Labs 20 Russian Arms Exporters Project

By Andrea Christelle, Ian Forrest, David Hang, Curtis McKendrick, & Andre Savkin

Overview

- Review of Labs 18 Project
- Labs 20 Project - ‘DESCRIPTION_GOOD’ Column Text Analysis
- Corporate Registry Data
- Data Cleaning
- Russian Text Preprocessing
- Text Vectorization
- KMeans Modeling & FIndings
- Predictive Function
- Results
- Additional Testing (Ukrainian Dataset)

Review of Labs 18 Project

- Good analysis and thought
- Used list of known Russian arms exporters carried over from Labs 16
- Valid results!
- No analysis of underlying data - 1s and 0s
- Issues with list - missing INNs, wrong INNs
- Different direction - analysis of ‘DESCRIPTION_GOOD’ column

[2]:
```
# reads in encoded dataframe from s3
df = dd.read_csv('s3://labs18-arms-bucket/data/trade_encoded0.csv/*.csv')
df.head()
```

[2]:

	DECLARATION_NUMBER	CUSTOMS_POST_CODE	CUSTOMS_POST_NAME	STAT	CONSIGNOR_INN	CONSIGNEE_NAME_nosuffix	CONSIGNEE_COUNTRY	PAYEE_INN	DECLARANT_INN	TOTAL
0	0	1	1	1.0	5032136476	0	1	0	0	
1	0	0	1	1.0	2348034933	0	0	0	0	
2	0	0	1	1.0	7705935687	0	1	0	0	
3	0	0	1	1.0	7811071894	0	1	0	0	
4	0	1	1	1.0	6165178580	0	1	0	0	

5 rows × 46 columns

Corporate Registry Data

- Unable to incorporate
- Messy
- Time constraints

Data Cleaning

- PySpark SageMaker Notebook linked to EMR Cluster
  - Manage large dataset
  - Limit columns - Consignor Name, Consignor INN, Description of Goods
  - Exports only
  - Initial punctuation cleaning in ‘DESCRIPTION_GOOD’ column

- Python3 SageMaker Data Cleaning
  - Python Dictionary
  - Saved over 177,000 rows of data!

Russian Text Preprocessing

- Make ‘DESCRIPTION_GOOD’ column text more useable
- Tokenization - isolate individual Russian words for analysis
- Punctuation removal
- Stop word removal
- Adding trade-specific stop words

Trade-Specific stop words added to list:

```
#create stem and stopwords list
mystem = Mystem()
russian_stopwords = stopwords.words("russian")

# https://stackoverflow.com/questions/5511708/adding-words-to-nltk-stoplist
# add trade-specific stopwords to list
newStopWords = ['г', '№', '10', '1', '20', '30', 'кг', '5', 'см',
                '100', '80', '2', 'х', 'l', 'м', '00', '000'
                '1.27', '2011.10631', '4', '12', '3', 'фр', 'количество',
                'становиться', 'мм', 'вид', 'упаковка', 'получать',
                'прочий', 'использование', 'масса', 'размер', 'черный',
                '6', '8', '7', '50', '40', '25', 'коробка', 'поддон',
                'вдоль', '250', '65', '85', '15', '35', '40', '45',
                '55', '60', '70', '75', 'м3', '13', '0', '14',
                '16', '18', 'm2', 'n', 'p', 't', 'тип', 'являться',
                'размер', 'cm', 'm', '01', '02', '03', '04', '05',
                '06', '07', '08', '09', '24', '27']
russian_stopwords.extend(newStopWords)
```

Text Vectorization

- Converting ‘DESCRIPTION_GOOD’ column to a series of numbers
- Value of number corresponds to how our model should use a given word to define the description as a whole
- Zeros/low decimals indicate a word shouldn’t define a given description
- High decimals indicate a word can define a given trade description

	CONSIGNOR_NAME	CONSIGNOR_INN	PROCESSED_TEXT	00	10	11	27	848686	88104см	90	...	черновой	швейный	шина	шип	шлифовать	шт
0	АОЧЕРЕПОВЕЦКИЙ ФАНЕРНО-МЕБЕЛЬНЫЙ КОМБИНАТ	3528006408	пиломатериалыдоска еловый picea abies обрезная...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.360739	0.0	0.000000
1	АО ПИКАЛЁВСКАЯ СОДА	4715022874	сульфат калий калий серонокислый технический к...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.000000
2	ООО ИНТЕР-ТРАНС	6324057625	швеллер бампер ваз 21900280301501 шт	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.604300
3	ООО ЗЕЛЕНЫЙ СВЕТ	3849029492	лесоматериал праспиливать вдольнестроганыенелу...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.000000
4	ОАО АВИАКОМПАНИЯ УРАЛЬСКИЕ АВИАЛИНИИ	6608003013	телефонный проводной трубка сбор связь бортпро...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.000000	0.0	0.345262

KMeans Modeling & Findings

- Grouped into 10 clusters

```
Cluster 0:
Russian:  шт , English: PCS
Russian:  часть , English: part
Russian:  ваз , English: vases
Russian:  изделие , English: product
Russian:  марка , English: mark
Russian:  гост , English: guest
Russian:  материал , English: material
Russian:  назначение , English: appointment
Russian:  сталь , English: steel
Russian:  вес , English: weight
Russian:  новый , English: new
Russian:  длина , English: length
Russian:  рулон , English: roll
Russian:  металл , English: metal
Russian:  диаметр , English: diameter
```

```
Cluster 4:
Russian:  трикотажный , English: knitted
Russian:  обхват , English: coverage
Russian:  вязание , English: knitting
Russian:  машинный , English: machine
Russian:  машинный вязание , English: knitting machine
Russian:  женский , English: female
Russian:  нить , English: a thread
Russian:  пряжа , English: yarn
Russian:  рост , English: height
Russian:  грудь , English: chest
Russian:  хлопчатобумажный , English: cotton
Russian:  обхват грудь , English: chest girth
Russian:  хлопчатобумажный пряжа , English: cotton yarn
Russian:  одежда , English: clothes
Russian:  класс люкс , English: luxury
```

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
Known Russian Arms Exporter Trades per Cluster (in training dataset)	55,151	9,961	81	42	13	23	8,723	1,356	348	0
Percent of Total Trades	72.86%	13.16%	0.11%	0.06%	0.02%	0.03%	11.52%	1.79%	0.46%	0.00%

Predictive Function

- Runs through entire process
- Calculates trades per cluster for every INN
- Convert to percentages

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
INN X	5,500	31	12	332	22	23	1,123	12	3	32
INN X - Percent of Total Trades	77.57%	0.44%	0.17%	4.68%	0.31%	0.32%	15.84%	0.17%	0.04%	0.45%

- Then compare to percentages of known arms exporters

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
INN X - Percent of Total Trades	77.57%	0.44%	0.17%	4.68%	0.31%	0.32%	15.84%	0.17%	0.04%	0.45%
Known Russian Arms Exporters - Percent of Total Trades	72.86%	13.16%	0.11%	0.06%	0.02%	0.03%	11.52%	1.79%	0.46%	0.00%

Predictive Function

- Score measured using inverse Euclidean distance
- Smaller cluster similarity increased score too much

	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 7	Cluster 8	Cluster 9
INN X - Percent of Total Trades	0.17%	4.68%	0.31%	0.32%	0.17%	0.04%	0.45%
Known Russian Arms Exporters - Percent of Total Trades	0.11%	0.06%	0.02%	0.03%	1.79%	0.46%	0.00%

- Removed from final analysis

	Cluster 0	Cluster 1	Cluster 6
INN X - Percent of Total Trades	77.57%	0.44%	15.84%
Known Russian Arms Exporters - Percent of Total Trades	72.86%	13.16%	11.52%

Results

- Ran on test dataframe
- 32,000 to 3,000!
- Link to results dataframe: http://localhost:8891/notebooks/notebooks/results.ipynb

	CONSIGNOR_INN	clust0	clust1	clust6	pdist_score
1213	3123138830	0.775862	0.086207	0.137931	0.935134
2763	610210391297	0.726829	0.063415	0.121951	0.935869
2772	6119005430	0.723735	0.073930	0.147860	0.937712
3091	6166093748	0.773333	0.133333	0.066667	0.938020
3138	6167129059	0.666667	0.115226	0.123457	0.939363
4797	7814413641	0.755682	0.079545	0.142045	0.939392
4877	7839447850	0.761468	0.082569	0.091743	0.940256
2814	6145001168	0.781250	0.109375	0.109375	0.945649
2348	5075018950	0.694074	0.175451	0.121340	0.946850
3002	6164021988	0.692308	0.137652	0.072874	0.946891
2877	6154135457	0.681303	0.110482	0.094901	0.947317
3724	7536146773	0.684685	0.162162	0.103604	0.948108
2898	6154141429	0.686275	0.156863	0.137255	0.948798
3619	7302040482	0.770115	0.098659	0.112069	0.949568
4546	7736207543	0.685185	0.148148	0.129630	0.953639
592	2311196470	0.741379	0.086207	0.124138	0.954208
1304	332701210896	0.706897	0.172414	0.120690	0.955529
1333	3442093695	0.725490	0.156863	0.078431	0.957165
2780	6123015760	0.725490	0.156863	0.078431	0.957165
3169	6168094673	0.722772	0.118812	0.079208	0.962778
2277	5042145938	0.714286	0.166667	0.119048	0.963331
3409	6659007785	0.700000	0.150000	0.125000	0.965847
2935	6161053692	0.715074	0.148897	0.091912	0.968969
4792	7814121367	0.707317	0.121951	0.121951	0.976295
3962	7706216371	0.742857	0.142857	0.102857	0.978465

```
test.shape
```

```
(2996, 5)
```

Additional Testing

- Tested on Ukrainian dataset, results were positive!
- Returned 65 EDRPOUs of interest

	SHIPPER_EDRPOU	clust0	clust1	clust6	pdist_score
940	2781504255	0.837838	0.162162	0.0	0.860791
1834	32502825	0.759494	0.240506	0.0	0.860918
1497	31489175	0.761194	0.238806	0.0	0.861515
2840	36767366	0.833333	0.166667	0.0	0.862335
2755	36417791	0.837989	0.117318	0.0	0.862404
3468	38884736	0.766667	0.233333	0.0	0.863279
3298	382272	0.829268	0.170732	0.0	0.863586
3774	39792657	0.826923	0.173077	0.0	0.864244
697	24363204	0.770950	0.229050	0.0	0.864483
1587	31804036	0.773869	0.226131	0.0	0.865212
3138	377511	0.822496	0.177504	0.0	0.865355
383	20925875	0.824859	0.169492	0.0	0.865891
3001	37310549	0.820043	0.179957	0.0	0.865894
4307	413475	0.818671	0.181329	0.0	0.866172
626	23510703	0.785714	0.214286	0.0	0.867365
2059	33240672	0.790476	0.209524	0.0	0.867853
3214	37987502	0.791980	0.208020	0.0	0.867961
3291	382102	0.804348	0.195652	0.0	0.868001
3725	39695169	0.794872	0.205128	0.0	0.868107
4299	41330519	0.755102	0.040816	0.0	0.870267
1226	30638249	0.752747	0.219780	0.0	0.871756
2375	34589850	0.816667	0.116667	0.0	0.872738
4428	41633830	0.804245	0.101415	0.0	0.876327
2654	35947117	0.806387	0.127745	0.0	0.877883
4741	692096	0.756098	0.134146	0.0	0.894051

```
test.shape
```

```
(65, 5)
```

Conclusions

- Break up Cluster0
- Make vectorizer more sensitive
- Expand list of known Russian arms exporters - improve model strength
- Create list of known non-arms exporters - improve product performance

Thank You!