# RAPORT DATA MINING – PODOBIEŃSTWO COSINUSOWE W ŚRODOWISKU R

EMIL FILIPOWICZ

## 1. ZAŁADOWANE PAKIETY

- library("topicmodels")
- library("igraph")

## 2. ZBIÓR DANYCH

Proces przedstawiony w raporcie powstał przy pomocy takich danych, jak:

- korpus – zawierający 10 dokumentów tekstowych
- macierz DTM przedstawiająca rozkład słów w poszczególnych dokumentach zawartych w korpusie

## 3. PODOBIEŃSTWO COSINUSOWE – DOKUMENTY

### 3.1 BUDOWANIE MACIERZY – ODLEGŁOŚCI COSINUSOWYCH

| | Dangerous AI.txt | finance AI.txt | Healt care and criminal justice.txt | Myths AI.txt | |
|---|---|---|---|---|---|
| finance AI.txt | 0.14119562 | | | | |
| Healt care and criminal justice.txt | 0.14782532 | 0.24279690 | | | |
| Myths AI.txt | 0.36618801 | 0.11494648 | | 0.10838243 | |
| Qualities AI.txt | 0.16450006 | 0.18116817 | | 0.22113522 | 0.15852092 |
| safe AI.txt | 0.16805083 | 0.08158032 | | 0.12630509 | 0.22094498 |
| Security AI.txt | 0.20427894 | 0.18993794 | | 0.17914981 | 0.13884470 |
| Smart citiesAI.txt | 0.03669879 | 0.15503673 | | 0.16410658 | 0.03112110 |
| Start AI.txt | 0.11901266 | 0.20034784 | | 0.22005544 | 0.11503216 |
| Transportation AI.txt | 0.10816079 | 0.15623083 | | 0.15526725 | 0.09554365 |

| | Qualities AI.txt | safe AI.txt | Security AI.txt | Smart citiesAI.txt | Start AI.txt |
|---|---|---|---|---|---|
| finance AI.txt | | | | | |
| Healt care and criminal justice.txt | | | | | |
| Myths AI.txt | | | | | |
| Qualities AI.txt | | | | | |
| safe AI.txt | 0.07328973 | | | | |
| Security AI.txt | 0.29903694 | 0.11840764 | | | |
| Smart citiesAI.txt | 0.15394232 | 0.06941756 | 0.15751683 | | |
| Start AI.txt | 0.29934767 | 0.11486973 | 0.20936858 | 0.16117930 | |
| Transportation AI.txt | 0.14233966 | 0.06277237 | 0.17886997 | 0.12189275 | 0.07993410 |

Utworzona macierz przedstawia podobieństwo między każdym dokumentem zawartym w korpusie. Jednak nie wszystkie uzyskane wyniki są odpowiednio istotne, aby brać je pod uwagę podczas badania. Dlatego też macierz została wyczyszczona z danych nie istotnych dla dalszej analizy z progiem odległości przyjętym na poziomie min_cos = 0,15 (każdej mniejszej wartości od docelowej została przypisana wartość 0).

| | Dangerous AI.txt | finance AI.txt | Healt care and criminal justice.txt | | |
|---|---|---|---|---|---|
| finance AI.txt | 0.000 | | | | |
| Healt care and criminal justice.txt | 0.000 | 0.243 | | | |
| Myths AI.txt | 0.366 | 0.000 | 0.000 | | |
| Qualities AI.txt | 0.165 | 0.181 | 0.221 | | |
| safe AI.txt | 0.168 | 0.000 | 0.000 | | |
| Security AI.txt | 0.204 | 0.190 | 0.179 | | |
| Smart citiesAI.txt | 0.000 | 0.155 | 0.164 | | |
| Start AI.txt | 0.000 | 0.200 | 0.220 | | |
| Transportation AI.txt | 0.000 | 0.156 | 0.155 | | |

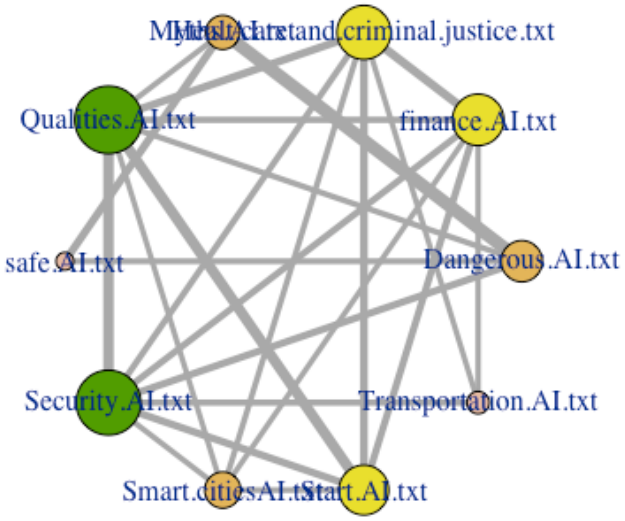
| | Myths AI.txt | Qualities AI.txt | safe AI.txt | Security AI.txt | Smart citiesAI.txt |
|---|---|---|---|---|---|
| finance AI.txt | | | | | |
| Healt care and criminal justice.txt | | | | | |
| Myths AI.txt | | | | | |
| Qualities AI.txt | 0.159 | | | | |
| safe AI.txt | 0.221 | 0.000 | | | |
| Security AI.txt | 0.000 | 0.299 | 0.000 | | |
| Smart citiesAI.txt | 0.000 | 0.154 | 0.000 | 0.158 | |
| Start AI.txt | 0.000 | 0.299 | 0.000 | 0.209 | 0.161 |
| Transportation AI.txt | 0.000 | 0.000 | 0.000 | 0.179 | 0.000 |

| | Start AI.txt |
|---|---|
| finance AI.txt | |
| Healt care and criminal justice.txt | |
| Myths AI.txt | |
| Qualities AI.txt | |
| safe AI.txt | |
| Security AI.txt | |
| Smart citiesAI.txt | |
| Start AI.txt | |
| Transportation AI.txt | 0.000 |

Dane z kroku pierwszego uległy drastycznej zmianie, niemal połowa z wcześniej wykazanych w macierzy odległości została wyzerowana. W macierzy pozostały odległości istotne dla analizy cosinusowej, a ich liczba wynosi 23. W drugim kroku widoczne jest, że największym podobieństwem charakteryzują się dokumenty „Myths AI.txt" oraz „Dangerous AI.txt" z wartością odległości równej 0,366, a najmniejszym „Smart cities AI.txt" i „Qualities AI.txt".

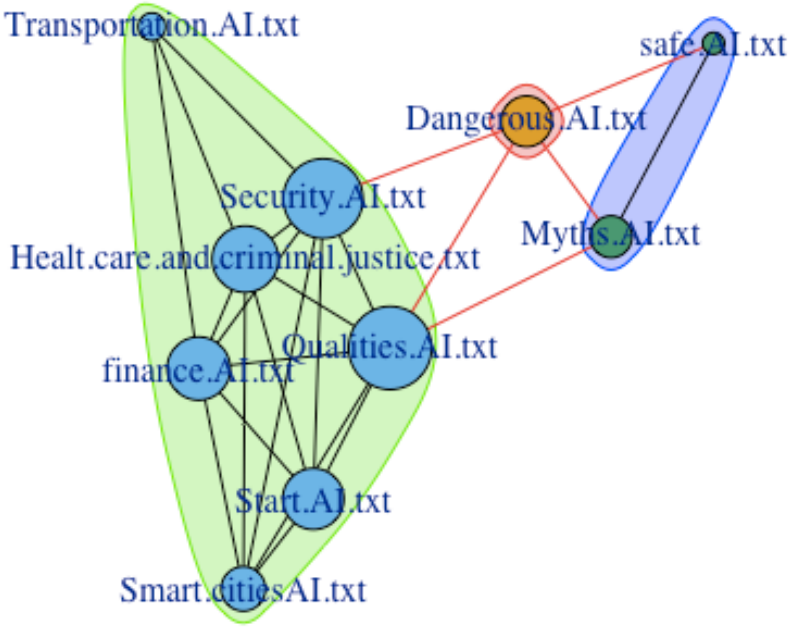
## 3.2 PRZEDSTAWIENIE GRAFICZNE PODOBIEŃSTWA COSINUSOWEGO DOKUMENTÓW NA TLE CAŁEGO KORPUSU.

Do przedstawienia podobieństwa został użyty wykres lay_2, który w najlepszy sposób przedstawia moc powiązań między poszczególnymi dokumentami, gdzie czym grubsza linia połączenia tym większe podobieństwo tematyczne dokumentów.




Dokumenty zostały podzielone na trzy grupy. W kolorze zielonym widnieją dokumenty, których podobieństwo względem całości jest najsilniejsze, następne teksty w hierarchii zaznaczone zostały na żółto oraz te, których podobieństwo jest najmniejsze na pomarańczowo.

## 3.3 COMMUNITY DETECTION – ALGORYTMY WYKRYWANIA SPOŁECZNOŚCI.



W tym przypadku analizowany wykres to algorytm 1. Przedstawia on 3 grupy relacyjne dokumentów. Widoczne jest, że grupa zielono-niebieska zawiera w sobie najwięcej dokumentów, czyli można stwierdzić, że te dokumenty wykazują między sobą większe podobieństwo od reszty grup. Druga grupa zawiera jedynie 2 dokumenty. Na wykresie widnieje także jeden dokument tworzący osobny obszar, co pokazuję jego odosobnienie relacyjne od reszty dokumentów w korpusie.

Podział grupowy dokumentów:

| | |
|---|---|
| Dangerous.AI.txt | finance.AI.txt |
| 1 | 2 |
| Healt.care.and.criminal.justice.txt | Myths.AI.txt |
| 2 | 3 |
| Qualities.AI.txt | safe.AI.txt |
| 2 | 3 |
| Security.AI.txt | Smart.citiesAI.txt |
| 2 | 2 |
| Start.AI.txt | Transportation.AI.txt |
| 2 | 2 |

# 4. PODOBIEŃSTWO COSINUSOWE – SŁOWA

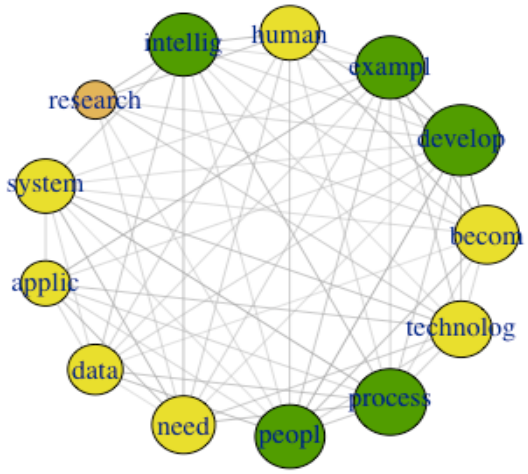
## 4.1 BUDOWANIE MACIERZY – ODLEGŁOŚCI COSINUSOWYCH

| | becom | develop | exampl | human | intellig | research | system | applic | data | need | peopl |
|---|---|---|---|---|---|---|---|---|---|---|---|
| develop | 0.915 | | | | | | | | | | |
| exampl | 0.751 | 0.791 | | | | | | | | | |
| human | 0.744 | 0.744 | 0.489 | | | | | | | | |
| intellig | 0.488 | 0.474 | 0.396 | 0.751 | | | | | | | |
| research | 0.485 | 0.471 | 0.000 | 0.676 | 0.702 | | | | | | |
| system | 0.413 | 0.379 | 0.415 | 0.000 | 0.588 | 0.000 | | | | | |
| applic | 0.000 | 0.441 | 0.759 | 0.000 | 0.000 | 0.000 | 0.505 | | | | |
| data | 0.370 | 0.472 | 0.419 | 0.490 | 0.584 | 0.000 | 0.432 | 0.459 | | | |
| need | 0.000 | 0.393 | 0.485 | 0.508 | 0.719 | 0.452 | 0.749 | 0.724 | 0.358 | | |
| peopl | 0.706 | 0.857 | 0.626 | 0.547 | 0.418 | 0.404 | 0.367 | 0.454 | 0.740 | 0.000 | |
| process | 0.454 | 0.567 | 0.483 | 0.434 | 0.717 | 0.000 | 0.833 | 0.500 | 0.663 | 0.724 | 0.648 |
| technolog | 0.465 | 0.521 | 0.609 | 0.000 | 0.381 | 0.590 | 0.683 | 0.630 | 0.000 | 0.615 | 0.536 |
| | process | | | | | | | | | | |
| develop | | | | | | | | | | | |
| exampl | | | | | | | | | | | |
| human | | | | | | | | | | | |
| intellig | | | | | | | | | | | |
| research | | | | | | | | | | | |
| system | | | | | | | | | | | |
| applic | | | | | | | | | | | |
| data | | | | | | | | | | | |
| need | | | | | | | | | | | |
| peopl | | | | | | | | | | | |
| process | | | | | | | | | | | |
| technolog | 0.552 | | | | | | | | | | |

Macierz przedstawia podobieństwo występowania słów w poszczególnych dokumentach. Macierz ta została już wyczyszczona, a jej cechy są analogiczne do cech macierzy stworzonej przy analizie dokumentów.

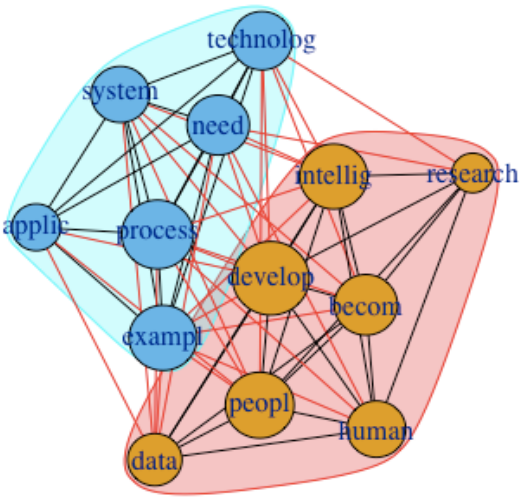
## 4.2 PRZEDSTAWIENIE GRAFICZNE PODOBIEŃSTWA COSINUSOWEGO DOKUMENTÓW NA TLE CAŁEGO KORPUSU.

Do wyjaśnienia tego przypadku został użyty podobnie jak w poprzednim wypadku wykres lay_2, który ukazuję poszczególne powiązania słów w najbardziej czytelny sposób.




Analogicznie jak przy analizie dokumentów, utworzone zostały 3 grupy powiązań słów. Grupa zielona charakteryzująca się najsilniejszymi podobieństwami względem całego korpusu, żółta – pośrednia oraz pomarańczowa cechująca się najsłabszym powiązaniem między innymi słowami w dokumentach korpusu.

### 4.3 COMMUNITY DETECTION




Aby zbadać community detection został wybrany algorytm 3. Jako jedyny z trzech podzielił słowa na dwie grupy, kiedy pozostałe dwa tworzyły jedną.

## 5. PODSUMOWANIE PODOBIEŃSTWA COSINUSOWEGO NA PODSTAWIE MACIERZY DTM ORAZ TDM.

| Metoda | Źródło | Community 1 | Community 2 | Community 3 |
|---|---|---|---|---|
| Community detection | Podobieństwo cosinusowe - DTM | Ryzyko związane z korzystaniem AI | Rozwój oraz zastosowanie AI w różnych sektorach życia | Mity związane z AI oraz ich wytłumaczenie |
| Community detection | Podobieństwo cosinusowe - TDM | Słowa związane z AI w sferze technologicznej | Słowa związane z AI w sferze rozwoju i społeczeństwa | - |

- Dokumenty przy użyciu metody Community detection na podstawie podobieństwa cosinusowego wykazało dobre i sensowne pogrupowanie dokumentów na grupy tematyczne.
- Słowa pogrupowane za pomocą tych metod zostały podzielone na dwie społeczności - pierwszą technologiczną oraz drugą społeczną. Jest to odpowiedni podział, jednak dla doprecyzowania powinny powstać jeszcze jedna grupa po rozbiciu pierwszej aby rozdzielić dokumenty z treścią bardziej ogólną od tych z szczegółową wiedzą technologiczną z obszaru AI.

## 6. TOPIC MODELING

- podział słów na 2 tematy:

```
      Topic 1    Topic 2
 [1,] "human"    "system"
 [2,] "intellig" "data"
 [3,] "goal"     "technolog"
 [4,] "research" "citi"
 [5,] "machin"   "advanc"
 [6,] "car"      "autonom"
 [7,] "mani"     "learn"
 [8,] "control"  "vehicl"
 [9,] "develop"  "exampl"
[10,] "robot"    "deploy"
```

- podział słów na 3 tematy:

```
      Topic 1     Topic 2    Topic 3
 [1,] "system"    "human"    "data"
 [2,] "technolog" "intellig" "citi"
 [3,] "learn"     "goal"     "becom"
 [4,] "advanc"    "machin"   "exampl"
 [5,] "autonom"   "research" "deploy"
 [6,] "vehicl"    "mani"     "state"
 [7,] "artifici"  "control"  "imag"
 [8,] "speed"     "robot"    "use"
 [9,] "new"       "risk"     "develop"
[10,] "car"       "posit"    "applic"
```

- podział słów na 4 tematy:

```
      Topic 1     Topic 2     Topic 3    Topic 4
 [1,] "technolog" "citi"      "human"    "intellig"
 [2,] "car"       "becom"     "goal"     "data"
 [3,] "system"    "exampl"    "research" "system"
 [4,] "need"      "deploy"    "mani"     "speed"
 [5,] "learn"     "state"     "machin"   "process"
 [6,] "advanc"    "imag"      "robot"    "artifici"
 [7,] "autonom"   "use"       "control"  "human"
 [8,] "inform"    "peopl"     "risk"     "capabl"
 [9,] "vehicl"    "algorithm" "posit"    "decis"
[10,] "invest"    "applic"    "ant"      "nation"
```

Po sprawdzeniu 3 alternatyw podziału tematycznego, w niniejszej analizie wybrana została opcja grupująca na trzy tematy.

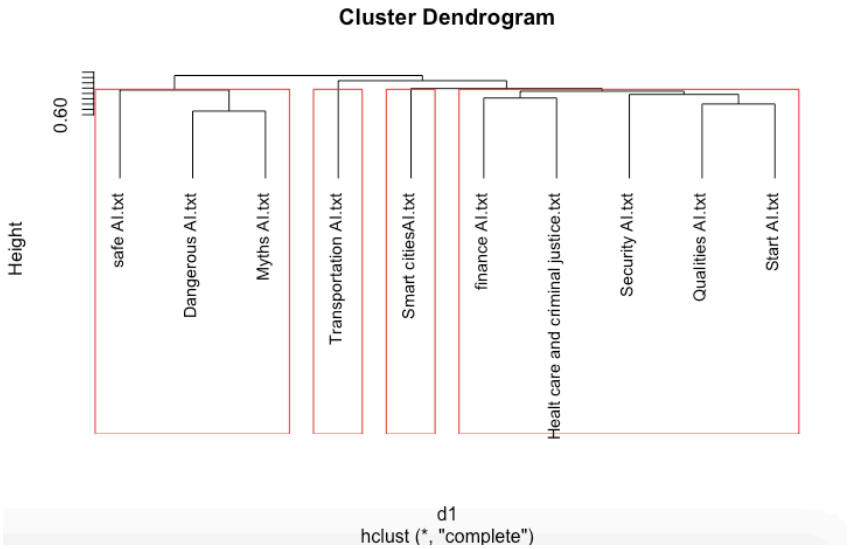
```
           V1        V2        V3                                    [,1]
1  0.20570571 0.5525526 0.24174174 Dangerous AI.txt                     2
2  0.31948566 0.1770524 0.50346192 finance AI.txt                       3
3  0.24183007 0.1879085 0.57026144 Healt care and criminal justice.txt  3
4  0.07859848 0.8257576 0.09564394 Myths AI.txt                         2
5  0.38568935 0.2652705 0.34904014 Qualities AI.txt                     1
6  0.18954248 0.5669935 0.24346405 safe AI.txt                          2
7  0.60164609 0.1917695 0.20658436 Security AI.txt                      1
8  0.25157233 0.1836478 0.56477987 Smart citiesAI.txt                   3
9  0.23291492 0.2705718 0.49651325 Start AI.txt                         3
10 0.70983936 0.1495984 0.14056225 Transportation AI.txt                1
```

W czerwonych ramkach zaznaczone zostało prawdopodobieństwo przynależności określonego dokumentu do określonego tematu.

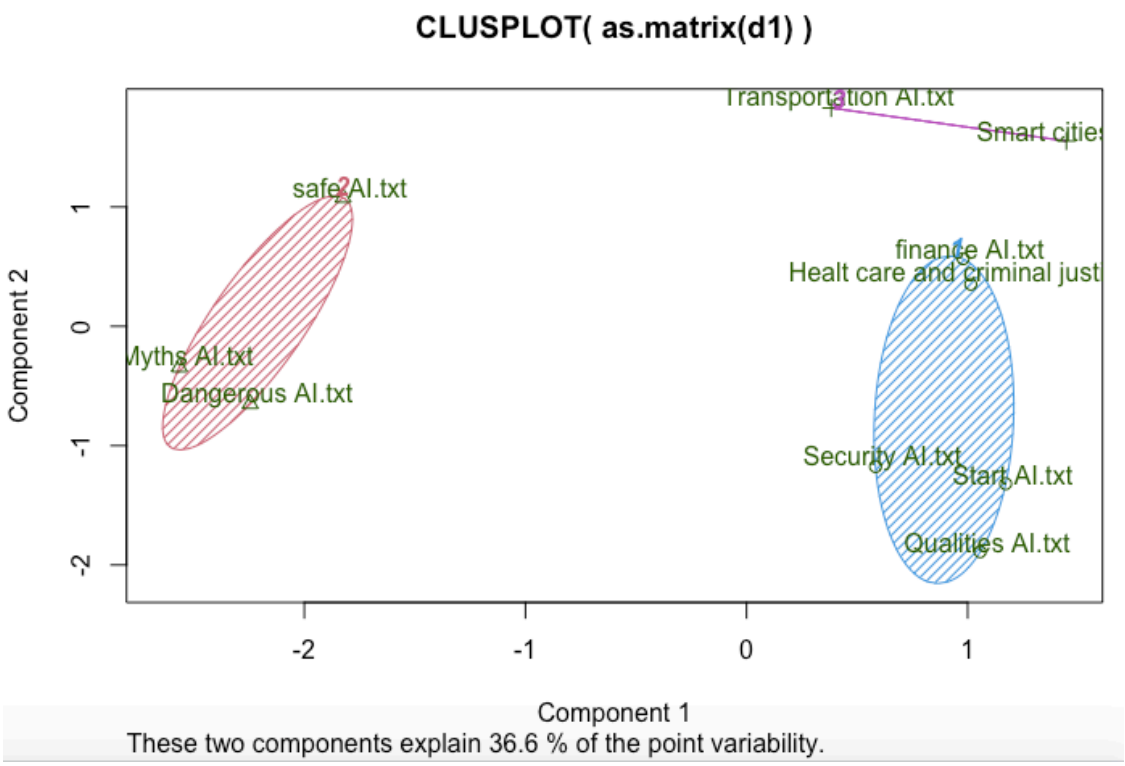
| # | Topic 1 | Topic description | Example | % of Documents, mostly associated with This Topic |
|---|---|---|---|---|
| 1 | Rozwój technologii oraz możliwości idące za rozwojem AI | Opis technologii AI w kontekście rozwoju transportu, obronności narodowej oraz cechy i możliwości sztucznej inteligencji | "The big data analytics associated with AI will profoundly affect intelligence analysis, as massive amounts of data are sifted in near real time—if not eventually in real time—thereby providing commanders and their staffs a level of intelligence analysis and productivity heretofore unseen." Security.txt | 30% |
| 2 | Zagrożenia, ryzyko i mity powstałe wraz z rozwojem AI | Obawy społeczne związane z wprowadzaniem sztucznej inteligencji w życiu codziennym, narastające mity oraz wytłumaczenie realnych zagrożeń związanych ze sztuczną inteligencją | „Typically, these articles are accompanied by an evil-looking robot carrying a weapon, and they suggest we should worry about robots rising up and killing us because they've become conscious and/or evil." **Myths.txt** | 30% |
| 3 | Rozwój AI w kontekście użyteczności społecznej | Sztuczna inteligencja rozwój i wprowadzanie jej w różnych sektorach gospodarki oraz sfery publicznej. | "AI tools are helping designers improve computational sophistication in health care. For example, Merantix is a German company that applies deep learning to medical issues. It has an application in medical imaging that "detects lymph nodes in the human body in Computer Tomography (CT) images" **Healt care and criminal justice.txt** | 40% |

## 7. K-MEANS I HIERARCHICAL CLUSTERING/COSINE SIMILARITY

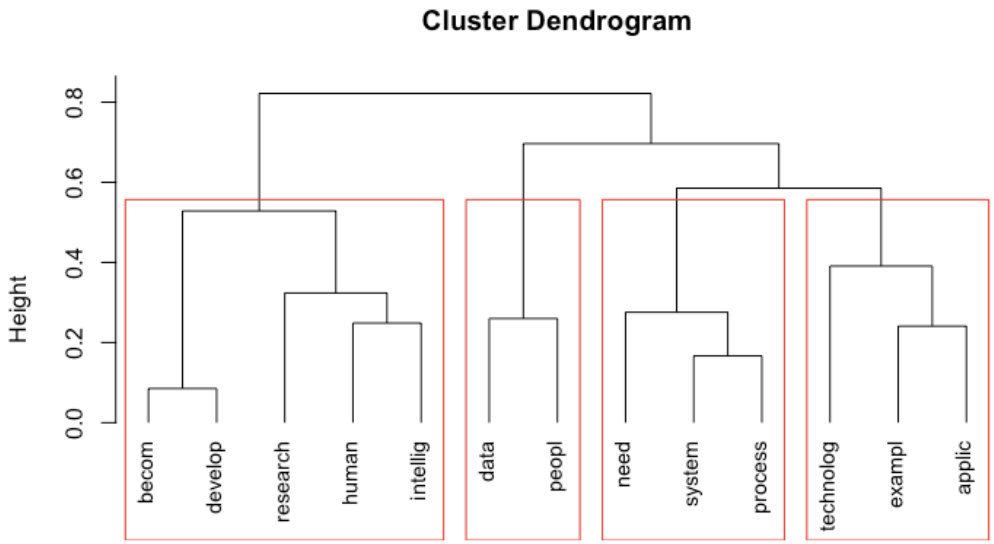
- **Dokumenty (DTM)**
  - Hierarchical Clustering



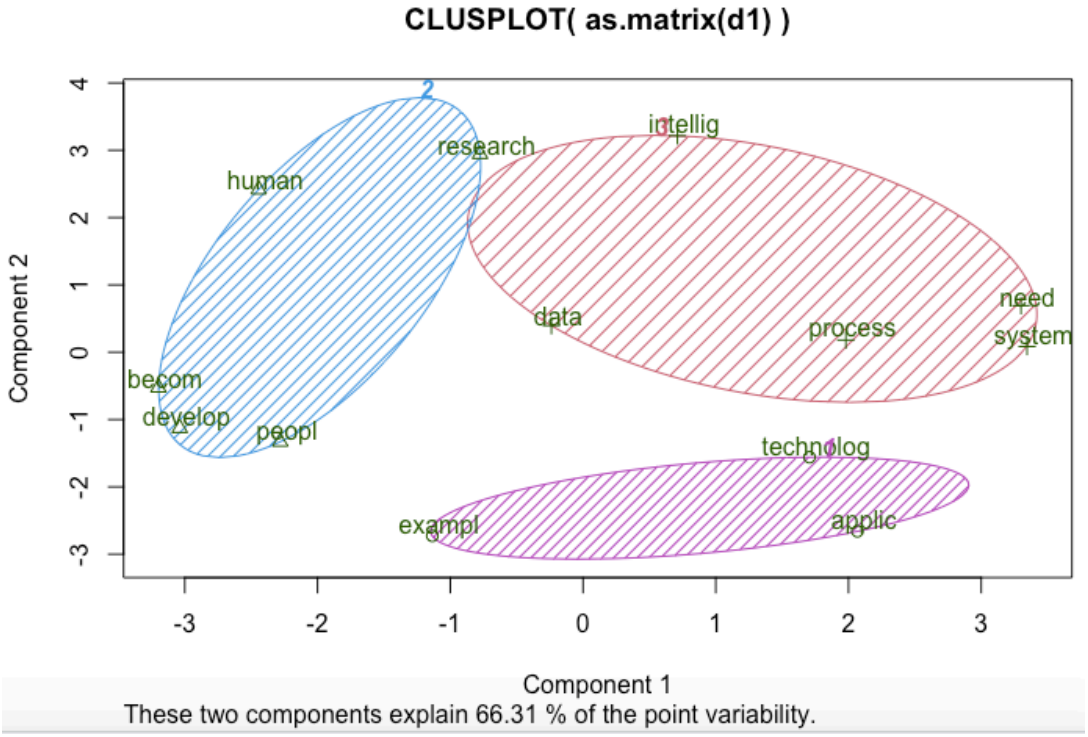

  - K-means



- **Słowa (TDM)**
  - Hierarchical Clustering




  - K-means



## 8. PODSUMOWANIE WYNIKÓW GRUPOWANIA

W tabeli zostały zebrane oraz podsumowane wszystkie metody grupowania dokumentów oraz słów w badanym korpusie 10 dokumentów powiązanych z obszarem sztucznej inteligencji.

| Metoda | Źródło | 1 | 2 | 3 | 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 | - | 5 |
| Hierarchical Clustering | DTM | Mity oparte na sztucznej inteligencji | Rozwój AI w transporcie | Rozwój bezpieczeństwa narodowego | Obszary ogólne i społeczne kierunki rozwoju AI | 6 |
| K-means | Cosine similarity / DTM | Obszary ogólne, cechy oraz ochrona AI | Rozwój AI w kontekście użyteczności społecznej | Zagrożenia i mity powstałe wraz z rozwojem AI | Kontrola i rozwój AI w bezpiecznym kierunku | 2 |
| Hierarchical Clustering | Cosine similarity / DTM | Zagrożenia, bezpieczeństwo oraz mity powstałe wraz z rozwojem AI | Rozwój AI w transporcie | AI w rozwoju inteligentnych miast | Rozwój AI w kontekście użyteczności społecznej | 4 |
| Community detection | Cosine similarity/ DTM | Ryzyko związane z korzystaniem AI | Rozwój oraz zastosowanie AI w różnych sektorach życia | Mity związane z AI oraz ich wytłumaczenie | - | 3 |
| Topic modelling | DTM | Rozwój technologii oraz możliwości idące za rozwojem AI | Zagrożenia, ryzyko i mity powstałe wraz z rozwojem AI | Rozwój AI w kontekście użyteczności społecznej | - | 1 |

| Metoda | Źródło | 1 | 2 | 3 | 4 | Rating |
|---|---|---|---|---|---|---|
| Terms | | | | | | |
| K-means | TDM | Słowa występujące najczęściej | Dane | Słownictwo z zakresu technologii | - | 5 |
| Hierarchical Clustering | TDM | Słowa występujące najczęściej definiujące tematykę korpusu | System | Słownictwo z zakresu technologii kształtujące pod tematy korpusu | - | 4 |
| K-means | Cosine similarity / TDM | Słownictwo z zakresu technologii | Słowa związane z badaniami nad AI | Procesy zachodzące oraz systemy stosowane w obszarach AI | | 1 |
| Hierarchical Clustering | Cosine similarity / TDM | Słowa związane z badaniami nad AI | Słowa związane z danymi oraz cechami ludzkimi | Procesy zachodzące w systemach AI | Słownictwo z zakresu technologii | 2 |
| Community detection | Cosine similarity/ DTM | Słowa związane z AI w sferze technologicznej | Słowa związane z AI w sferze rozwoju i społeczeństwa | - | - | 3 |