Mateusz Marciniewicz
Przemysław Bedełek

Inżynieria Uczenia Maszynowego
Preprocessing

Etap: 1
Iteracja: 1
Technologia: SAS 4GL

1. Import plików .jsonl

Uruchomienie poniższego kodu pozwoliło na zaimportowanie ustrukturyzowanych plików tekstowych w formie zrozumiałej dla środowiska *Enterprise Guide*.

```
%let path=C:\Users\splmum\Desktop\IUM;
libname ium "&path";


%macro importAllJSONL(path);
      options nonotes nosource;
      filename filelist pipe "dir /b /s &path\*.jsonl";

      data _null_;
            infile filelist truncover;
            input filename $100.;
            put filename=;
            call execute(cats('%importJSONL(', filename, ')'));
      run;

      options notes source;
%mend;

%macro importJSONL(file);
%let setName = %scan(&file, -2,/.\);
%put &=setName;
      filename jsonl "&file";
      filename json temp;
      filename map temp;

      data _null_;
            infile jsonl end=eof;
            file json;
            put '[{"records":['; /* Required for JSON formatting */
            do until (eof);
                  input;
                  line = STRIP(_infile_);
                  if (eof=0) then
                        line=cats(line, ',');

                  put line;
            end;
            put ']}]'; /* Required for JSON formatting */
            stop;
      run;
      libname json json automap=create map=map;
      proc print data=json.records; run;


      data ium.&setName;
            set json.records;
      run;


%mend;
```

2. Preprocessing tabeli USERS

a. Nagłówek tabeli nieprzetworzonej

Obs.	user_id	name	city	street
1	102	Jędrzej Murach	Konin	aleja Kwiatowa 05/25
2	103	Iwo Kubasiak	Radom	plac Floriana 81/08
3	104	Ewelina Litwa	Kutno	al. Żeromskiego 261
4	105	Daniel Wojewodzic	Radom	ulica Spółdzielcza 78
5	106	Krzysztof Doroszuk	Kutno	aleja Willowa 14/49
6	107	Agnieszka Miąsko	Szczecin	plac Głowackiego 009
7	108	Tymoteusz Majsterek	Szczecin	plac Słowianska 29/93
8	109	Rozalia Litwa	Szczecin	aleja Cyprysowa 868
9	110	Gustaw Kudłacz	Gdynia	ulica Wierzbowa 026
10	111	Konrad Rapa	Gdynia	pl. Dębowa 94/87
11	112	Justyna Parzyszek	Mielec	ulica Pomorska 12
12	113	Błażej Buśko	Konin	aleja Jaśminowa 20/09
13	114	Anna Maria Sołek	Szczecin	aleja Zamkowa 88
14	115	Arkadiusz Sadza	Police	ulica Pomorska 20/40
15	116	Cyprian Skierka	Gdynia	aleja Reja 108

b. Obserwacje

- Zmienna *name* nie ma i nie powinna mieć żadnego wpływu na czas dostawy, a zatem może zostać wykluczona ze zbioru zmiennych.
- Należy ujednolicić zapis rodzajów adresów aby uniknąć ich błędnej interpretacji przez model uczenia - al. Żeromskiego oraz aleja Żeromskiego to te same ulice
- Należy oddzielić numerację adresów od nazwy ulicy – al. Reja 108 oraz al. Reja 110 odnoszą się do tej samej ulicy.
- Należy ujednolicić sposób numerowania adresów tak, aby mieć pewność, że przykładowo pl. Dębowy 9 będzie jednoznaczny z pl. Dębowym 009.
- Należy usunąć ze zbioru rozpatrywanych zmiennych numer mieszkania – różnica w czasie dostawy między mieszkaniami na jednej ulicy powinna być pomijalna w kontekście czasu dostawy szacowanego w jednostce godzin.

c. Kod preprocessingu

```
data users_preprocessed;
length streetName $ 50;
      set ium.users(drop=name);

      /* Process street's type (ul. pl. al.) */
      streetType=scan(street, 1, ' ');

      if find(streetType, '.') = 0 then
            streetType = cats(substr(streetType, 1, 2), '.');

      /* Process street's name */
      do word=2 to countw(street, ' ')-1;
      (' ', streetName, scan(street, word, ' '));
      end;

      /* Process street's number - drop the apartament number */
      streetNumber = scan(street, -1, ' ');
      streetNumber = prxchange('s/(0*)([1-9]+0*)(\/\d+)?/$2/io', 1, streetNumber);

      /* Concatenate unified street type and it's name */
      street = catx(' ', streetType, streetName);

      /* Keep just the necessary variables */
      keep user_id city street streetNumber;
run;
```

d. Nagłówek tabeli przetworzonej

Obs.	user_id	city	street	streetNumber
1	102	Konin	al. Kwiatowa	5
2	103	Radom	pl. Floriana	81
3	104	Kutno	al. Żeromskiego	261
4	105	Radom	ul. Spółdzielcza	78
5	106	Kutno	al. Willowa	14
6	107	Szczecin	pl. Głowackiego	9
7	108	Szczecin	pl. Słowianska	29
8	109	Szczecin	al. Cyprysowa	868
9	110	Gdynia	ul. Wierzbowa	26
10	111	Gdynia	pl. Dębowa	94
11	112	Mielec	ul. Pomorska	12
12	113	Konin	al. Jaśminowa	20
13	114	Szczecin	al. Zamkowa	88
14	115	Police	ul. Pomorska	20
15	116	Gdynia	al. Reja	108

3. Preprocessing tabeli DELIVERIES

a. Nagłówek tabeli nieprzetworzonej

Obs.	purchase_id	purchase_timestamp	delivery_timestamp	delivery_company
1	20001	2020-08-21T17:22:16	2020-08-20T12:11:31.068226	620
2	20002	2020-10-10T17:13:56	2020-10-10T01:02:44.951840	360
3	20003	2020-10-02T02:39:14	2020-10-05T08:05:06.777539	360
4	20004	2020-08-29T00:49:38	2020-08-28T23:00:50.029353	620
5	20005	2020-08-29T19:24:01	2020-08-30T08:51:49.267894	.
6	20006	2020-09-25T02:59:07	2020-09-27T18:09:24.633484	620
7	20007	2020-09-24T02:33:57	2020-09-20T23:13:03.258877	620
8	20008	2020-01-07T18:04:44	2020-01-07T12:05:35.768876	360
9	20009	2020-01-05T06:53:12	2020-01-04T17:28:35.608405	620
10	20010	2020-04-30T01:54:45	2020-04-28T14:44:58.059691	360
11	20011	2020-08-05T18:15:26	2020-08-09T23:07:55.654692	360
12	20012	2020-06-14T09:00:43	2020-06-16T01:30:08.388811	360
13	20013	2020-04-29T00:43:52		360
14	20014	2020-02-09T21:27:20	2020-02-11T01:17:02.131515	620
15	20015	2020-02-12T13:11:35	2020-02-13T14:14:48.255407	360

b. Obserwacje

- Na zmienne *purchase_timestamp* oraz *delivery_timestamp* należy przeformatować tak, aby pozwolić na zrozumiały odczyt daty i pory dnia
- Należy obliczyć czas dostawy. Najrozsądniejszym byłby czas dostawy zaokrąglony do godzin.
- Przeformatowane zmienne należy rozdzielić na miesiące, dni, godziny oraz dni tygodnia. Są to zmienne kategoryczne
- Rok jest zmienną kategoryczną o nieskończonym zbiorze, więc należy tę zmienną pominąć.
- Delivery company należy traktować jak zmienną kategoryczną
- Wszelkie rekordy z brakującą wartością *purchase_timestamp* lub *delivery_timestamp* należy usunąć, gdyż nie jesteśmy w stanie obliczyć dla nich czasu dostawy
- Wszelkie rekordy z ujemną wartością czasu dostawy należy pominąć

c. Kod preprocessingu

```
/* Create a set of companies categorized names */
proc sql noprint;
 create table companiesFormat as
       select distinct
             'companies' as fmtname,
             delivery_company as start,
             cats('C', delivery_company) as label
       from ium.deliveries
             where delivery_company is not missing;
quit;

proc freq data=ium.deliveries;
 tables delivery_company;
 format delivery_company companies.;
run;

proc format cntlin=companiesFormat;
run;

data deliveries_preprocessed;
 set ium.deliveries(rename=(delivery_timestamp=deliverypurchase_timestamp=purchase);
 where delivery_company ne .;

 /* Informat datetimes */
 delivery = scan(delivery, -2, '.');
 purchase = input(purchase, anydtdtm19.);
 delivery = input(delivery, anydtdtm19.);

 /* Extract date and time parts */
 purchase_month = month(datepart(purchase));
 purchase_day = day(datepart(purchase));
 purchase_hour = hour(purchase);
 purchase_weekday = weekday(datepart(purchase));
 delivery_month = month(datepart(delivery));
 delivery_day = day(datepart(delivery));
 delivery_hour = hour(delivery);
 delivery_weekday = weekday(datepart(delivery));

 /* Extract the Y from data */
 shipping_time_in_hours = round((delivery-purchase)/3600, 1);

 /* Output observations with positive delivery time value */
 if shipping_time_in_hours > 0 then  output;

 /* Format values some of the variables */
 format
       purchase_month delivery_month roman5.
       purchase_weekday delivery_weekday downame12.
       delivery_company companies.;

 /* Drop unnecessary variables */
 drop purchase delivery;
run;
```

d. Nagłówek tabeli przetworzonej

Obs.	purchase_id	delivery_company	purchase_month	purchase_day	purchase_hour	purchase_weekday
1	20003	C360	X	2	2	Thursday
2	20006	C620	IX	25	2	Thursday
3	20011	C360	VIII	5	18	Tuesday
4	20012	C360	VI	14	9	Saturday
5	20014	C620	II	9	21	Saturday
6	20015	C360	II	12	13	Tuesday
7	20017	C620	II	27	9	Wednesday
8	20018	C620	IX	27	22	Saturday
9	20019	C516	VII	13	7	Sunday
10	20020	C360	IX	4	2	Thursday
11	20021	C360	VIII	15	16	Friday
12	20022	C516	I	19	8	Saturday
13	20023	C360	V	8	6	Thursday
14	20024	C620	I	18	9	Friday
15	20025	C516	VII	3	6	Thursday

Obs.	delivery_month	delivery_day	delivery_hour	delivery_weekday	shipping_time_in_hours
1	X	5	8	Sunday	77
2	IX	27	18	Saturday	63
3	VIII	9	23	Saturday	101
4	VI	16	1	Monday	40
5	II	11	1	Monday	28
6	II	13	14	Wednesday	25
7	II	28	2	Thursday	17
8	IX	29	5	Monday	31
9	VII	14	14	Monday	31
10	IX	4	21	Thursday	19
11	VIII	18	7	Monday	62
12	I	20	23	Sunday	40
13	V	8	14	Thursday	8
14	I	21	3	Monday	66
15	VII	5	9	Saturday	51

4. Preprocessing tabeli SESSIONS

a. Nagłówek tabeli nieprzetworzonej

Obs.	session_id	timestamp	user_id	product_id	event_type	offered_discount	purchase_id
1	100001	2020-08-21T17:19:51	102	1318	VIEW_PRODUCT	0	.
2	100001	2020-08-21T17:22:16	102	1318	BUY_PRODUCT	0	20001
3	100002	2020-10-10T17:13:27	102	1278	VIEW_PRODUCT	5	.
4	100002	2020-10-10T17:13:56	102	1278	BUY_PRODUCT	5	20002
5	100003	2020-10-02T02:24:55	102	1011	VIEW_PRODUCT	10	.
6	100003	2020-10-02T02:28:35	102	1013	VIEW_PRODUCT	10	.
7	100003	2020-10-02T02:30:25	102	1007	VIEW_PRODUCT	10	.
8	100003	2020-10-02T02:31:18	102	1012	VIEW_PRODUCT	10	.
9	100003	2020-10-02T02:32:06	102	1009	VIEW_PRODUCT	10	.
10	100003	2020-10-02T02:33:02	102	1010	VIEW_PRODUCT	10	.
11	100003	2020-10-02T02:33:07	102	1006	VIEW_PRODUCT	10	.
12	100003	2020-10-02T02:36:43	102	1008	VIEW_PRODUCT	10	.
13	100003	2020-10-02T02:37:57	102	1004	VIEW_PRODUCT	10	.
14	100003	2020-10-02T02:39:14	102	1004	BUY_PRODUCT	10	20003
15	100004	2020-08-29T00:44:22	102	1281	VIEW_PRODUCT	15	.

b. Obserwacje

- Zmienna *session_id* w kontekście zadania nie jest potrzebna. Można ją usunąć
- Zmienna *offered_discount* również jest niepotrzebna i zostanie usunięta
- Zmienna *timestamp* pokrywa się ze zmienną *purchase_timestamp* z tabeli *DELIVERIES*, zatem może zostać usunięta
- W kontekście zadania interesującymi rekordami są te, których zmienna *event_type* przyjmuje wartość *„BUY_PRODUCT”*, a wartości zmiennych *user_id* oraz *purchase_id* są niepuste. Rekordy niespełniające wspomnianych warunków powinny zostać usunięte
- Po odfiltrowaniu rekordów, należy odrzucić zmienną *event_type*

c. Kod preprocessingu

```
data sessions_preprocessed;
    set ium.sessions(drop=session_id offered_discount timestamp);
    where upcase(event_type) = "BUY_PRODUCT"
       and purchase_id ne .
       and user_id ne .;
    drop event_type;
run;
```

d. Nagłówek tabeli przetworzonej

Obs.	user_id	product_id	purchase_id
1	102	1318	20001
2	102	1278	20002
3	102	1004	20003
4	102	1278	20004
5	102	.	20005
6	102	1233	20006
7	102	1035	20007
8	102	1234	20008
9	102	1048	20009
10	102	1234	20010
11	102	1283	20011
12	102	1076	20012
13	102	1001	20013
14	102	1317	20014
15	102	1005	20015

5. Preprocessing tabeli PRODUCTS

a. Nagłówek tabeli nieprzetworzonej

Obs.	product_id	product_name	category_path	price
1	1001	Telefon Siemens Gigaset DA310	Telefony i akcesoria;Telefony stacjonarne	58.97
2	1002	Kyocera FS-1135MFP	Komputery;Drukarki i skanery;Biurowe urządzenia wielofunkcyjne	2048.50
3	1003	Kyocera FS-3640MFP	Komputery;Drukarki i skanery;Biurowe urządzenia wielofunkcyjne	7639.00
4	1004	Fallout 3 (Xbox 360)	Gry i konsole;Gry na konsole;Gry Xbox 360	49.99
5	1005	Szalone Króliki Na żywo i w kolorze (Xbox 360)	Gry i konsole;Gry na konsole;Gry Xbox 360	49.99
6	1006	Call of Duty 4 Modern Warfare (Xbox 360)	Gry i konsole;Gry na konsole;Gry Xbox 360	59.90
7	1007	Dead Space 3 (Xbox 360)	Gry i konsole;Gry na konsole;Gry Xbox 360	89.99
8	1008	Tom Clancy's Rainbow Six Vegas (Xbox 360)	Gry i konsole;Gry na konsole;Gry Xbox 360	-49.99
9	1009	Kinect Joy Ride (Xbox 360)	Gry i konsole;Gry na konsole;Gry Xbox 360	69.00
10	1010	BioShock 2 (Xbox 360)	Gry i konsole;Gry na konsole;Gry Xbox 360	89.99
11	1011	BioShock Infinite (Xbox 360)	Gry i konsole;Gry na konsole;Gry Xbox 360	-139.99
12	1012	Fallout New Vegas (Xbox 360)	Gry i konsole;Gry na konsole;Gry Xbox 360	69000000.00
13	1013	LA Noire (Xbox 360)	Gry i konsole;Gry na konsole;Gry Xbox 360	129.99
14	1014	Lego Batman 2 DC Super Heroes (Xbox 360)	Gry i konsole;Gry na konsole;Gry Xbox 360	69.99
15	1015	Max Payne 3 (Xbox 360)	Gry i konsole;Gry na konsole;Gry Xbox 360	89.99

b. Obserwacje

- Zmienna *product_name* nie powinna wpływać na czas dostawy produktu, zatem może zostać usunięta
- Zmienna *price* nie powinna wpływać na czas dostawy produktu, zatem może zostać usunięta
- Zmienna *category_path* powinna zostać zmodyfikowana w taki sposób, aby zawierała tylko ostatni człon

c. Kod preprocessingu

```
data products_preprocessed;
 set ium.products(keep=product_id category_path);
 category_path = scan(category_path, -1, ';');
run;
```

d. Nagłówek tabeli przetworzonej

Obs.	product_id	category_path
1	1001	Telefony stacjonarne
2	1002	Biurowe urządzenia wielofunkcyjne
3	1003	Biurowe urządzenia wielofunkcyjne
4	1004	Gry Xbox 360
5	1005	Gry Xbox 360
6	1006	Gry Xbox 360
7	1007	Gry Xbox 360
8	1008	Gry Xbox 360
9	1009	Gry Xbox 360
10	1010	Gry Xbox 360
11	1011	Gry Xbox 360
12	1012	Gry Xbox 360
13	1013	Gry Xbox 360
14	1014	Gry Xbox 360
15	1015	Gry Xbox 360

6. Scalenie tabel

Warto zwrócić uwagę na to, że zmienne *purchase_id*, *session_id*, *user_id* oraz *product_id* niosą wartość jedynie w przypadku łączenia tabel, dlatego po otrzymaniu tabeli wynikowej, mogą zostać pominięte.

W celu zachowania estetyki dokumentu, w raporcie zamieszczony został jedynie kod źródłowy, a tabela wynikowa pokazana będzie w notebook'u.

```
proc sql;
 create table
      dataset_processed(drop=session_id user_id product_id purchase_id)
 as
      select *
           from sessions_preprocessed S
                inner join users_preprocessed U
                     on S.user_id = U.user_id
                inner join deliveries_preprocessed D
                     on S.purchase_id = D.purchase_id
                left join products_preprocessed P
                     on S.product_id = P.product_id;
quit;

ods csvall file="&path/ium_preprocessed.csv";

proc print data=dataset_processed;
run;

ods csvall close;
```