SMART INDIA HACKATHON 2019

MINISTRY OF RAILWAYS

PROBLEM STATEMENT: DE-DUPLICATION OF PL NUMBERS

TEAM NAME: SHAREWARE_MAGNETS

INSTITUTE: I.I.T. KHARAGPUR

OUR UNDERSTANDING

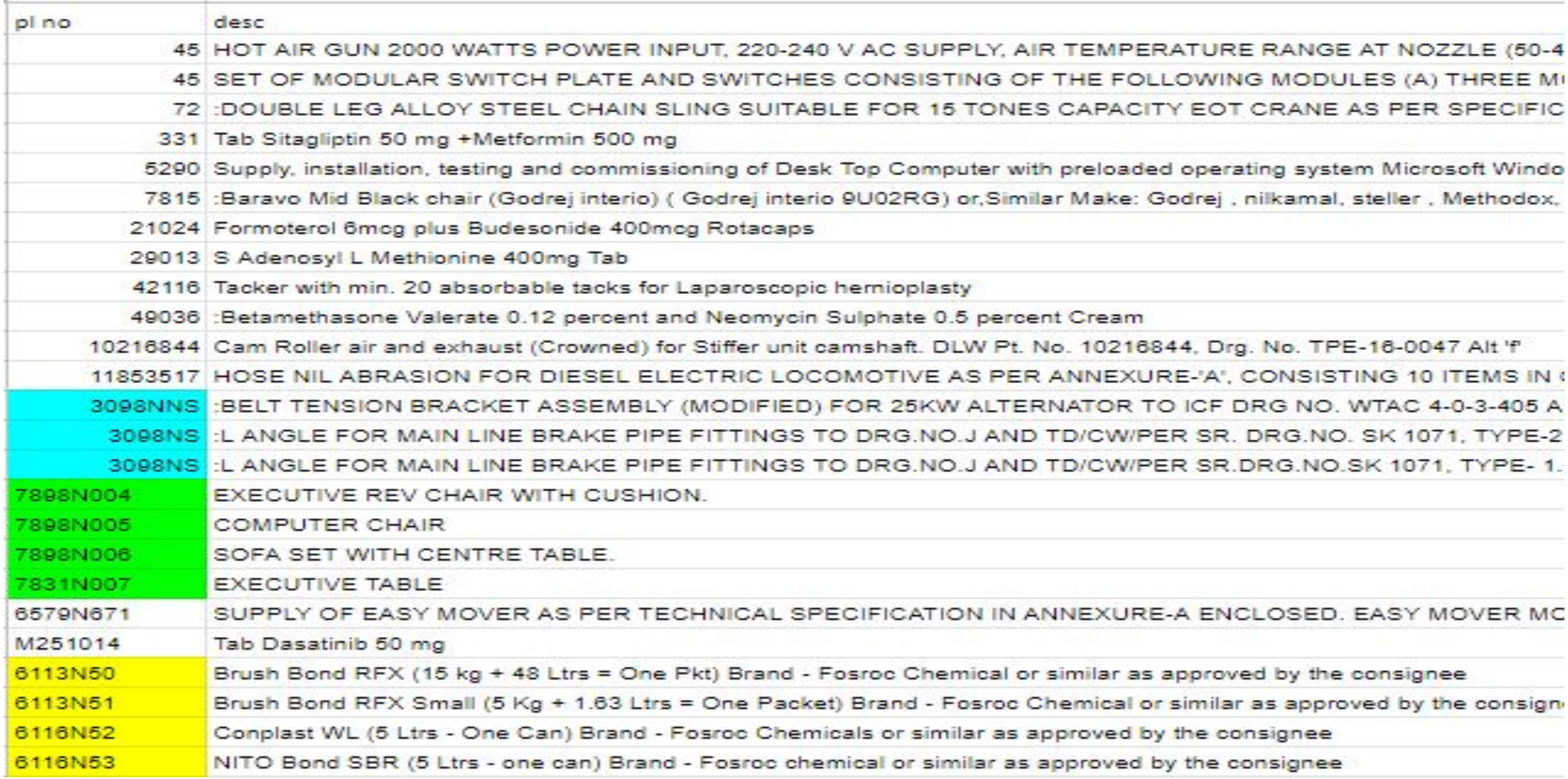
To every item Railway procures, a PL Number is allotted to it. Previously the database system was not centralized and various intermediate bodies had the freedom to allot PL Numbers. This led to

- Same PL Numbers - Different Descriptions
- Same PL Numbers - Same Descriptions
- Similar PL Numbers - Different Descriptions
- Different PL Numbers - Same Descriptions
- Different PL Numbers - Similar Descriptions
- Different PL Numbers - Different Descriptions

Initial Approach

- Used NLP to filter the descriptions
- Data Preprocessing
  - Tokenized the description
  - Removed stop words
  - Stemming and Lemmatization
  - Spell checker & Case checker
- Used tf-idf to transform the descriptions into vectors and calculated the cosine similarity between them.
- Tested our approach on the sample dataset which didn’t give the results as expected by observation.

Initial data analysis and observations



pl no	desc
45	HOT AIR GUN 2000 WATTS POWER INPUT, 220-240 V AC SUPPLY, AIR TEMPERATURE RANGE AT NOZZLE (50-4
45	SET OF MODULAR SWITCH PLATE AND SWITCHES CONSISTING OF THE FOLLOWING MODULES (A) THREE M
72	:DOUBLE LEG ALLOY STEEL CHAIN SLING SUITABLE FOR 15 TONES CAPACITY EOT CRANE AS PER SPECIFIC
331	Tab Sitagliptin 50 mg +Metformin 500 mg
5290	Supply, installation, testing and commissioning of Desk Top Computer with preloaded operating system Microsoft Windo
7815	:Baravo Mid Black chair (Godrej interio) (Godrej interio 9U02RG) or,Similar Make: Godrej , nilkamal, steller , Methodox,
21024	Formoterol 6mcg plus Budesonide 400mcg Rotacaps
29013	S Adenosyl L Methionine 400mg Tab
42116	Tacker with min. 20 absorbable tacks for Laparoscopic hernioplasty
49036	:Betamethasone Valerate 0.12 percent and Neomycin Sulphate 0.5 percent Cream
10216844	Cam Roller air and exhaust (Crowned) for Stiffer unit camshaft. DLW Pt. No. 10216844, Drg. No. TPE-16-0047 Alt 'f'
11853517	HOSE NIL ABRASION FOR DIESEL ELECTRIC LOCOMOTIVE AS PER ANNEXURE-'A', CONSISTING 10 ITEMS IN
3098NNS	:BELT TENSION BRACKET ASSEMBLY (MODIFIED) FOR 25KW ALTERNATOR TO ICF DRG NO. WTAC 4-0-3-405 A
3098NS	:L ANGLE FOR MAIN LINE BRAKE PIPE FITTINGS TO DRG.NO.J AND TD/CW/PER SR. DRG.NO. SK 1071, TYPE-2
3098NS	:L ANGLE FOR MAIN LINE BRAKE PIPE FITTINGS TO DRG.NO.J AND TD/CW/PER SR.DRG.NO.SK 1071, TYPE- 1.
7898N004	EXECUTIVE REV CHAIR WITH CUSHION.
7898N005	COMPUTER CHAIR
7898N006	SOFA SET WITH CENTRE TABLE.
7831N007	EXECUTIVE TABLE
6579N671	SUPPLY OF EASY MOVER AS PER TECHNICAL SPECIFICATION IN ANNEXURE-A ENCLOSED. EASY MOVER MO
M251014	Tab Dasatinib 50 mg
6113N50	Brush Bond RFX (15 kg + 48 Ltrs = One Pkt) Brand - Fosroc Chemical or similar as approved by the consignee
6113N51	Brush Bond RFX Small (5 Kg + 1.63 Ltrs = One Packet) Brand - Fosroc Chemical or similar as approved by the consign
6116N52	Conplast WL (5 Ltrs - One Can) Brand - Fosroc Chemicals or similar as approved by the consignee
6116N53	NITO Bond SBR (5 Ltrs - one can) Brand - Fosroc chemical or similar as approved by the consignee

Initial data analysis and observations

- We observed that the PL no. consists of large number of variations:
  - It may be any number ranging from 3 - 8 digits.
  - It may also be alphanumeric.
- In a 8- digit standardised PL Number
  - First 2 digits represents main group to which the item belongs.
    E.g. 20-30 represents loco-spare parts.
  - Next 2 digits represents sub-group of the item.
  - Next 3 digits represents serial no. of item within the sub-group.
- In case the PL no. is of 3,4 or 5 digits, we either need more data or the significance of the digits.
- Threshold Analysis for Description(Based on hit and trial on dataset):
  - For similar descriptions : 0.7 < similarity score < 0.9
  - For very similar descriptions(difference in size, etc.): similarity score > 0.9

Improvisation

- We used **Word2Vec** to convert each word of the description to a vector and thus calculated the vector for the entire description.
- Then used Cosine similarity to calculate the similarity score between the descriptions.
- We also took into consideration the similarity of the PL Numbers along with description similarity.
- Some descriptions consist of both words and numbers, but Word2Vec can only handle words. So we separated both the parts and then took the intersection of numbers in both the descriptions to get the similarity score of numbers. The main examples being materials which are in different sizes. They only have numeric differences in their description and 100% textual similarity. These data can be sorted this way.
- We took a weighted sum of both the similarity scores(word and numbers) for getting the final similarity score.

Similarity Matrix

- Created an NxN similarity matrix for N descriptions.
- Created a list for each PL Number, of descriptions which have similarity scores greater than the threshold 1(determined by trial and error on the data set)
- Assigned the current PL Number to the description from the list which has the highest similarity score with the current description.
- Removed assigned description from the list.
- If the list had more than one element this would be a case of **different pl numbers** but **similar descriptions.**
- Now each element of the list has 2 possibilities:

1. **First 2 digits of the PL Numbers are different:**
    a. If threshold 1 < similarity score < threshold 2:
        i. Assign first 4 digits of PL1 to new PL no. and **alpha-numeric** to rest of the digits

If similarity score> threshold 2:

Assign first 6 digits of PL1 to new PL no. and **alpha-numeric** to rest of the digits.

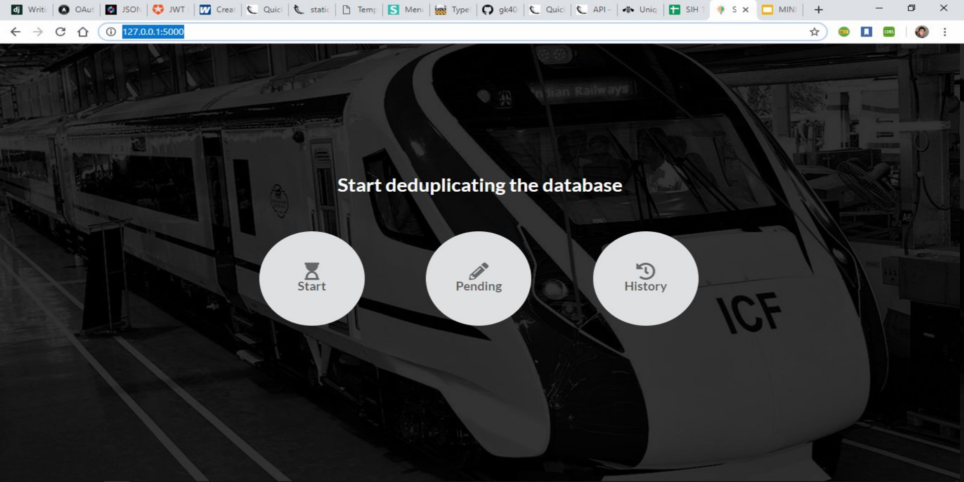
2. First 2 digits of the PL Numbers are same:

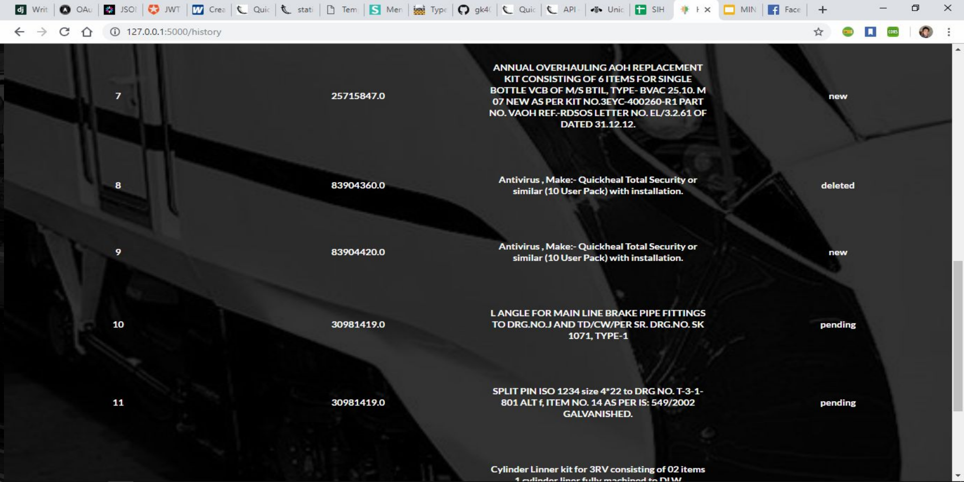
-> If next 2 are also same, assign PL2 to the description.

-> If next 2 are different:

If threshold 1 < similarity score < threshold 2 - assign PL2

If threshold 2 < similarity score - assign first 6 digits of PL1 to new PL





7	25715847.0	ANNUAL OVERHAULING AOH REPLACEMENT KIT CONSISTING OF 6 ITEMS FOR SINGLE BOTTLE VCB OF M/S BTIL, TYPE- BVAC 25.10. M 07 NEW AS PER KIT NO.3EYC-400260-R1 PART NO. VAOH REF.-RDSOS LETTER NO. EL/3.2.61 OF DATED 31.12.12.	new
8	83904360.0	Antivirus , Make:- Quickheal Total Security or similar (10 User Pack) with installation.	deleted
9	83904420.0	Antivirus , Make:- Quickheal Total Security or similar (10 User Pack) with installation.	new
10	30981419.0	L ANGLE FOR MAIN LINE BRAKE PIPE FITTINGS TO DRG.NO.J AND TD/CW/PER SR. DRG.NO. SK 1071, TYPE-1	pending
11	30981419.0	SPLIT PIN ISO 1234 size 4*22 to DRG NO. T-3-1-801 ALT f, ITEM NO. 14 AS PER IS: 549/2002 GALVANISHED.	pending
		Cylinder Linner kit for 3RV consisting of 02 items 1 cylinder liner fully machined to DLW	

Same PL different descriptions

Send the descriptions into pending case.

#Pending cases are too difficult to differentiate which requires human intervention.We have built an **interactive user interface** offering semi-automation for the pending cases using human intervention.

#These need to be dealt with Machine Learning or Deep Learning Techniques on the already created database.

Same PL numbers and similar descriptions:

Send the descriptions into pending. Unless the descriptions match 100%, we are unsure whether they are same product or similar product. Hence assignment of PL Number is problematic.

Different PL numbers and same descriptions

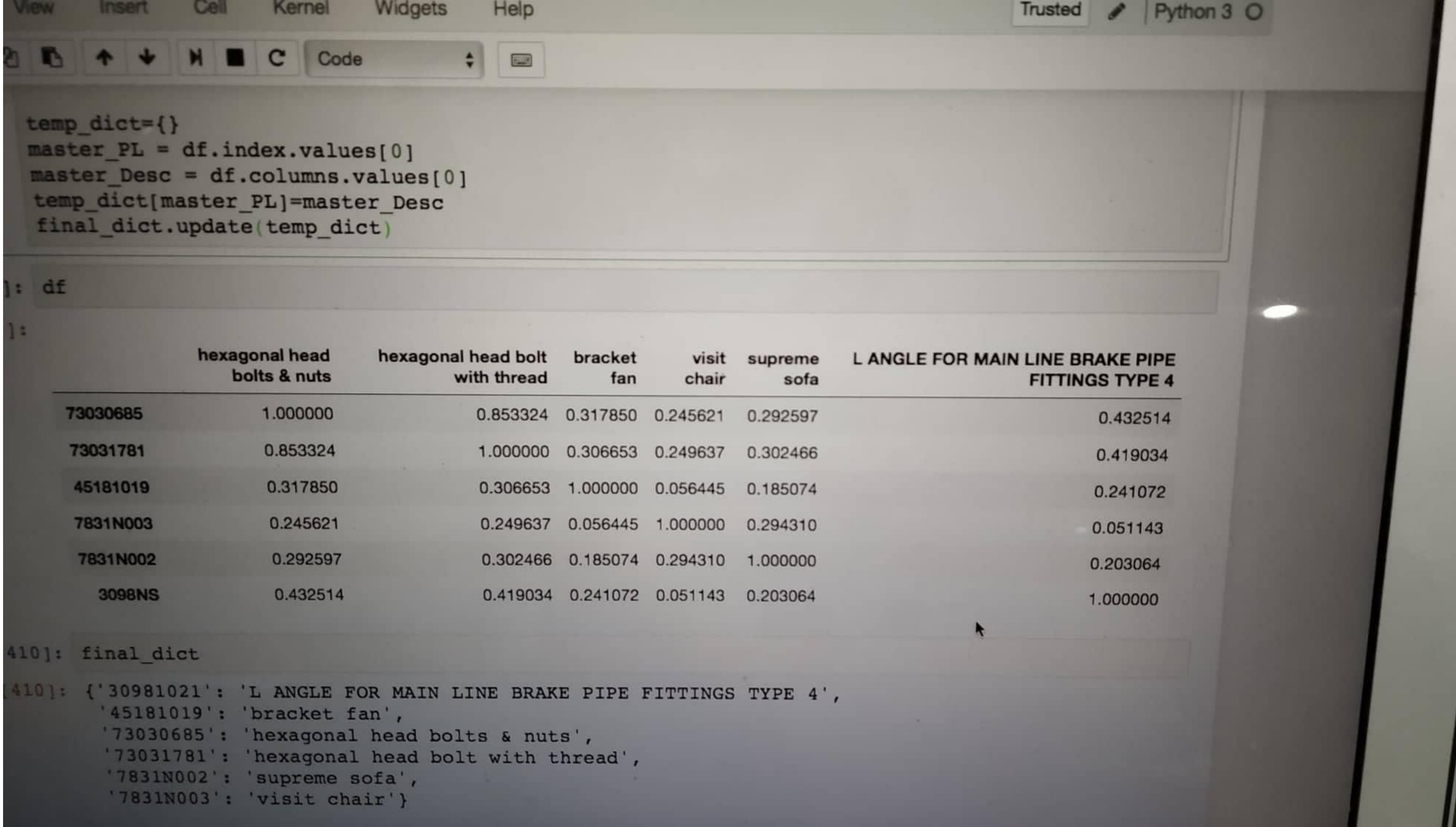
Solved earlier. We have fixed one PL number as a basis of our database and removed the other descriptions.

Assumptions / Constraints

- Limited knowledge about the categorization of PL Numbers : Main group and Sub group
- The size of data is very less to implement Supervised Machine Learning methods.
- Standard PL Number is of 8 digits and since we have PL numbers of variable lengths in the scraped data, we assumed that first 4 digits of PL Number represents main group and sub group .
- The descriptions may be sometimes either too large or refers to some other annexures.

Possible Future development

- Due to high randomness and unavailability of data the problem statement is only partially solvable.
  Considering our accumulated data [IREPS.Gov.in] :
- There are PL nos. < 6 digits, but some of them are clearly depicting medicinal products. To club and allocate PL no. to these products we need the information regarding the PL no. nomenclature (i.e. idea of main groups or subgroups that Ministry of Railways has) and a detailed database of significant keywords to retrieve and apply relevant machine learning models.
- Some of the data sets go to the pending section because of same pl numbers. We can automate such cases by using our newly created database to train our model by using Machine Learning techniques.



```
temp_dict={}
master_PL = df.index.values[0]
master_Desc = df.columns.values[0]
temp_dict[master_PL]=master_Desc
final_dict.update(temp_dict)
```

```
]: df
```

```
]:
```

	hexagonal head bolts & nuts	hexagonal head bolt with thread	bracket fan	visit chair	supreme sofa	L ANGLE FOR MAIN LINE BRAKE PIPE FITTINGS TYPE 4
73030685	1.000000	0.853324	0.317850	0.245621	0.292597	0.432514
73031781	0.853324	1.000000	0.306653	0.249637	0.302466	0.419034
45181019	0.317850	0.306653	1.000000	0.056445	0.185074	0.241072
7831N003	0.245621	0.249637	0.056445	1.000000	0.294310	0.051143
7831N002	0.292597	0.302466	0.185074	0.294310	1.000000	0.203064
3098NS	0.432514	0.419034	0.241072	0.051143	0.203064	1.000000

```
410]: final_dict
```

```
[410]: {'30981021': 'L ANGLE FOR MAIN LINE BRAKE PIPE FITTINGS TYPE 4',
        '45181019': 'bracket fan',
        '73030685': 'hexagonal head bolts & nuts',
        '73031781': 'hexagonal head bolt with thread',
        '7831N002': 'supreme sofa',
        '7831N003': 'visit chair'}
```