RAPORT DATA MINING – TRANSFORMACJA IT-IDF, ANALIZA KLASTRA PRZY UZYCIU R

Emil Filipowicz

Raport eksploracji korpusu zawierającego dane tekstowe przedstawione w języku angielskim, o tematyce opartej na sztucznej inteligencji (ang. AI). Raport przedstawia analizę klastra krok po kroku oraz wnioski z wykonanego badania.

1. ZAŁADOWANE PAKIETY

- library(tm)
- library(SnowballC)
- library(ggplot2)
- library(wordcloud)
- library(cluster)
- library(fpc)

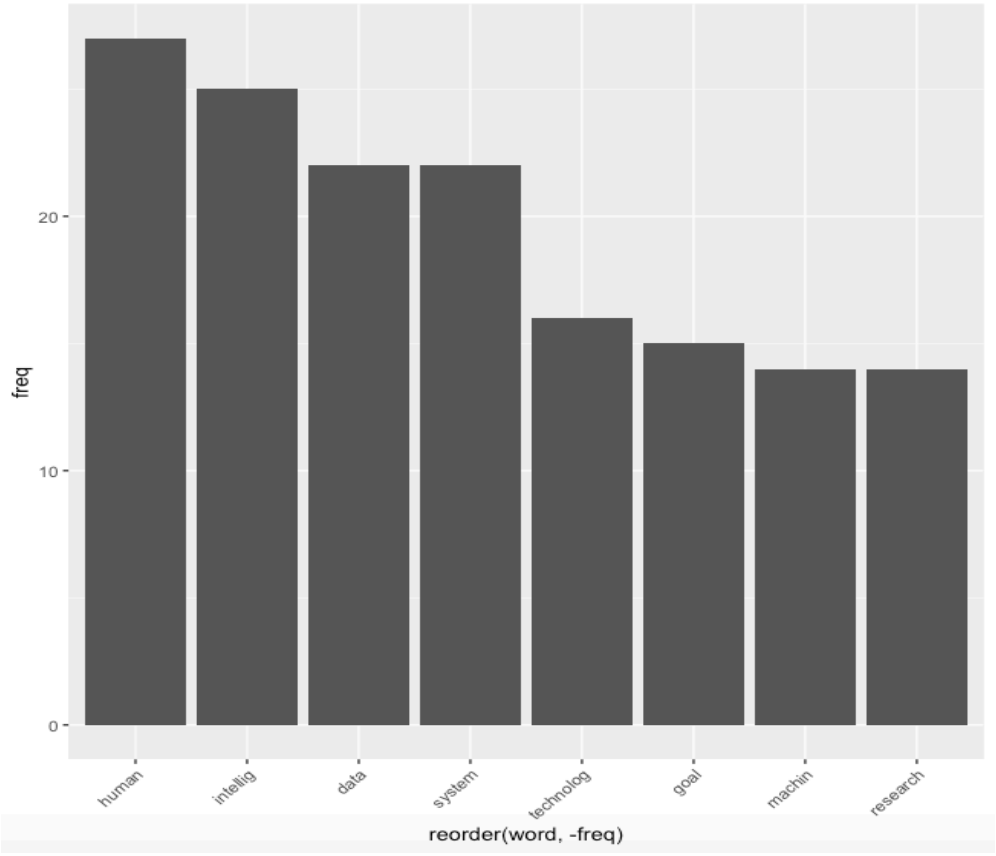
2. PRZYGOTOWANIE ZBIORU DANYCH

Zbiór danych wykorzystany w raporcie to zestawienie 10 plików tekstowych związanych ze sztuczną inteligencją. Do przeprowadzenia badań została również użyta macierz dtm powstała podczas czyszczenia oraz eksploracji danych tekstowych zawartych w korpusie.

```
getwd()
setwd("/Users/emilfilipowicz/Desktop/korpus")
```

```
MyData <-read.csv("/Users/emilfilipowicz/Desktop/DocumentTermMatrix.csv",
                  header = TRUE, sep = ",",
                  strip.white = TRUE, fill = TRUE,
                  comment.char = "#",
                  stringsAsFactors = FALSE)
```

3. WYKRES CZĘSTOTLIWOŚCI



Wykres przedstawiający 8 słów występujących z największą częstotliwością w badanym korpusie.

4. TWORZENIE MACIERZY TF-IDF - proces techniczny transformacji macierzy

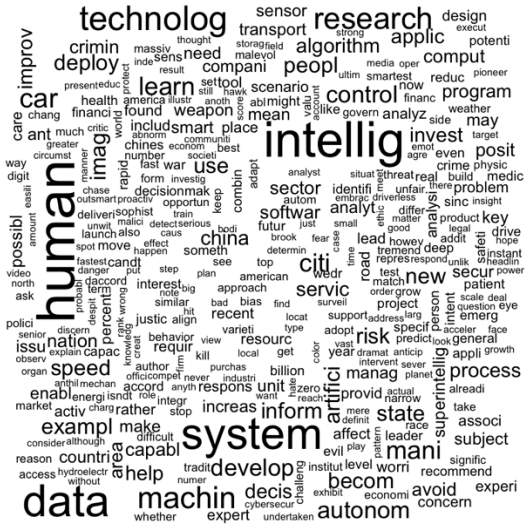
```
> tf[1:5,1:5]
          Dangerous AI.txt finance AI.txt Healt care and criminal justice.txt Myths AI.txt Qualities AI.txt
accomplish                1             0                                   0           0                0
achiev                    1             0                                   0           0                0
advanc                    1             3                                   1           0                0
agre                      1             0                                   0           0                1
aid                       1             0                                   0           0                0
> idf[1:5]
accomplish     achiev     advanc       agre        aid
 2.3025851  2.3025851  0.9162907  1.6094379  2.3025851

> head(idf_sort, 15)
    human  intellig       can   develop  research    system      make technolog       use     becom    exampl
0.2231436 0.2231436 0.3566749 0.3566749 0.3566749 0.3566749 0.3566749 0.3566749 0.3566749 0.5108256 0.5108256
     take    applic      data      need
0.5108256 0.5108256 0.5108256 0.5108256
> tail(idf_sort, 15)
       steer       suffer         surg     surround      suspend   systemsdus         taxi        truck
    2.302585     2.302585     2.302585     2.302585     2.302585     2.302585     2.302585     2.302585
        uber       unless       vehicl vehiclesdcar        vital        volvo         went
    2.302585     2.302585     2.302585     2.302585     2.302585     2.302585     2.302585
```

```
> tf[349:352,1:3]
      Dangerous AI.txt finance AI.txt Healt care and criminal justice.txt
way                 0              2                                    0
without             0              1                                    0
world               0              1                                    0
zero                0              1                                    0
> tf_idf_t<-t(tf_idf)
> tf_idf_t[349:352,1:3]
     Dangerous AI.txt finance AI.txt Healt care and criminal justice.txt
[1,]                0       1.832581                                    0
[2,]                0       1.609438                                    0
[3,]                0       1.609438                                    0
[4,]                0       1.609438                                    0
```

5. WORDCLOUDS - stworzony w oparciu o macierz TF-IDF

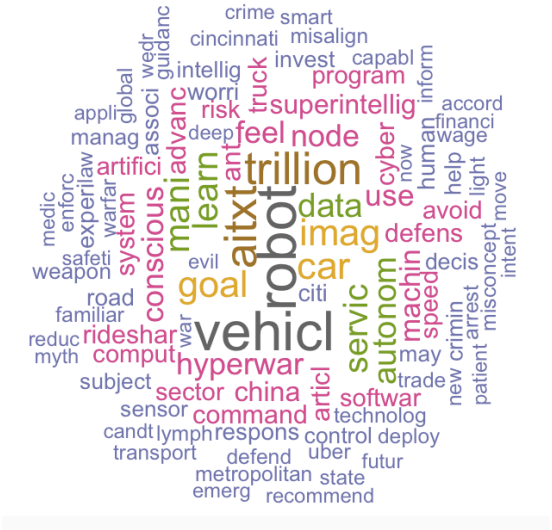
- **WordClouds TF**

WordClouds zbudowany na podstawie o macierz TF-IDF różnią się od tych tworzonych na podstawie TF. Główna różnica polega na celowości korzystania z tych metod. Po przeanalizowaniu WordClouds TF dla odbiorcy widoczny jest temat całego korpusu, innymi słowy na jakiej tematyce skupia się przyjęty zbiór danych. Tworząc korpus oparty na tematyce sztucznej inteligencji (ang. AI - Artifficial Intelligence) można przeczuć jakie słowa mogą występować najczęściej. Jednak to pozwala jedynie na poznanie ogólnej tematyki badanego zbioru dokumentów i nie pozwala na poznanie podtematów zawartch w korpusie.



- **WordClouds TF – IDF**

Do szerszego poznania korpusu służy TF-IDF. Ta metoda pozwala zbadać w ilu dokumentach pojawia się dane słowo, dzięki czemu powstaje możliwość redukcji słów wystepujących najczęściej we wszystkich dokumentach, a to za to skutkuję możliwością bardziej szczegółowego poznania problematyki poszczególnych dokumentów.



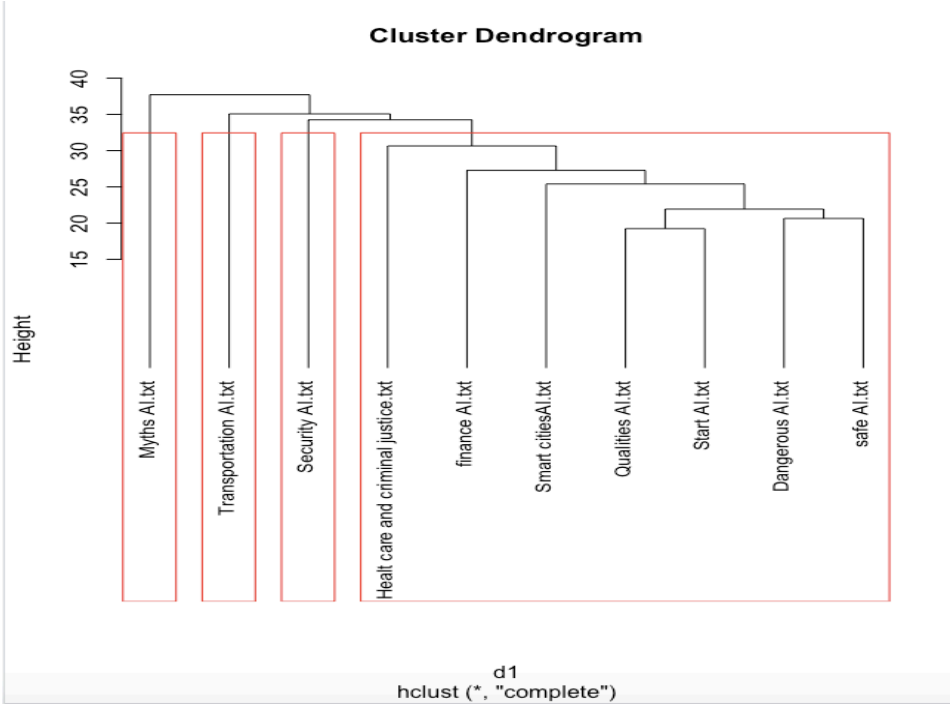
Na załączonych chmurach opartych na TF – IDF widać, że zniknęły słowa występujące najczęściej w dokumencie, które przedstawiały główną tematykę korpusu jak „human", „intellig" czy „system". Stało się tak ponieważ słowa występujące w każdym tekście korpusu poprzez transformacje TF-IDF otrzymują wagę 0 przez co stają się nie istotne w dalszym badaniu analizy pod tematów zawartych w dokumentach. Na ich miejscu pojawiły się słowa takiej jak „vehicl" czy „robot". Na podstawie tego można wywnioskować, że tematyka dokumentów omawia sztuczną inteligencję w kategoriach technicznego zastosowania jej w życiu codziennym. Dodatkowo można też zauważyć zwiększenie wagi takich słów jak „autonom", „superintellig" i „aitxt", które potwierdzają te stwierdzenie.

6. KLASTROWANIE HIERARCHICZNE

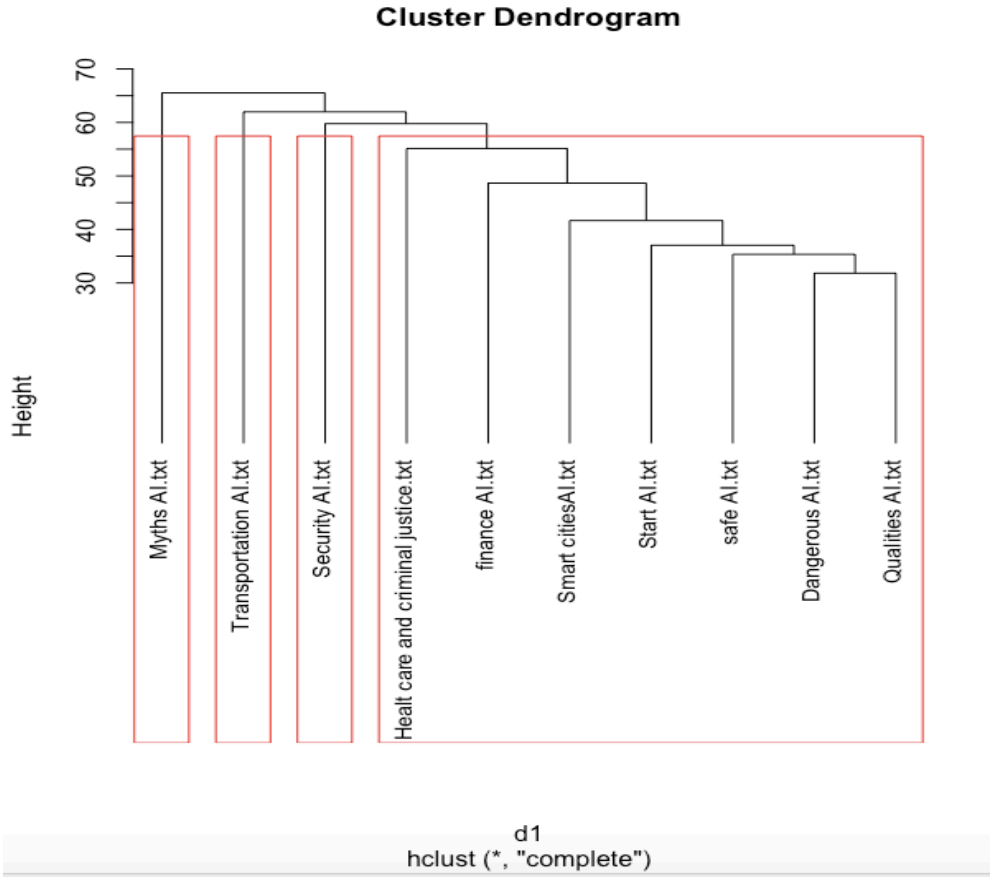
Metoda grupowania, w której tworzone klastry pozostają w pewnej hierarchii, z której wyszczególnić można nadrzędne grupy oraz ich elementy, czyli klastry niższego rzędu.

Dla dokumentów:

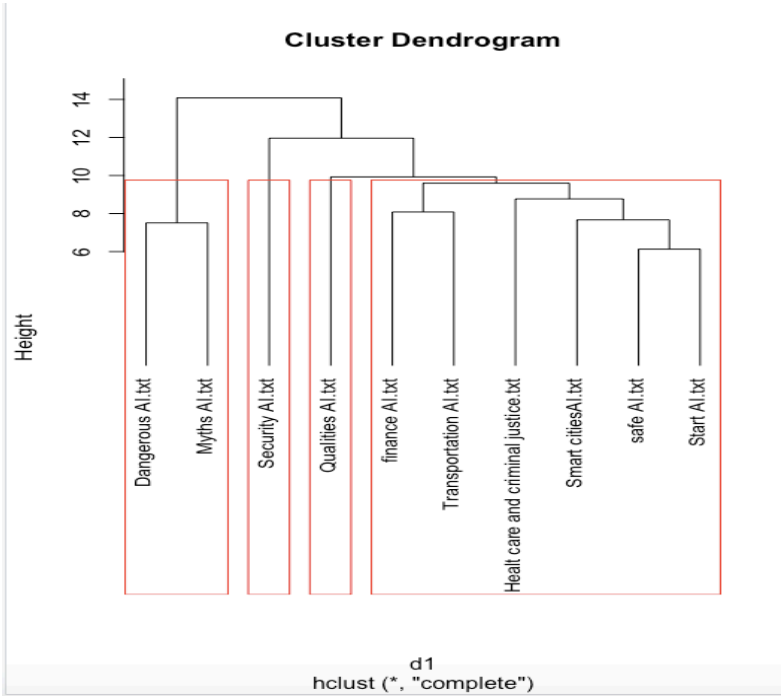
- **DTM**



- **TF-IDF**

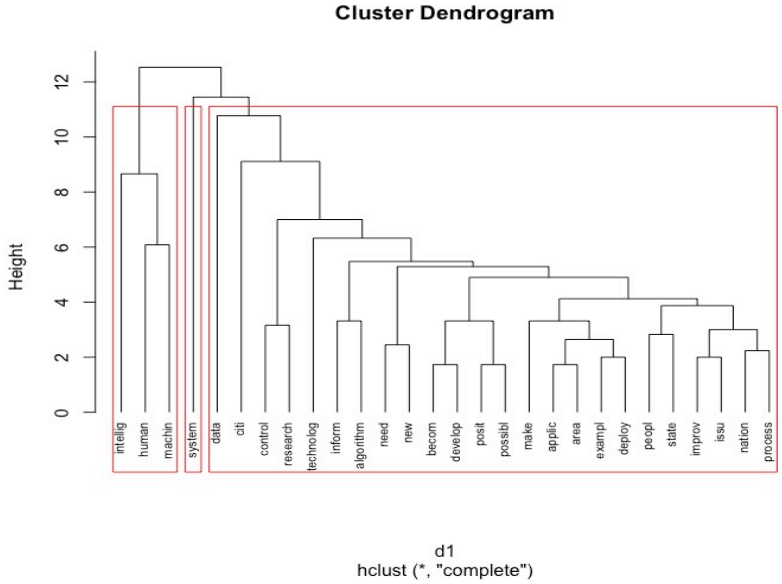


- **TF-IDF reducing sparsity**

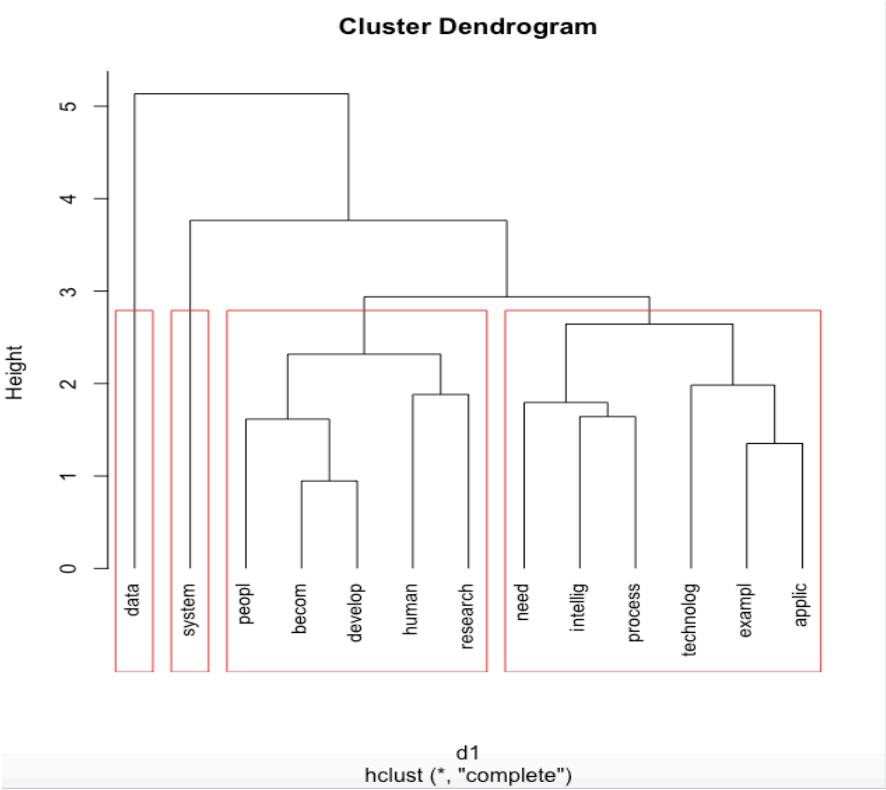


Dla słów:

- **TDM**



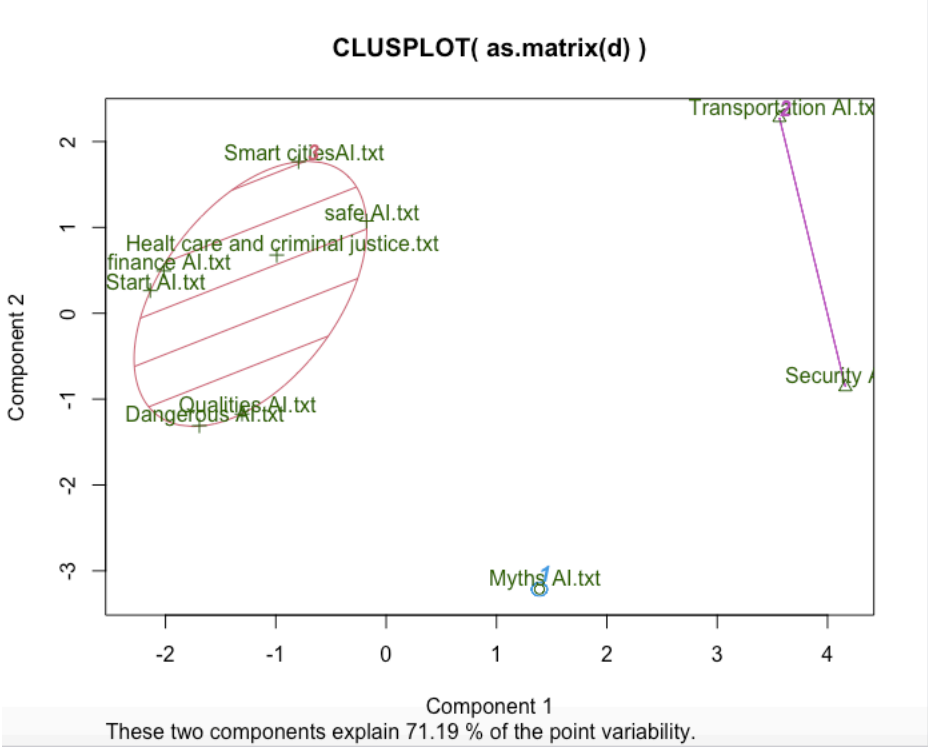
- **T(TF-IDF)**



7. K-MEANS CLUSTERING

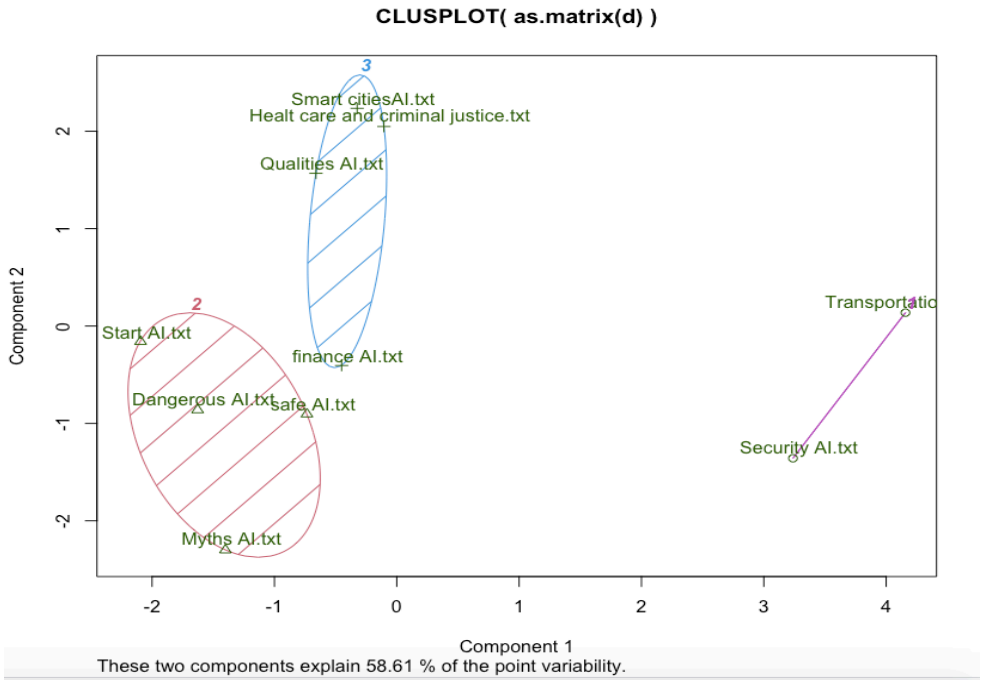
Dla dokumentów:

- **DTM**

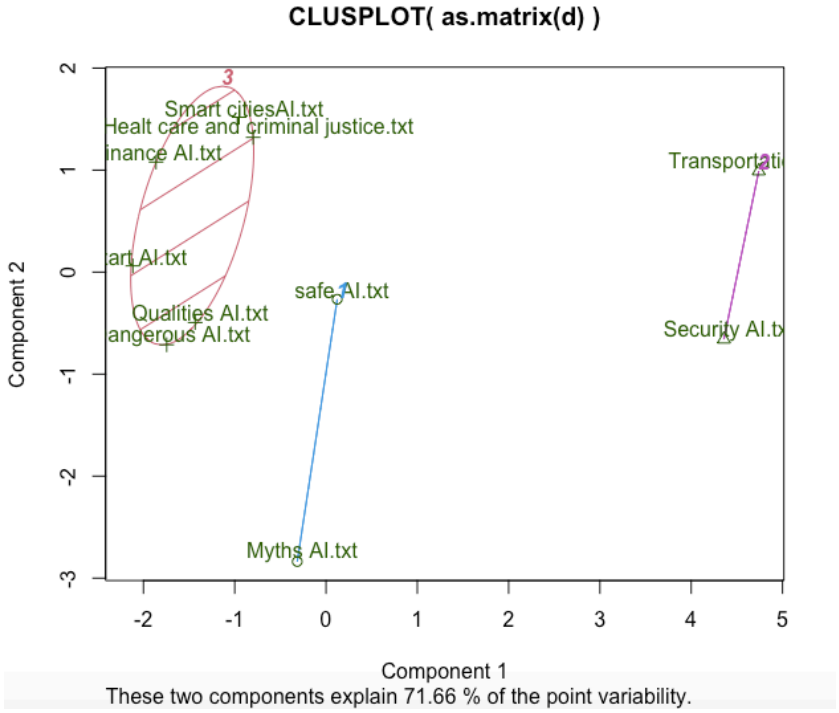


Klasteryzacja DTM

- **TF-IDF**

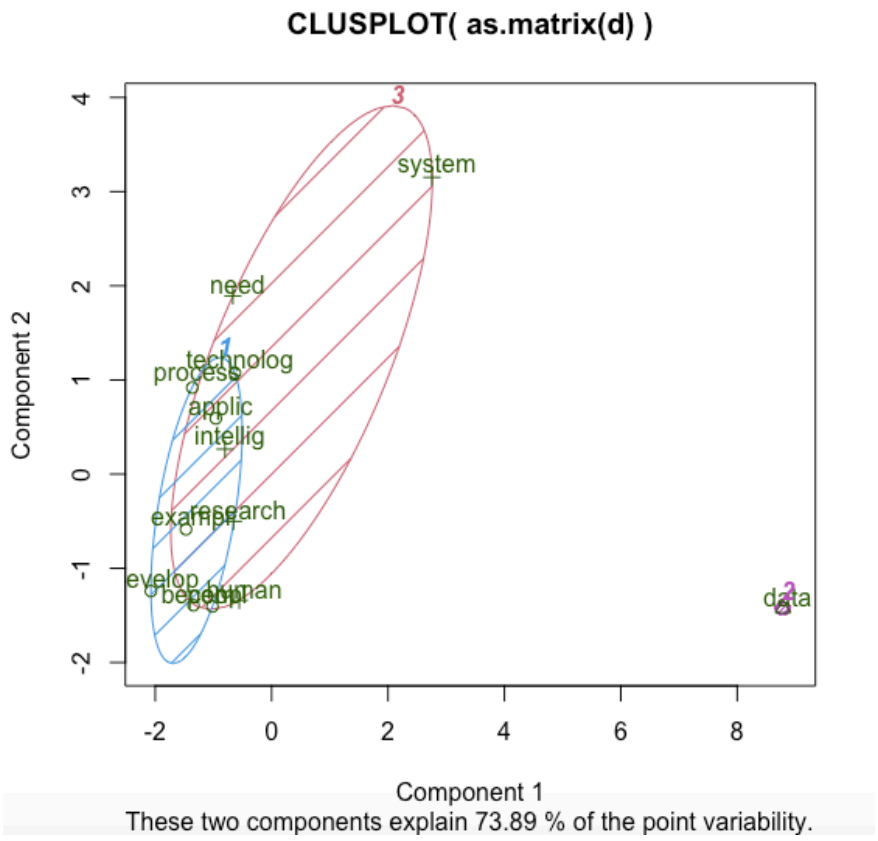


- **TF-IDF sparsity**

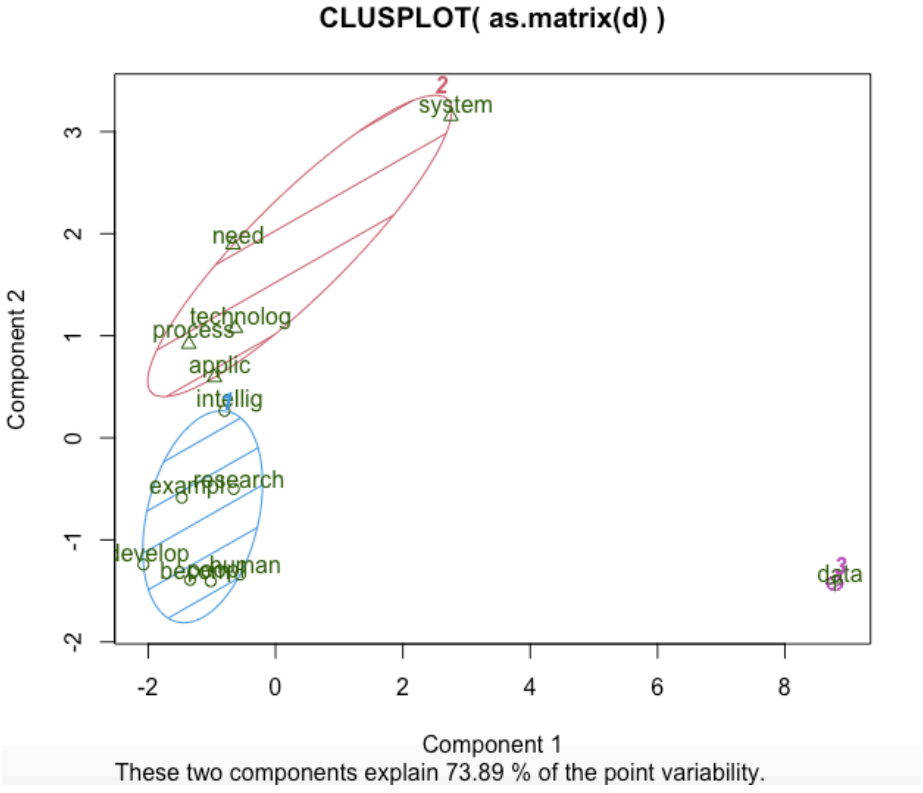


Dla słów:

- **TDM**


CLUSPLOT(as.matrix(d))
system
need
technolog
process
applic
intellig
research
data
Component 2
Component 1
These two components explain 73.89 % of the point variability.

- **T(TF-IDF)**


CLUSPLOT(as.matrix(d))
system
need
technolog
process
applic
intellig
research
data
Component 2
Component 1
These two components explain 73.89 % of the point variability.

Method	Source	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Rating
Documents						
K-means	DTM	Mity oparte na sztucznej inteligencji	Obszary ogólne i Społeczne kierunki rozwoju AI	Technologiczne rozwiązania AI w transporcie i bezpieczeństwie	-	3
K-means	TF-IDF	Technologiczne rozwiązania AI w transporcie i bezpieczeństwie	Ogólny obszary zainteresowań AI	Społeczne kierunki rozwoju AI	-	1
K-means	TF-IDF-S	Mity i ich wyjaśnienie	Technologiczne rozwiązania AI w transporcie i bezpieczeństwie	Społeczne kierunki rozwoju AI	-	2
Hierarchical Clustering	DTM	Mity oparte na sztucznej inteligencji	Rozwój transportu	Rozwój bezpieczeństwa narodowego	Obszary ogólne i społeczne kierunki rozwoju AI	2
Hierarchical Clustering	TF-IDF D	Mity oparte na sztucznej inteligencji	Rozwój transportu	Rozwój bezpieczeństwa narodowego	Obszary ogólne i społeczne kierunki rozwoju AI	2
Hierarchical Clustering	TF-IDF-S	Mity i zagrożenia AI	Rozwój bezpieczeństwa narodowego	Możliwości rozwijania AI	Technologiczne i Społeczne kierunki rozwoju AI	1
Terms						
K-means	TDM-S	Słowa występujące najczęściej	Dane	Słownictwo z zakresu technologii	-	1
K-means	T(TF-IDF-S)	Słowa występujące najczęściej	Słownictwo z zakresu technologii	Dane	-	1
Hierarchical Clustering	TDM-S	Słowa występujące najczęściej definiujące tematykę korpusu	System	Słownictwo z zakresu technologii kształtujące pod tematy korpusu	-	1

Hierarchical Clustering	T(TF-IDF-S)	Dane	System	Słowa związane z społeczeństwem i jego rozwojem	Słownictwo z zakresu technologii	2

- **Wnioski z analizy klasteryzacji - dokumenty**

W niniejszej analizie przypadku dokumentów klasteryzacja hierarchiczna jaki i ta wykorzystująca metodę k-means wykazały bardzo podobne wyniki. Niemal w każdym przypadku dokumenty zostały pogrupowane w odpowiedni sposób. Lecz, pierwszy z wyników w metodzie k-means oparty na DTM pogrupował dokumenty w sposób zbyt ogólny przez co wyniki tego przypadku stają się nie jasne. Jednak kolejne metody zadziałały poprawnie grupując kolejne dokumenty w sposób bardziej szczegółowy. Co do metody klasteryzacji hierarchicznej wyniki wykazały odpowiednie i sensowne pogrupowanie poszczególnych dokumentów. W każdej z przeprowadzonych metod pojawiły się dwie najlepiej grupujące, a były nimi:

- TF-IDF – k-means
- TF-IDF-S – klasteryzacja hierarchiczna

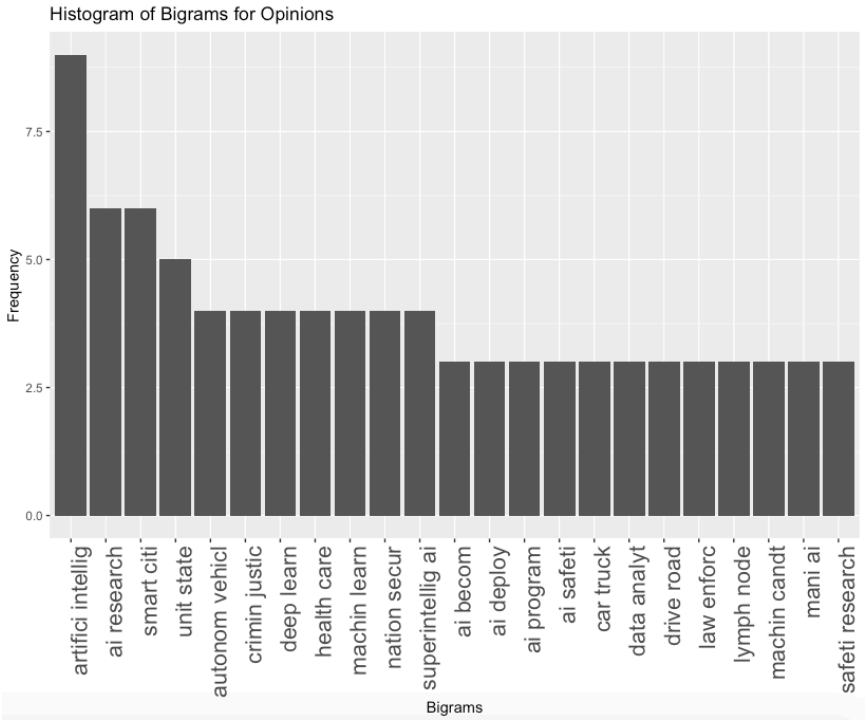
Jednak z tych dwóch ta druga okazała się tą najbardziej szczegółową, grupując dokumenty z najlepszym rezultatem.

- **Wnioski z analizy klasteryzacji - dokumenty**

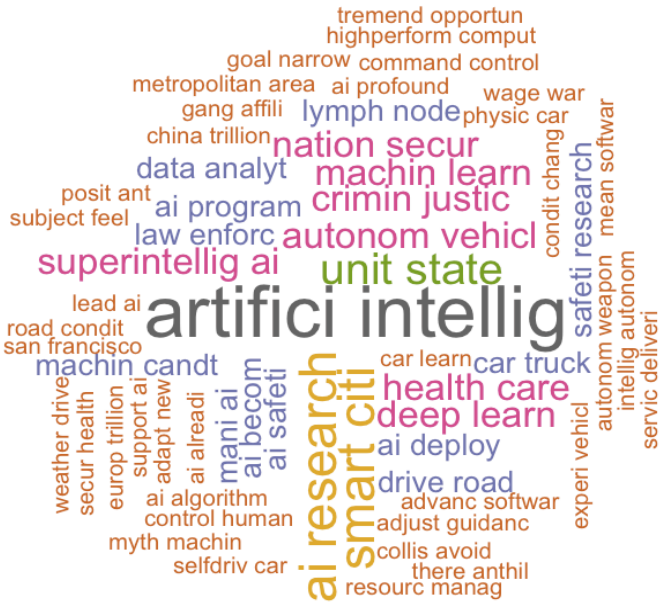
Jak przy dokumentach analiza klasteryzacji słów wykazały podobne rezultaty oraz cechy w odniesieniu pogrupowania słów według określonego schematu na słowa definiujące ogólny temat korpusu oraz na tę objaśniające jego podtematy. Wyjątkiem była jedynie klasteryzacja hierarchiczna obliczana na podstawie T(TF-IDF-S), która przedstawiła nie jasne wyniki w odniesieniu do reszty wykorzystanych metod.

8. BIOGRAMS

- **Zipf's**

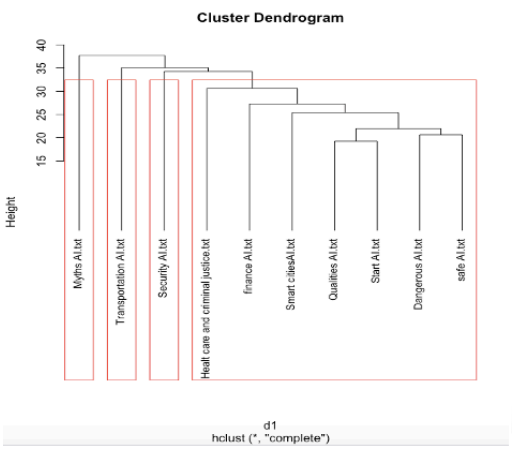


- **WordClouds**

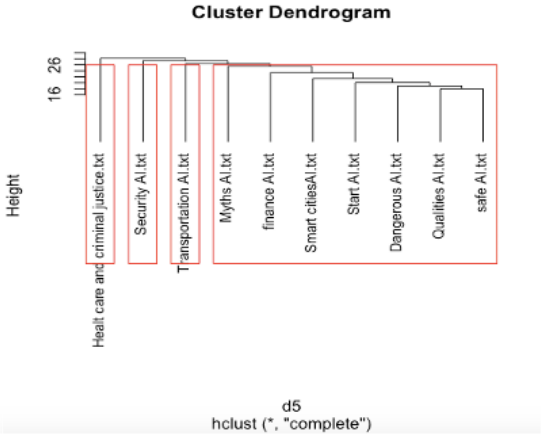


9. DTM vs DTM-n

- **Klasteryzacja hierarchiczna**



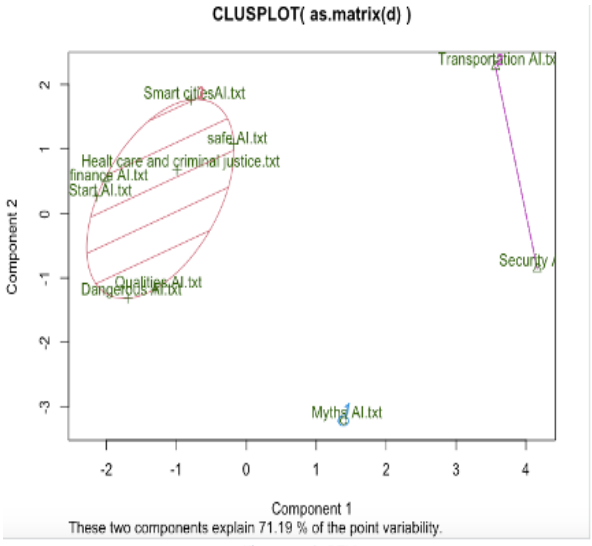
Rysunek 2 Klasteryzacja DTM



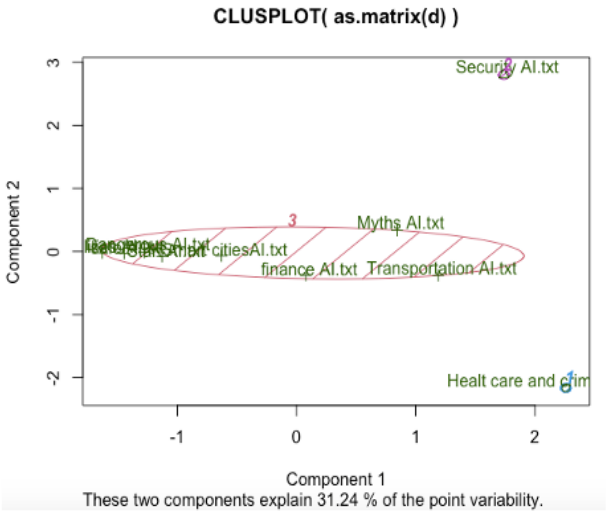
Rysunek 3 Klasteryzacja DTM_n

W przypadku badanego korpusu jaśniejsze wyniki grupowania pojawiają się przy klasteryzacji hierarchicznej na podstawie macierzy DTM, charakteryzujące się lepszym połączeniem tematycznym w przedstawionych grupach.

- **Klasteryzacja metodą k-means**



Klasteryzacja DTM



Klasteryzacja DTM_n

Taka sama sytuacja pojawia się także w tym przypadku. Dla badanego korpusu efektywniejsze jest korzystanie z pierwszej metody niż z biogramów.

10. Wnioski

Analiza korpusu dokumentów i zawartych w nich pod tematów dobiegła końca. Szczegółowe poznanie tematyki i słów kluczowych obecnych w pod tematach, a także połączenie określonych tematów w grupy, pozwoliły na szersze poznanie tematyki badanego korpusu. Już nie tylko ogólnej, lecz także problematyki szczegółowej. Z czego wynika, że badany korpus zawiera trzy podstawowe grupy tematyczne, ogólną zawierającą pobieżną wiedzę na temat AI, społeczną opartą na rozwoju systemów związanych z życiem codziennym, a także grupę trzecią skupiającą się na konkretnych technologiach wykorzystywanych w transporcie i w obronności.