

## 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 użyliśmy algorytmu klasyfikacji i regresji Generalize linear 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 zmienną 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 nie ma, np. „\_”) Jest to spowodowane różnicą filtrów, tokenów w różnych procesach.

ExampleSet (+ Stemming)

ExampleSet (Multiply)

ExampleSet (+ Filter Stopwords & Tokens)

ExampleSet (+ Generate n-grams)

History

ExampleSet (Process Documents from Data)

Open in

Turbo Prep

Auto Model

Filter (499 / 499 examples): all

Row No.	Category	JobText
1	customer ser...	OGPlanet (www.ogplanet.com) is an online game publisher that specializes in high-quality, ...
2	customer ser...	Our company is right now looking for a full-time (40 hours per week) reproduction/photocopi...
3	customer ser...	 
4	customer ser...	Due to rapid growth, we currently have several openings for the role of Financial Services Re...
5	customer ser...	 
6	customer ser...	Direct Promotions & Customer Service
7	food/beverag...	Really!! A Whole \$13.00 an hour... for those responsibilities! Ha! What a joke! Bakery manag...
8	food/beverag...	We have expanded and we are looking for Managers and Assistant Managers. If you would ...
9	food/beverag...	I know hey!
10	food/beverag...	We currently have opening for day and evening shifts ranging from 20-40/hrs per week.
11	food/beverad...	Looking for a chef to do a sit down meal for approximately 13 adults.

ExampleSet (499 examples, 1 special attribute, 1 regular attribute)

History

ExampleSet (Process Documents from Data)

Open in

Turbo Prep

Auto Model

Filter (499 / 499 examples): all

Row No.	text
1	ogplanet www ogplanet com is an online game publisher that specializes in high quality downloadable free to...
2	our company is right now looking for a full time hours per week reproduction photocopier expert to join our fast...
3	are you ready to take your career to the next level we are looking for the right candidates to help us expand in th...
4	due to rapid growth we currently have several openings for the role of financial services representative this is a...
5	our growing coffee company is looking for an customer service rep to help grow our business responsibilities ...
6	direct promotions customer service our company is looking for smart fun dynamic people to join our growing t...
7	really a whole an hour for those responsibilities ha what a joke bakery management comes at a price and you...
8	we have expanded and we are looking for managers and assistant managers if you would like a new challeng...
9	i know hey it s crazy that people think they can get away with paying others that i understand that starting up a b...
10	we currently have opening for day and evening shifts ranging from hrs per week applicants must be very custo...
11	looking for a chef to do a sit down meal for approximately adults this will include appetizers main meal and a d...

ExampleSet (499 examples, 1 special attribute, 0 regular attributes)

ExampleSet (+ Filter Stopwords & Tokens)

ExampleSet (+ Generate n-grams)

ult History

ExampleSet (Process Documents from Data)

Open in

Turbo Prep

Auto Model

Filter (499 / 499 examples): all

Row No.	text
1	ogplanet ogplanet_www ogplanet_www_ogplanet www www_ogplanet www_ogplanet_com ogplanet o...
2	company company_looking company_looking_full looking looking_full looking_full_time full full_time ful...
3	ready ready_take ready_take_career take take_career take_career_level career career_level career_lev...
4	due due_rapid due_rapid_growth rapid rapid_growth rapid_growth_currently growth growth_currently gr...
5	growing growing_coffee growing_coffee_company coffee coffee_company coffee_company_looking co...
6	direct direct_promotions direct_promotions_customer promotions promotions_customer promotions_c...
7	hour hour_responsibilities hour_responsibilities_ha responsibilities responsibilities_ha responsibilitie...
8	expanded expanded_looking expanded_looking_managers looking looking_managers looking_manag...
9	i_know i_know_hey know know_hey know_hey_s hey hey_s hey_s_crazy s_crazy s_crazy_people crazy ...
10	currently currently_opening currently_opening_day opening opening_day opening_day_evening day day...
11	lookina lookina_chef lookina_chef_sit chef chef_sit chef_sit meal sit sit_meal sit_meal approximatelv ...

ExampleSet (499 examples, 1 special attribute, 0 regular attributes)

ExampleSet (+ Stemming)

ExampleSet (Multiply)

ExampleSet (+ Filter Stopwords & Tokens)

ExampleSet (+ Generate n-grams)

esult History

ExampleSet (Process Documents from Data)

Open in

Turbo Prep

Auto Model

Filter (499 / 499 examples): all

Row No.	text
1	ogplanet www ogplanet com onlin game publish special qualiti download free plai casual onlin multiplay game e...
2	compani look full time hour week reproduct photocopi expert join fast grow compani work industri size copier sca...
3	readi take career level look candid help expand vancouv greater area market offer person attent profession advant...
4	due rapid growth current open role financi servic repres self employ opportun success candid build financi practic...
5	grow coffe compani look custom servic rep help grow busi respons cold call establish busi prospect abil work fas...
6	direct promot custom servic compani look smart fun dynam peopl join grow team know remain competit offer rew...
7	hour respons ha joke bakeri manag come price lead baker north burnabi date pst repli job wgkva craigslist org er...
8	expand look manag assist manag challeng get paid train work fun environ look join domino s pizza team pleas se...
9	i know hei s crazi peopl think get pai other i understand start busi expens cost manag review hr s gener labour qu...
10	current open dai even shift rang hr week applic custom servic focus orient energi friendli enthusiast enjoi work fas...

b2) 20 najczęściel występujących wyrażen (Total Occurencies), malejaco

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

Brak zmian (?)

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

leSet (Multiply)

ExampleSet (Multiply)


**ExampleSet (Similarity to Data)**


y

% PerformanceVector (Performance)

% PerformanceVector (Performance (2))

Open in

 Turbo Prep

 Auto Model

Filter (65,792 / 65,792 examples): 

all

Row No.	FIRST_ID	SECOND_ID	SIMILARITY
1	1	2	0.040
2	1	3	0.078
3	1	4	0.028
4	1	5	0.002
5	1	6	0.002
6	1	7	0.035
7	1	8	0
8	1	9	0.034
9	1	10	0.033
10	1	11	0.013
11	1	12	0
12	1	13	0.042
13	1	14	0.060

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

ExampleSet (Multiply)

ExampleSet (Multiply)

ExampleSet (Similarity to Data)

PerformanceVector (Performance)

PerformanceVector (Performance (2))

Open in

Turbo Prep

Auto Model

Filter (257 / 257 examples): all

Row No.	ID	1	2	3	4	5	6
1	1	1	0.040	0.078	0.028	0.002	0.002
2	2	0.040	1	0.082	0.014	0.002	0.010
3	3	0.078	0.082	1	0.026	0.004	0.001
4	4	0.028	0.014	0.026	1	0.012	0.003
5	5	0.002	0.002	0.004	0.012	1	0.002
6	6	0.002	0.010	0.001	0.003	0.002	1
7	7	0.035	0.028	0.131	0.017	0.179	0
8	8	0	0.013	0.013	0.089	0	0
9	9	0.034	0.063	0.071	0.025	0.006	0.003
10	10	0.033	0.007	0.048	0.016	0.007	0.002
11	11	0.013	0.028	0.030	0.016	0.011	0.025
12	12	0	0.004	0	0.011	0.006	0

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)

**Cluster 7**  Average Distance: 0.598  
**trainers** is on average **8,466.67%** larger, **trainer** is on average **8,466.67%** larger, **entrepreneur** is on average **6,810.41%**

Korzystając z Centroid Table próbujemy stwierdzić, 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 Średnie 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.