Zadanie 4.1, 228884, Mateusz Guściora-Sprawozdanie

4.1.1

a) i b) Udało zainstalować rozszerzenia do programu RapidMiner włącznie z Meaning Cloud.

c) Przykładowe procesy operacji na danych tekstowych

Zapoznajemy się z procesem Web analytics, który przewiduje „high low value” (true/false), usuwa atrybuty mocno ze sobą skorelowane oraz dodaje wagi(rosnąca korelacja atrybutów). W wynikach widzimy bardzo dużą dokładność predykcji 99,29%. Zapoznajemy się także z danymi, w danych obserwujemy takie zmienne jak visit time, period. W tym procesie użytwaliśmy algorytmu klasyfikacji i regresji Genaralize lienear model.

Kolejny przykładowy proces opiera się na danych opinii historycznych i przewiduje czy są pozytywne i negatywne. Dane składają się z tekstu i oceny opinii (positive i negative). Proces ustala zmienna sentyment jako label. Następnie podproces prowadzi do walidacji krzyżowej oraz do zapisania tekstu w innym pliku i dodania własny tekst. W walidacji używamy algorytmu używamy algorytmu SMV.

Zapoznajemy job post processing oraz job post processing solutions(w którym zastosowano uwagi z instrukcji)

4.1.2

a) Został stworzony plik excel z oryginalnymi danymi JobPosts. Aby nie używać retrieve tylko operatora Read Excel.

b)

Poniżej znajduje się screen z oryginalnymi danymi, oraz screeny z poszczególnych procesów:

Możemy zobaczyć jak wyglądają komórki Jobtext i że się nieznacznie różnią(niektóre słowa są inne albo ich niema , np. „\_”) Jest to spowodowane różnością filtrów, tokenów w różnych procesach.

Obraz zawierający zrzut ekranu

Opis wygenerowany automatycznie

Obraz zawierający zrzut ekranu

Opis wygenerowany automatycznie

Obraz zawierający zrzut ekranu

Opis wygenerowany automatycznie

Obraz zawierający zrzut ekranu

Opis wygenerowany automatycznie

b2) 20 najczęściel występujących wyrażeń (Total Occurencies), malejąco

|  |  |  |  |
| --- | --- | --- | --- |
| Filter Stopwords&Tokens | Process Documents from Data | Stemming | Generate n-grams |
| |  |  |  | | --- | --- | --- | | experience |  | 566.0 | | work |  | 471.0 | | please |  | 394.0 | | skills |  | 307.0 | | looking |  | 281.0 | | team |  | 271.0 | | time |  | 256.0 | | resume |  | 247.0 | | position |  | 246.0 | | sales |  | 246.0 | | service |  | 206.0 | | job |  | 195.0 | | working |  | 192.0 | | company |  | 188.0 | | customer |  | 176.0 | | ability |  | 174.0 | | environment |  | 174.0 | | required |  | 168.0 | | years |  | 161.0 | | knowledge |  | 150.0 | | |  |  |  | | --- | --- | --- | | and |  | 3200.0 | | to |  | 2213.0 | | a |  | 1790.0 | | the |  | 1701.0 | | in |  | 1225.0 | | for |  | 1149.0 | | of |  | 1109.0 | | with |  | 999.0 | | you |  | 797.0 | | is |  | 723.0 | | we |  | 665.0 | | are |  | 616.0 | | be |  | 616.0 | | experience |  | 566.0 | | our |  | 472.0 | | work |  | 471.0 | | or |  | 459.0 | | have |  | 448.0 | | an |  | 439.0 | | will |  | 424.0 | | |  |  |  | | --- | --- | --- | | work |  | 688.0 | | experi |  | 579.0 | | pleas |  | 396.0 | | requir |  | 353.0 | | skill |  | 338.0 | | posit |  | 325.0 | | look |  | 313.0 | | time |  | 312.0 | | servic |  | 300.0 | | custom |  | 292.0 | | team |  | 281.0 | | manag |  | 279.0 | | resum |  | 271.0 | | sale |  | 260.0 | | year |  | 222.0 | | s |  | 216.0 | | commun |  | 214.0 | | job |  | 214.0 | | includ |  | 211.0 | | applic |  | 206.0 | | |  |  |  | | --- | --- | --- | | experience |  | 566.0 | | work |  | 471.0 | | please |  | 394.0 | | skills |  | 307.0 | | looking |  | 281.0 | | team |  | 271.0 | | time |  | 256.0 | | resume |  | 247.0 | | position |  | 246.0 | | sales |  | 246.0 | | service |  | 206.0 | | job |  | 195.0 | | working |  | 192.0 | | company |  | 188.0 | | customer |  | 176.0 | | ability |  | 174.0 | | environment |  | 174.0 | | required |  | 168.0 | | years |  | 161.0 | | knowledge |  | 150.0 | |

c1)

Zmienione zostały miejsca docelowe store:testing set, wordlist, model na mój local repository:data

A po uruchomieniu procesu został stworzony proces do zapisu tych plików store w folderze na moim komputerze.

c4)

Uruchomiono procesy bez breakpoint.

c5\*) 20 najcześciej używanych wordlist -malejąco

Brak zmian (?)

|  |  |  |
| --- | --- | --- |
| 1gram(domyślny) | 2gram | 5gram |
| |  |  | | --- | --- | | work | 527.0 | | experi | 492.0 | | pleas | 323.0 | | requir | 294.0 | | skill | 269.0 | | posit | 267.0 | | look | 251.0 | | time | 242.0 | | custom | 232.0 | | servic | 232.0 | | team | 227.0 | | manag | 222.0 | | resum | 218.0 | | sale | 207.0 | | s | 186.0 | | year | 182.0 | | applic | 171.0 | | commun | 167.0 | | job | 165.0 | | includ | 162.0 | | |  |  | | --- | --- | | work | 527.0 | | experi | 492.0 | | pleas | 323.0 | | requir | 294.0 | | skill | 269.0 | | posit | 267.0 | | look | 251.0 | | time | 242.0 | | custom | 232.0 | | servic | 232.0 | | team | 227.0 | | manag | 222.0 | | resum | 218.0 | | sale | 207.0 | | s | 186.0 | | year | 182.0 | | applic | 171.0 | | commun | 167.0 | | job | 165.0 | | includ | 162.0 | | |  |  |  |  | | --- | --- | --- | --- | | work | | 527.0 | | | experi | | 492.0 | | | pleas | | 323.0 | | | requir | | 294.0 | | | skill | | 269.0 | | | posit | | 267.0 | | | look | | 251.0 | | | time | | 242.0 | | | custom | | 232.0 | | | servic | | 232.0 | | | team | | 227.0 | | | manag | | 222.0 | | | resum | | 218.0 | | | sale | | 207.0 | | | s | | 186.0 | | | year | | 182.0 | | | applic | | 171.0 | | | commun | | 167.0 | | | job | 165.0 | | | includ | 162.0 | | |

**Accuracy dla n gram-5**

80,24%, a kappa 0,606 dla performance1

93% dla performance2

**Accuracy dla n gram-2**

83,58%, kappa 0,672 dla performance1

93% dla performance2 (proces zbioru testowego)

**Accuracy dla n gram-1**

86,33%, a kappa 0,727 dla performance1

93% 0,596 Dla performance2

c6)

Zamieniamy algorytm SMV na **Decission tree** i powtarzamy porównanie:

|  |  |  |
| --- | --- | --- |
| 1gram(domyślny) | 2gram | 5gram |
| |  |  | | --- | --- | | work | 527.0 | | experi | 492.0 | | pleas | 323.0 | | requir | 294.0 | | skill | 269.0 | | posit | 267.0 | | look | 251.0 | | time | 242.0 | | custom | 232.0 | | servic | 232.0 | | team | 227.0 | | manag | 222.0 | | resum | 218.0 | | sale | 207.0 | | s | 186.0 | | year | 182.0 | | applic | 171.0 | | commun | 167.0 | | job | 165.0 | | includ | 162.0 | | |  |  | | --- | --- | | work | 527.0 | | experi | 492.0 | | pleas | 323.0 | | requir | 294.0 | | skill | 269.0 | | posit | 267.0 | | look | 251.0 | | time | 242.0 | | custom | 232.0 | | servic | 232.0 | | team | 227.0 | | manag | 222.0 | | resum | 218.0 | | sale | 207.0 | | s | 186.0 | | year | 182.0 | | applic | 171.0 | | commun | 167.0 | | job | 165.0 | | includ | 162.0 | | |  |  | | --- | --- | | work | 527.0 | | experi | 492.0 | | pleas | 323.0 | | requir | 294.0 | | skill | 269.0 | | posit | 267.0 | | look | 251.0 | | time | 242.0 | | custom | 232.0 | | servic | 232.0 | | team | 227.0 | | manag | 222.0 | | resum | 218.0 | | sale | 207.0 | | s | 186.0 | | year | 182.0 | | applic | 171.0 | | commun | 167.0 | | job | 165.0 | | includ | 162.0 | |

**Accuracy dla n gram-5**

80,38%, a kappa 0,6063 dla performance1

89% a kappa 0,497 dla performance2

**Accuracy dla n gram-2**

81,25%, kappa 0,62 dla performance1

94% a kappa 0,694 dla performance2 (proces zbioru testowego)

**Accuracy dla n gram-1**

82,39%, a kappa 0,642 dla performance1

93% a kappa 0,596 Dla performance2

Obserwujemy, że w zakładce wordlist (zapisane w plikach excel wyniki ngrams) nie występują zmiany przy total occurencies. Zmienia się natomiast dokładność i wskaźnik kappa.

d1) d2) d3) Zastosowano operatory podane w poleceniu, w Sample probability=0.5. Zapisano wyniki, operator performance 1 to operator performance distance cluster. Performance 2 to operator performance distinct cluster.

Fragment wyniku Similarity to data(z parametrem long table)

Obraz zawierający zrzut ekranu

Opis wygenerowany automatycznie

Fragment wyniku Similarity to data(z parametrem matrix), gdzie pokazano są prawdopodobieństwa.

Obraz zawierający biały, komputer, duży, laptop

Opis wygenerowany automatycznie

Otrzymano wyniki:

W performance distance cluster:

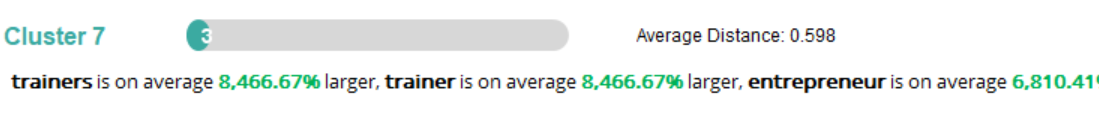
Avg. within centroid distance**: -0.762**

W performance distinct cluster :

Avg. within cluster similarity**: 1.661**

d4)\*

d5) Dla Cluster 7 (mało liczne bo -3 items)



Korzystając z Centroid Table próbujemy swierdzić, które słowa wydają się decydować o similarity(uznaniu ofert pracy za podobne w tym clustrze), słowa:

**entrepreneur(0,222), trainer (0.146), trainers(0.195), train(0.103),**

d6)

Zastosowano operator generate n-grams dla 2 grams, otrzymano wyniki:

W performance distance cluster:

Avg. within centroid distance**: -0.552**

W performance distinct cluster :

Avg. within cluster similarity**: 6.185**

Średni dystans w clustrze zmniejszył się oraz Srednie podobieństwa w clustrze zwiększyły się co oznacza że operator ngrams poprawił wskaźniki jakości grupowania.

e) Korzystamy z zbioru wos. Wos1 dla zbioru uczącego i dla testowego zbiór Wos2.

Zostały utworzone dwa procesy(dwa pliki). Drugi plik/proces został dodany algorytm decision tree z walidacją krzyżową

e2)\*Nie udało się uruchomić dla pliku pdf.