata sull'IA nell'ambito dell'impresa e del suo business
- cl.2: Applicazione di sistemi e macchine basati sull'IA per l'automatizzazione di compiti
- cl.3: La possibilità da parte dei lavoratori di utilizzare la tecnologia come strumento di supporto.
- cl.4: La creazione id opportunità dovute alla nascita di un tipo di impiego nuovo, che necessita abilità dal punto di vista tecnologico.
- cl.5: L'importanza dei dati, anche in prospettiva futura, nell'ambito dell'IA.
- cl.6: L'influenza in termini di scala dell'IA sul lavoro delle persone.

L'analisi dell'indice di silhouette in questo clustering mostra un valore medio di 0.23, il che indica una qualità di clustering piuttosto bassa.

I cluster variano in termini di coesione e separabilità. Il cluster 5 ha il valore medio di silhouette più alto (0.34), suggerendo una buona separazione, mentre il cluster 3 ha il valore più basso (0.15), indicando una sovrapposizione maggiore con altri cluster.

Il numero di elementi nei cluster varia significativamente: il cluster 1 è il più grande (9 elementi), mentre il cluster 5 è il più piccolo (2 elementi).

La maggior parte dei valori di silhouette individuali sono inferiori a 0.5, suggerendo che molti punti sono vicini ai confini tra cluster, il che indica la caratteristica dei dati di avere una scarsa distinzione tra gruppi.

3.6 LSA Effettuazione della Latent Semantic Analysis

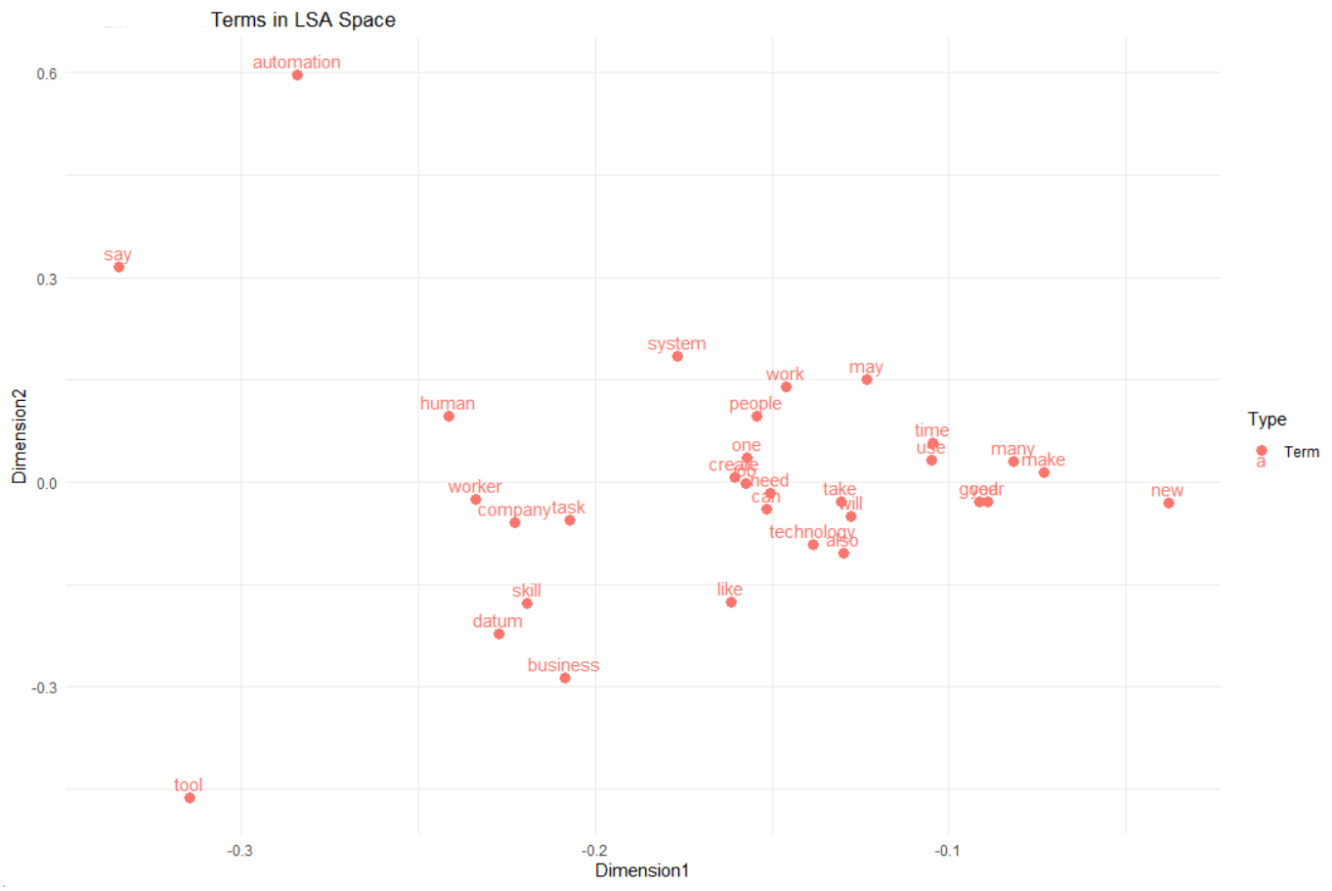
In questa fase si vuole mettere in luce l'eventuale presenza di termini diversi che, ricorrendo in contesti simili, sono semanticamente legati. Si noti che la Correspondence Analysis potrebbe non aver catturato questo legame semantico perchè "latente" rispetto alla varibilità linguistica del corpus.

```
library(tm)
library(lsa)

## Caricamento del pacchetto richiesto: SnowballC

##
## Caricamento pacchetto: 'lsa'
```

```
## Il seguente oggetto è mascherato da 'package:koRpus':
##
##     query

library(ggplot2)
library(textstem)

#il dataset viene ri-caricato perchè precedentemente era stato modificato

library(readxl)
gruppo_articoli<-
read_excel("C:/Users/1jaco/OneDrive/Desktop/SMA/Progetto_SMA/Dataset_finale/Datase
testeso.xlsx")

#Inizialmente si preprocessa il corpus
mycorpus_lsa<-Corpus(VectorSource(gruppo_articoli$testo))
mycorpus_lsa <- tm_map(mycorpus_lsa, content_transformer(tolower))

## Warning in tm_map.SimpleCorpus(mycorpus_lsa, content_transformer(tolower)):
## transformation drops documents

mycorpus_lsa <- tm_map(mycorpus_lsa, removeNumbers)

## Warning in tm_map.SimpleCorpus(mycorpus_lsa, removeNumbers): transformation
## drops documents

mycorpus_lsa <- tm_map(mycorpus_lsa, removePunctuation)

## Warning in tm_map.SimpleCorpus(mycorpus_lsa, removePunctuation): transformation
## drops documents

remove_custom_punctuation <- function(text) {gsub("[[:punct:]]", " ", text) }
# Rimuove tutti i segni di punteggiatura comprese le virgolette (trasformandoli in
spazi)
# Applicare la funzione al corpus
mycorpus_lsa <- tm_map(mycorpus_lsa,
content_transformer(remove_custom_punctuation))

## Warning in tm_map.SimpleCorpus(mycorpus_lsa,
## content_transformer(remove_custom_punctuation)): transformation drops documents

mycorpus_lsa <- tm_map(mycorpus_lsa, stripWhitespace)

## Warning in tm_map.SimpleCorpus(mycorpus_lsa, stripWhitespace): transformation
## drops documents

mycorpus_lsa<-tm_map(mycorpus_lsa, removeWords, stopwords("en"))

## Warning in tm_map.SimpleCorpus(mycorpus_lsa, removeWords, stopwords("en")):
## transformation drops documents

theme_words <- c( "ai", "artificial", "intelligence")
mycorpus_lsa <- tm_map(mycorpus_lsa, removeWords, theme_words)
```

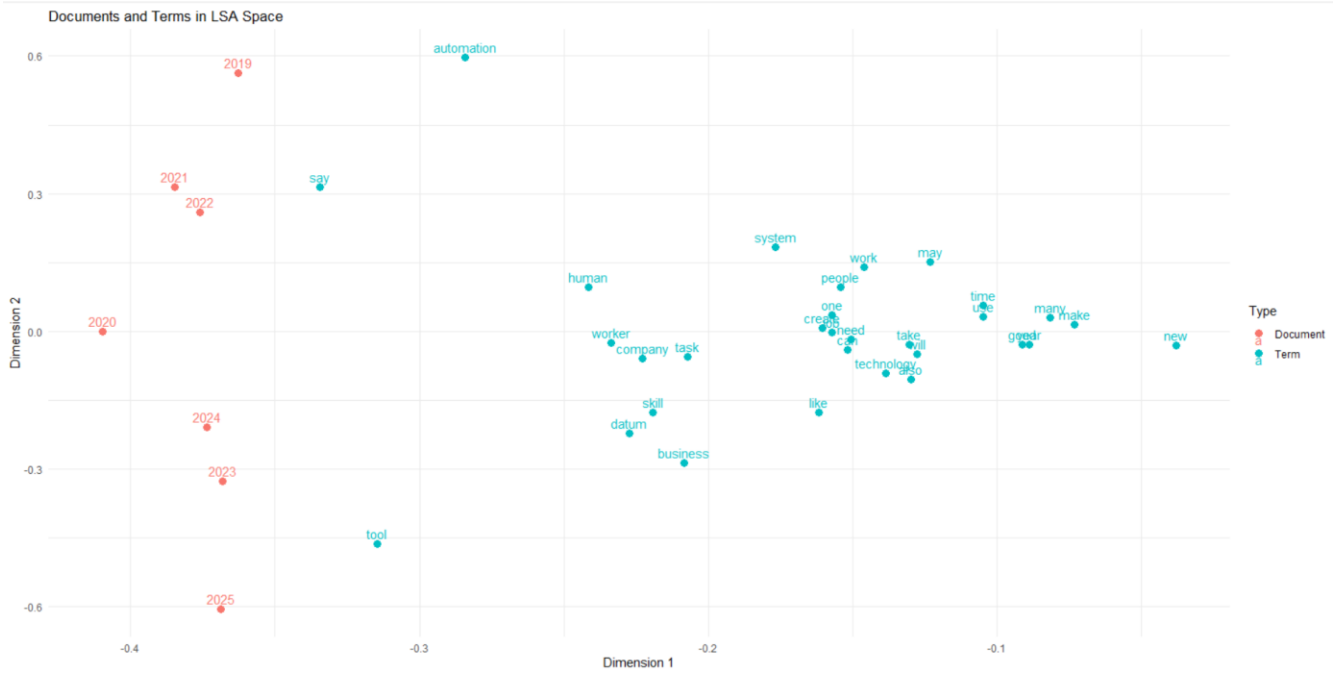
```
## Warning in tm_map.SimpleCorpus(mycorpus_lsa, removeWords, theme_words):
## transformation drops documents

#si parte da una term-document matrix che utilizza un sistema di pesi basato sulle
frequenza assolute dei termini, per poi estrarre i termini maggiormente
ricorrenti, su cui effettuare la lsa.

# Lemmatizzazione
mycorpus_lsa_lem <- tm_map(mycorpus, lemmatize_strings)

## Warning in tm_map.SimpleCorpus(mycorpus, lemmatize_strings): transformation
## drops documents

##########mycorpus <- mycorpus_lem
tdm_raw <- TermDocumentMatrix(mycorpus_lsa_lem, control = list(weighting =
weightTf))

## Warning in TermDocumentMatrix.SimpleCorpus(mycorpus_lsa_lem, control =
## list(weighting = weightTf)): custom functions are ignored

#Imposizione di una soglia minima di occorrenze nel corpus
threshold <- 120

#####120 -> 31 termini + frequenti
freq_terms_lsa <- findFreqTerms(tdm_raw,lowfreq = threshold)

# Creazione della Term-DocumentMatrix (TDM) con la funzione di pesi TF-IDF,
includendo solo i termini + frequenti
tdm_tfidf<- TermDocumentMatrix(mycorpus_lsa_lem, control = list(weighting=
weightTfIdf, dictionary = freq_terms_lsa))

## Warning in TermDocumentMatrix.SimpleCorpus(mycorpus_lsa_lem, control =
## list(weighting = weightTfIdf, : custom functions are ignored

tdm_tfidf_matrix<- as.matrix(tdm_tfidf)

dim(tdm_tfidf_matrix)

## [1] 31 70

TL_tfidf<-matrix(nrow=31,ncol=7) #inizialemnte è una matrice vuota, con tante
righe quanti sono i vocaboli diversi e tante colonne quanti sono gli anni in cui
gli articoli sono stati scritti.
for (j in 1:31)
{TL_tfidf[j,]<-tapply(tdm_tfidf_matrix[j,],gruppo_articoli$anno, mean)
TL_tfidf<-rbind(TL_tfidf)}              ###nel creare una tabella lessicale
aggregata, contenente i valori di ogni termine per ogni anno, essendo stato
utilizzato in questo caso un sistema di pesi di tipo tf-idf, è opportuno
effettuare la media
colnames(TL_tfidf)<-c("2019","2020", "2021", "2022", "2023", "2024", "2025")
row.names(TL_tfidf)<-row.names(tdm_tfidf_matrix)
```

```r
# Effettuazione della LatentSemantic Analysis (LSA)
lsa_result<- lsa(TL_tfidf)

#coordinate termini
term_coords<- as.data.frame(lsa_result$tk)
rownames(term_coords) <- rownames(tdm_tfidf_matrix)

#coordinate documenti
document_coords <- as.data.frame(lsa_result$dk)
rownames(document_coords) <- paste( 2019:2025)

term_labels<- rownames(term_coords)
doc_labels <- rownames(document_coords)

# Combine term data into  a single data frame

#per maggiore chiarezza, inizialmente si estraggono unicamente le coordinate dei termini nello spazio semantico.
terms_data<-
  (cbind(term_coords[, 1:2], Type= "Term", Label = term_labels)) #utilizzo di unicamente le prime due dimensioni dello spazio ridotto, creazione di un dataframe contenente le diverse parole con le rispettive coordinate nelle prime due dimensioni del sottospazio ridotto


colnames(terms_data)[1:2] <- c("Dim1", "Dim2")
terms_data$Dim1 <- as.numeric(as.character(terms_data$Dim1))
terms_data$Dim2 <- as.numeric(as.character(terms_data$Dim2))

# Visualizzazione di termini nello spazio semantico
ggplot(terms_data,aes(x = Dim1, y = Dim2, color = Type, label = Label)) +
  geom_point(size = 3) +
  geom_text(hjust= 0.5, vjust= -0.5) +
  labs(
    title= "Terms in LSA Space",
    x = "Dimension1",
    y = "Dimension2"
  ) +
  theme_minimal()
```




```r
#combinazione dei dati su termini e documenti

combined_data <- rbind(
  cbind(document_coords[, 1:2], Type = "Document", Label = doc_labels),
  cbind(term_coords[, 1:2], Type = "Term", Label = term_labels)
)


# Ensure combined_data has numeric columns for plotting
colnames(combined_data)[1:2] <- c("Dim1", "Dim2")
combined_data$Dim1 <-
  as.numeric(as.character(combined_data$Dim1))
combined_data$Dim2 <-
  as.numeric(as.character(combined_data$Dim2))
# Step 6: Visualize documents and terms in the LSA space
ggplot(combined_data, aes(x = Dim1, y = Dim2, color = Type,
                          label = Label)) +
  geom_point(size = 3) +
  geom_text(hjust = 0.5, vjust = -0.5) +
  labs(
    title = "Documents and Terms in LSA Space",
    x = "Dimension 1",
    y = "Dimension 2"
```

```
  ) +
  theme_minimal()
```



La rappresentazione dei risultati della LSA permette di individuare tre gruppi semantici che emergono dai testi degli articoli:

- La nascita di un nuovo modello di business fortemente basato sui dati e sulle abilità di gestirli, probabilmente nell'ottica di abilitare applicazioni IA.
- La necessità dei lavoratori delle aziende di acquisire nuove abilità inerenti all'utilizzo delle tecnologie dell'IA.
- Un nuovo tipo di lavoro in cui si ha la coesistenza di persone e sistemi di tipo tecnologico basati sull'IA.

Le posizioni degli anni non sono altamente rappresentative, questo è probabilmente dovuto al fatto che sono stati ottenuti da un'aggregazione a partre dagli articoli di loro competenza. Tuttavia, dal punto di vista della seconda dimensione si può affermare che il 2019 è molto prossimo al termine "automation", e dunque, alla possibilità di automatizzare compiti con l'utilizzo dell'IA; gli articoli del 2020 hanno come tema semantico centrale quello dei lavoratori ("worker"), mentre gli anni 2021 e 2022, essendo prossimi a "say" suggeriscono il fatto che negli articoli sono stati riportati pareri di terzi, probabilmente esperti. Sempre concentrandosi sulla seconda dimensione, gli anni 2023, 2024 e 2025 si concentrano, dal punto di vista semantico sui dati ("datum"), sul modo in cui questi possono influenzare il business e sulle abilità ("skill") e gli strumenti ("tool") necessary per gestirli ed in generale, per usare al meglio l'IA.

3.7 Topic modelling

Applicazione del modello LDA per il topic modelling del corpus di articoli. Lo scopo che ci si prefigge in questa fase è quello di far risaltare i temi maggiormente trattati all'interno del corpus. Benchè gli articoli sono stati selezionati basandosi sulla base della loro tematica principale, ossia l'IA ed il mondo del lavoro, è interessante determinare altre tematiche che, comparendo in questi documenti, sono collegate a quella principale.

```
library(topicmodels)
library(tm)
library(quanteda)

## Package version: 4.1.0
## Unicode version: 15.1
## ICU version: 74.1

## Parallel computing: 12 of 12 threads used.

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

##
## Caricamento pacchetto: 'quanteda'

## I seguenti oggetti sono mascherati da 'package:koRpus':
##
##     tokens, types

## Il seguente oggetto è mascherato da 'package:tm':
##
##     stopwords

## I seguenti oggetti sono mascherati da 'package:NLP':
##
##     meta, meta<-

library(readxl)
library(textstem)



#coercizione di mycorpus_lem: la funzione tokens accetta unicamente oggetti di
tipo: simple corpus
corpus_lem <- corpus(mycorpus_lem)

tok <- quanteda::tokens(corpus_lem, remove_punct = TRUE, remove_symbols = TRUE)
tok <- quanteda::tokens_remove(tok, stopwords("en"))



dtm <- dfm(tok, tolower = TRUE) #creazione della document-feature matrix a partire
da tok, in cui si convertono i termini al minuscolo.

 #eliminazione di termini che compaiono in meno di 2 documenti
 cdfm <- dfm_trim(dtm, min_docfreq = 2)
```

```r
 #eliminazione dei termini formati da meno di due caratteri
 cdfm<- dfm_clean <- dfm_select(cdfm, selection = "keep", min_nchar = 2)

 # estimate LDA with K topics
 set.seed(123)
 K <- 4
 lda <- LDA(cdfm, k = K, method = "Gibbs",
control = list(verbose=25L, seed = 123, burnin = 25, iter = 500))
```

```
## K = 4; V = 2766; M = 70
## Sampling 525 iterations!
## Iteration 25 ...
## Iteration 50 ...
## Iteration 75 ...
## Iteration 100 ...
## Iteration 125 ...
## Iteration 150 ...
## Iteration 175 ...
## Iteration 200 ...
## Iteration 225 ...
## Iteration 250 ...
## Iteration 275 ...
## Iteration 300 ...
## Iteration 325 ...
## Iteration 350 ...
## Iteration 375 ...
## Iteration 400 ...
## Iteration 425 ...
## Iteration 450 ...
## Iteration 475 ...
## Iteration 500 ...
## Iteration 525 ...
## Gibbs sampling completed!
```

```r
 #Estrazione dei termini più probabili per ciascun topic
 terms <- get_terms(lda, 15)

 print(terms)
```

```
##       Topic 1   Topic 2    Topic 3        Topic 4
##  [1,] "human"   "job"      "job"          "say"
##  [2,] "work"    "can"      "worker"       "use"
##  [3,] "may"     "skill"    "technology"   "company"
##  [4,] "one"     "role"     "many"         "datum"
##  [5,] "people"  "impact"   "new"          "hire"
##  [6,] "good"    "also"     "automation"   "tech"
##  [7,] "can"     "need"     "us"           "technology"
##  [8,] "machine" "future"   "create"       "software"
##  [9,] "make"    "increase" "work"         "business"
## [10,] "even"    "train"    "include"      "tool"
```

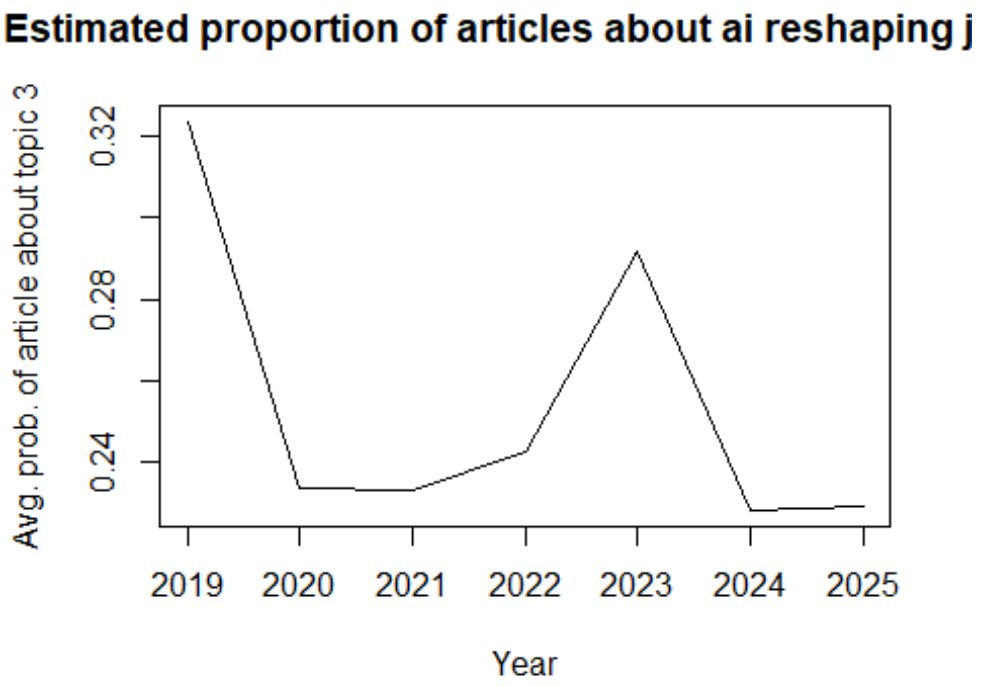
```
## [11,] "get"     "tool"     "year"         "see"
## [12,] "think"   "datum"    "robot"        "year"
## [13,] "take"    "help"     "little"       "firm"
## [14,] "like"    "become"   "replace"      "process"
## [15,] "task"    "employee" "productivity" "generative"
```

I temi individuati possono essere così riassunti:

- Topic 1: Le nuove caratteristiche degli impieghi e i campi di capacità a cui devono attingere i lavoratori per svolgerli.
- Topic 2: Riflessione sull'interazione uomo macchina nel contesto lavorativo.
- Topic 3: Gli effetti dell'introduzione dell'IA sul mondo del lavoro, in particolare, sul fabbisogno di forza lavoro.
- Topic 4: La ricerca riguardo l'IA e le applicazioni dei risultati di questa nelle aree di business delle aziende.

```
#topic più probabile per ciascun documento
 topics <- get_topics(lda, 1)
 head(topics)

## text1 text2 text3 text4 text5 text6
##     1     1     3     3     3     1

 gruppo_articoli$pred_topic <- topics



  # Topic X
 X <- 3
 paste(terms[,X], collapse=", ")

## [1] "job, worker, technology, many, new, automation, us, create, work, include,
year, robot, little, replace, productivity"

 #Aggiunge una colonna al dataset, contenente, per ogni, articolo, la probabilità
di riguardare il topic X

 gruppo_articoli$prob_topic <- lda@gamma[,X]

 #Aggregazione a livello annuale, e calcolo della probabilità media di un articolo
di un certo anno di riguardare il topic X

 agg <- aggregate(gruppo_articoli$prob_topic, by=list(year=gruppo_articoli$anno),
FUN=mean)



#andamento negli anni della probabilità calcolata precedentemente
 plot(agg$year, agg$x, type="line", xlab="Year", ylab="Avg. prob. of article about
```

```
topic 3",
 main="Estimated proportion of articles about ai reshaping jobs")

## Warning in plot.xy(xy, type, ...): il tipo plot 'line' sarà troncato al primo
## carattere
```



Si nota come il tema del cambiamento apportato dall'IA al fabbisogno di forza lavoro viene trattato maggiormente nel 2019, primo anno dal punto di vista cronologico per cui si hanno articoli. La proporzione stimata di articoli su questo tema decresce nel 2020 e nel 2021 per poi ri-aumentare nel biennio 2022-2023 ed infine raggiungere livelli bassi nel 2024-2025. La probabilità relativamente elevata del 2019 può essere data dal fatto che inizialmente, non essendo quello dell'IA un ambito completamente conosciuto e delineato, si nutriva il timore proprio che contraddistingue ciò che non si conosce appieno. L'aumento della probabilità negli anni 2022/2023 potrebbe derivare dallo sviluppo di declinazione dell'IA come quella generativa, altamente impattanti anche nei contesti lavorativi "white-collar".