databricks part2

(https://databricks.com)

Problemn answers

Question 1

solution

```
skill2vec_list.distinct().count()
```

```
Out[180]: 50000
```

```
%sql
SELECT COUNT(DISTINCT id) AS number_of_job_descriptions
FROM skill2vec_list
```

Table

	number_of_job_descriptions
1	50000

1 row

discussion

```
%sql
SELECT skill_list
      , COUNT(*)
FROM skill2vec_list
GROUP BY skill_list
ORDER BY COUNT(*) DESC
```

Table

	skill_list	count(1)
1	["Human Resource"]	174
2	["WordPress", "Purchase Vendor Development"]	56
3	["Consulting"]	50
4	["Call Centre", "call center", "bpo", "customer support executive", "customer service executive", "call center bpo"]	32
5	["International BPO", "Salary", "Jobs", "Inbound", "Fresher", "Communication", "Day Shif", "Non Voice", "Hiring", "Domestic BPO", "Bpo", "Call Center", "Voice Process"]	29
6	["ITES", "IT Help Desk", "international voice", "Customer Service", "Operations", "Technical Services", "Bpo Hiring", "BPO", "Technical Support"]	27
7	["Reading Engineering Drawing, ..."]	26

10,000 rows | Truncated data

```
%sql
SELECT COUNT(DISTINCT skill_list) AS number_of_job_descriptions
FROM skill2vec_list
```

Table

	number_of_job_descriptions
1	46959

1 row

Question 2

solution

```
skill2vec_processed.groupBy('skill').count().sort(col('count').desc()).show(10)
```

```
+--------------------+-----+
|               skill|count|
+--------------------+-----+
|                Java| 1911|
|          Javascript| 1770|
|               Sales| 1705|
|Business Development| 1545|
|    Web Technologies| 1313|
|Communication Skills| 1305|
|         development| 1238|
|           Marketing| 1184|
|             Finance| 1078|
|                HTML| 1067|
+--------------------+-----+
only showing top 10 rows
```

```
%sql
SELECT skill
      , COUNT(*) AS freq
FROM skill2vec_processed
GROUP BY skill
ORDER BY freq DESC
LIMIT 10
```

Table

	skill	freq
1	Java	1911
2	Javascript	1770
3	Sales	1705
4	Business Development	1545
5	Web Technologies	1313
6	Communication Skills	1305
7	development	1238
8	Marketing	1184
9	Finance	1078
10	HTML	1067

10 rows

discussion

```
%sql
SELECT skill_replaced
      , COUNT(*) AS freq
FROM skill2vec_replaced
GROUP BY skill_replaced
ORDER BY freq DESC
LIMIT 10
```

Table

	skill_replaced	freq
1	Java	1911
2	Javascript	1770
3	Sales	1706
4	Business Development	1545
5	Web Technologies	1313
6	Communication Skills	1305

7	development	1238
8	Marketing	1185
9	Finance	1078
10	HTML	1067

10 rows

```
%sql
SELECT *
FROM skill2vec_processed
WHERE skill LIKE "%Marketing%" OR skill LIKE "%Sales%"
ORDER BY skill
```

Table

	skill	skill_lower
1	&amp;quot ; Assessment Sales&amp;quot ;	&amp;quot ; assessment sales&amp;quot ;
2	&amp;quot ; Insurance Sales&amp;quot ;	&amp;quot ; insurance sales&amp;quot ;
3	&amp;quot ; Marketing&amp;quot ;	&amp;quot ; marketing&amp;quot ;
4	&amp;quot ; Regional Sales Manager&amp;quot ;	&amp;quot ; regional sales manager&amp;quot ;
5	&amp;quot ; Sales	&amp;quot ; sales
6	&amp;quot ; Sales Executive	&amp;quot ; sales executive
7	- Sales &amp; Marketing	- sales &amp; marketing

10,000 rows | Truncated data

Question 3

solution

```
skill2vec_list.groupBy('size').count().sort(col('count').desc()).show(5)

+----+-----+
|size|count|
+----+-----+
|  10|10477|
|   5| 3432|
|   6| 3405|
|   1| 3386|
|   7| 3345|
+----+-----+
only showing top 5 rows
```

```
%sql
SELECT size AS number_of_skills
      , COUNT(*) AS freq
FROM skill2vec_list
GROUP BY size
ORDER BY freq DESC
LIMIT 5
```

Table

	number_of_skills	freq
1	10	10477
2	5	3432
3	6	3405
4	1	3386
5	7	3345

5 rows

discussion

```
%sql
SELECT skill_list
FROM skill2vec_list
WHERE size = 1
```

Table

	skill_list
1	["Storage Administrator"]
2	["sap sd"]
3	["Production Merchandiser Woven Nift &amp; Pearl"]
4	["Oracle EBS Developer"]
5	["IT Software - E - Commerce"]
6	["r&amp; d"]
7	["Production Department"]
8	["Technology"]
9	["Storage Administrator"]
10	["Content Writing"]
11	["Project Officer- Bal Ashram"]
12	["Junior Web Designer"]
13	["housekeeping"]
14	["MS Office"]

3,386 rows

```
%sql
SELECT skill_list
FROM skill2vec_list
WHERE size = 10
```

Table

	skill_list
1	["diploma", "machining", "cnc m", "mould", "conventional machines", "die making", "knowledge", "tool", "cipet", "assembly"]
2	["Compensation", "Benefits", "HR Functions", "Alm", "Payroll", "ESS", "Core HR", "QC", "QA", "SQL"]
3	["development", "information technology", "api", "business intelligence", "problem solving", "quality assurance", "soa", "siebel", "informatica", "microsoft certified"]
4	["Investment Banking", "Secretarial Activities", "Accounting", "Business Finance", "Company Secretary", "Auditing", "Taxation", "Credit Risk", "Risk Management", "Credit Control"]
5	["handling", "articles", "be", "i e", "3b2", "knowledge", "supervision", "check", "xml", "composing"]
6	["windows systems", "system configuration", "2010", "2013", "Office 365 and Exchange", "Systems Administrator", "Microsoft Exchange Servers", "Microsoft Exchange 2007", "Exchange Administration", "Exchange 2010"]
7	["good communication skill", "experience", "field", "area", "skills", "markets", "business development", "marketing", "plan", "strategies"]
8	["Financial Analysis", "Agri Finance", "Housing Finance", "Communication Skills", "Structured Finance", "Sales", "Banking", "Debt Syndication", "Commercial Vehicle", "Financial Services"]
9	["Business Development Management", "Communication Skills", "Sales", "Marketing", "Bde", "Brand Building", "Campaigns",

10,000 rows | Truncated data

Question 4

solution

```
skill2vec_processed.groupBy('skill_lower').count().sort(col('count').desc()).show(10)
```

```
+--------------------+-----+
|         skill_lower|count|
+--------------------+-----+
|                java| 2759|
|          javascript| 2738|
|               sales| 2680|
|business development| 2108|
|           marketing| 1809|
|                 sql| 1564|
|              jquery| 1547|
|                html| 1539|
|communication skills| 1537|
|                 bpo| 1530|
+--------------------+-----+
only showing top 10 rows
```

```
%sql
SELECT skill_lower AS skill
      , COUNT(*) AS freq
FROM skill2vec_processed
GROUP BY skill_lower
ORDER BY freq DESC
LIMIT 10
```

Table

	skill	freq
1	java	2759
2	javascript	2738
3	sales	2680
4	business development	2108
5	marketing	1809
6	sql	1564
7	jquery	1547
8	html	1539
9	communication skills	1537
10	bpo	1530

10 rows

discussion

```
%sql
SELECT skill_replaced_lower AS skill
      , COUNT(*) AS freq
FROM skill2vec_replaced
GROUP BY skill_replaced_lower
ORDER BY freq DESC
LIMIT 10
```

Table

	skill	freq
1	java	2759
2	javascript	2738
3	sales	2682
4	business development	2110
5	marketing	1810
6	sql	1564
7	jquery	1547
8	html	1539
9	communication skills	1537
10	bpo	1530

10 rows

Question 5

solution

```
skill2vec_processed.select('skill_lower').count()
```

```
Out[195]: 463803
```

```
%sql
SELECT COUNT(skill_lower) AS number_of_skills_before_join
FROM skill2vec_processed
```

Table

	number_of_skills_before_join
1	463803

1 row

```
skill2vec_processed.join(technology_processed, skill2vec_processed.skill_lower ==
technology_processed.example_lower).count()
```

```
Out[197]: 1101498
```

```
%sql
SELECT COUNT(s.skill_lower) AS number_of_skills_after_join
FROM skill2vec_processed s
  JOIN technology_processed t ON s.skill_lower = t.example_lower
```

Table

	number_of_skills_after_join
1	1101498

1 row

discussion

```
%sql
SELECT COUNT(skill_replaced_lower) AS number_of_skills_before_join
FROM skill2vec_replaced
```

Table

	number_of_skills_before_join
1	462847

1 row

```
%sql
SELECT COUNT(s.skill_replaced_lower) AS number_of_skills_after_join
FROM skill2vec_replaced s
     JOIN technology_processed t ON s.skill_replaced_lower = t.example_lower
```

Table

	number_of_skills_after_join
1	1101601

1 row

```
%sql
SELECT *
FROM skill2vec_processed
WHERE skill_lower = 'kubernetes'
```

Table

	skill	skill_lower
1	Kubernetes	kubernetes
2	Kubernetes	kubernetes
3	kubernetes	kubernetes

3 rows

```
%sql
SELECT *
FROM technology_processed
WHERE example_lower = 'kubernetes'
```

Table

	soc_code	example	commodity_code	commodity_title	hot_technology	in_demand	example_low
1	15-1252.00	Kubernetes	43232701	Application server software	Y	Y	kubernetes
2	15-1299.05	Kubernetes	43232701	Application server software	Y	Y	kubernetes
3	15-1299.07	Kubernetes	43232701	Application server software	Y	Y	kubernetes
4	15-1299.08	Kubernetes	43232701	Application server software	Y	Y	kubernetes
5	17-2112.02	Kubernetes	43232701	Application server software	Y	Y	kubernetes

5 rows

```
%sql
SELECT *
FROM technology_processed t
     JOIN skill2vec_processed s ON t.example_lower = s.skill_lower
WHERE t.example_lower = 'kubernetes'
```

Table

	soc_code	example	commodity_code	commodity_title	hot_technology	in_demand	example_low
1	15-1252.00	Kubernetes	43232701	Application server software	Y	Y	kubernetes
2	15-1252.00	Kubernetes	43232701	Application server software	Y	Y	kubernetes
3	15-1252.00	Kubernetes	43232701	Application server software	Y	Y	kubernetes
4	15-1299.05	Kubernetes	43232701	Application server software	Y	Y	kubernetes
5	15-1299.05	Kubernetes	43232701	Application server software	Y	Y	kubernetes
6	15-1299.05	Kubernetes	43232701	Application server software	Y	Y	kubernetes
7	15-1299.07	Kubernetes	43232701	Application server software	Y	Y	kubernetes
8	15-1299.07	Kubernetes	43232701	Application server software	Y	Y	kubernetes
9	15-1299.07	Kubernetes	43232701	Application server software	Y	Y	kubernetes
10	15-1299.08	Kubernetes	43232701	Application server software	Y	Y	kubernetes
11	15-1299.08	Kubernetes	43232701	Application server software	Y	Y	kubernetes
12	15-1299.08	Kubernetes	43232701	Application server software	Y	Y	kubernetes
13	17-2112.02	Kubernetes	43232701	Application server software	Y	Y	kubernetes
14	17-2112.02	Kubernetes	43232701	Application server software	Y	Y	kubernetes
15	17-2112.02	Kubernetes	43232701	Application server software	Y	Y	kubernetes

15 rows

Question 6

solution

```
skill2vec_processed.join(technology_processed, skill2vec_processed.skill_lower ==
technology_processed.example_lower).groupBy('commodity_title').count().sort(col('count').desc()).show(10)
```

```
+--------------------+------+
|     commodity_title| count|
+--------------------+------+
|Object or compone...|324521|
|Web platform deve...|298754|
|Operating system ...|190926|
|Development envir...| 53013|
|Data base managem...| 44132|
|Analytical or sci...| 33552|
|Web page creation...| 31682|
|Data base user in...| 29436|
|Spreadsheet software| 18568|
|File versioning s...| 13846|
+--------------------+------+
only showing top 10 rows
```

```
%sql
SELECT t.commodity_title
      , COUNT(*) AS freq
FROM skill2vec_processed s
  JOIN technology_processed t ON s.skill_lower = t.example_lower
GROUP BY t.commodity_title
ORDER BY freq DESC
LIMIT 10
```

Table

	commodity_title	freq
1	Object or component oriented development software	324521
2	Web platform development software	298754
3	Operating system software	190926
4	Development environment software	53013
5	Data base management system software	44132
6	Analytical or scientific software	33552
7	Web page creation and editing software	31682
8	Data base user interface and query software	29436
9	Spreadsheet software	18568
10	File versioning software	13846

10 rows

discussion

```
%sql
SELECT t.commodity_title
      , COUNT(*) AS freq
FROM skill2vec_replaced s
  JOIN technology_processed t ON s.skill_replaced_lower = t.example_lower
GROUP BY t.commodity_title
ORDER BY freq DESC
LIMIT 10
```

Table

	commodity_title	freq
1	Object or component oriented development software	324573
2	Web platform development software	298798
3	Operating system software	190930
4	Development environment software	53013

5	Data base management system software	44132
6	Analytical or scientific software	33552
7	Web page creation and editing software	31682
8	Data base user interface and query software	29439
9	Spreadsheet software	18568
10	File versioning software	13846

10 rows

Table

	commodity_title	skill_lower	freq
1	Object or component oriented development software	python	113740
2	Object or component oriented development software	c++	80242
3	Object or component oriented development software	c#	45656
4	Object or component oriented development software	jquery	35581
5	Object or component oriented development software	perl	28864
6	Object or component oriented development software	r	11954
7	Object or component oriented development software	objective c	3996
8	Object or component oriented development software	scala	3036
9	Object or component oriented development software	swift	1350

14 rows