SkillRecommendation

November 29, 2023

```python
[1]: from pyspark.ml import Pipeline
     from pyspark.ml.recommendation import ALS
     from pyspark.ml.evaluation import RegressionEvaluator
     from pyspark.sql import SparkSession
     from pyspark.sql.functions import col, array_contains
     from pyspark.sql.functions import monotonically_increasing_id, regexp_replace
     from pyspark.sql.functions import split, udf
     from pyspark.ml.feature import Tokenizer, CountVectorizer, StringIndexer
     from ast import literal_eval
     from pyspark.sql.types import ArrayType, StringType, FloatType
     from pyspark.sql.functions import explode, col
```

```python
[2]: # Create a Spark session
     spark = SparkSession.builder.appName("CourseRecommendation").getOrCreate()

     # Load courses data into a DataFrame
     courses_df = spark.read.option("multiLine", True).csv("coursera_courses.csv",
      ↪header=True, inferSchema=True).withColumn('course_description',
      ↪regexp_replace('course_description', '\n', ' '))

     # Load the courses mapping
     courses_mapping_df = spark.read.csv("Coursera_courses_mappings.csv",
      ↪header=True, inferSchema=True).drop("course_url")

     # Load the course reviews dataset
     reviews_df = spark.read.csv("Coursera_reviews.csv", header=True,
      ↪inferSchema=True)

     # Show the first few rows of the DataFrame
     print(f"Courses:\n{courses_df.head(2)}")

     print(f"\nMappings:\n{courses_mapping_df.head(2)}")

     print(f"\nReviews:\n{reviews_df.head(2)}")
```

```
Courses:
[Row(course_title='(ISC)² Systems Security Certified Practitioner (SSCP)',
course_organization='ISC2', course_certificate_type='Specialization',
```

course_time='3 - 6 Months', course_rating='4.7', course_reviews_num='492',
course_difficulty='Beginner',
course_url='https://www.coursera.org/specializations/sscp-training',
course_students_enrolled='6,958', course_skills="['Risk Management', 'Access
Control', 'Asset', 'Incident Detection and Response', 'Cloud Computing
Security', 'Wireless Security', 'Security Software']", course_summary='[]',
course_description='Pursue better IT security job opportunities and prove
knowledge with confidence. The SSCP Professional Training Certificate shows
employers you have the IT security foundation to defend against cyber attacks –
and puts you on a clear path to earning SSCP certification. Learn on your own
schedule with 120-day access to content aligned with the latest (ISC)2 SSCP exam
domains. We're offering the complete online self-paced program for only $1,000 –
a $200 savings when you get all domains bundled together. 3 Steps to Career
Advancement 1. Register for the course 2. Gain access for 120 days 3.
Register and sit for the SSCP certification exam Upon completing the SSCP
Professional Certificate, you will: 1. Complete six courses of preparing you to
sit for the Systems Security Certified Practitioner (SSCP) certification exam
Opens in a new tabas outlined below. Course 1 – Access Controls Course 2 –
Security Operations and Administration Course 3 – Risk Identification,
Monitoring, and Analysis/Incident Response and Recovery Course 4 – Cryptography
Course 5 – Network and Communication Security Course 6 – Systems and Application
Security 2. Receive a certificate of program completion. 3. Understand how to
implement, monitor and administer an organization's IT infrastructure in
accordance with security policies and procedures that ensure data
confidentiality, integrity and availability. Applied Learning Project Each
course includes a case study that will require students to put into practice the
knowledge they have gained throughout each course. Successful completion of
course projects will require the basic understanding of the topics covered and
the ability to relate those topics to the real world. The objective of each
project is to determine whether students have understood course concepts and are
able to use them in a real world setting. '), Row(course_title='.NET FullStack
Developer', course_organization='Board Infinity',
course_certificate_type='Specialization', course_time='1 - 3 Months',
course_rating='4.3', course_reviews_num='51', course_difficulty='Intermediate',
course_url='https://www.coursera.org/specializations/dot-net-fullstack',
course_students_enrolled='2,531', course_skills="['Web API', 'Web Development',
'Cascading Style Sheets (CSS)', 'HTML', 'React (Web Framework)', 'RESTful APIs',
'C# programming', 'API Integration', 'Model-View-Controller (MVC)', 'asp.net']",
course_summary='"[\'Master .NET full stack web dev: from .NET core basics to
advanced web application creation using .NET technologies. \', ""Dive into
React frontend development: from foundational HTML & CSS to dynamic UI design
using JavaScript and React\'s core concepts. ""', course_description=" 'Explore
backend with ASP.NET: Understand ASP.NET Core architecture")]

Mappings:
[Row(name='Machine Learning', institution='Stanford University',
course_id='machine-learning'), Row(name='Indigenous Canada',
institution='University of Alberta', course_id='indigenous-canada')]

Reviews:
[Row(reviews="Pretty dry, but I was able to pass with just two complete watches so I'm happy about that. As usual there were some questions on the final exam that were NO WHERE in the course, which is annoying but far better than many microsoft tests I have taken. Never found the suplimental material that the course references… but who cares… i passed!", reviewers='By Robert S', date_reviews='Feb 12, 2020', rating='4', course_id='google-cbrs-cpi-training'), Row(reviews='would be a better experience if the video and screen shots would sho on the side of the text that the instructor is going thru so that user does not have to go all the way to beginning of text to be able to view any slides instructor is showing.', reviewers='By Gabriel E R', date_reviews='Sep 28, 2020', rating='4', course_id='google-cbrs-cpi-training')]

```python
[3]: # Data Preprocessing

# Add a user_id column for Collaborative Filtering
courses_df = courses_df.withColumn('user_id', monotonically_increasing_id())

# Convert 'course_rating' column to float
courses_df = courses_df.withColumn('course_rating', col('course_rating').
 ↪cast('float'))

# Feature Engineering: StringIndexer for user and item (course) columns
course_indexer = StringIndexer(inputCol="course_id", outputCol="courseIndex")

# Convert the 'rating' column in reviews_df to float
reviews_df = reviews_df.withColumn('rating', reviews_df['rating'].
 ↪cast(FloatType()))

courses_df.head(2)
```

[3]: [Row(course_title='(ISC)² Systems Security Certified Practitioner (SSCP)', course_organization='ISC2', course_certificate_type='Specialization', course_time='3 - 6 Months', course_rating=4.699999809265137, course_reviews_num='492', course_difficulty='Beginner', course_url='https://www.coursera.org/specializations/sscp-training', course_students_enrolled='6,958', course_skills="['Risk Management', 'Access Control', 'Asset', 'Incident Detection and Response', 'Cloud Computing Security', 'Wireless Security', 'Security Software']", course_summary='[]', course_description='Pursue better IT security job opportunities and prove knowledge with confidence. The SSCP Professional Training Certificate shows employers you have the IT security foundation to defend against cyber attacks - and puts you on a clear path to earning SSCP certification. Learn on your own schedule with 120-day access to content aligned with the latest (ISC)2 SSCP exam domains. We're offering the complete online self-paced program for only $1,000 - a $200 savings when you get all domains bundled together. 3 Steps to Career

Advancement 1. Register for the course 2. Gain access for 120 days 3. Register and sit for the SSCP certification exam Upon completing the SSCP Professional Certificate, you will: 1. Complete six courses of preparing you to sit for the Systems Security Certified Practitioner (SSCP) certification exam Opens in a new tabas outlined below. Course 1 – Access Controls Course 2 – Security Operations and Administration Course 3 – Risk Identification, Monitoring, and Analysis/Incident Response and Recovery Course 4 – Cryptography Course 5 – Network and Communication Security Course 6 – Systems and Application Security 2. Receive a certificate of program completion. 3. Understand how to implement, monitor and administer an organization's IT infrastructure in accordance with security policies and procedures that ensure data confidentiality, integrity and availability. Applied Learning Project Each course includes a case study that will require students to put into practice the knowledge they have gained throughout each course. Successful completion of course projects will require the basic understanding of the topics covered and the ability to relate those topics to the real world. The objective of each project is to determine whether students have understood course concepts and are able to use them in a real world setting. ', user_id=0),
 Row(course_title='.NET FullStack Developer', course_organization='Board Infinity', course_certificate_type='Specialization', course_time='1 – 3 Months', course_rating=4.300000190734863, course_reviews_num='51', course_difficulty='Intermediate', course_url='https://www.coursera.org/specializations/dot-net-fullstack', course_students_enrolled='2,531', course_skills="['Web API', 'Web Development', 'Cascading Style Sheets (CSS)', 'HTML', 'React (Web Framework)', 'RESTful APIs', 'C# programming', 'API Integration', 'Model-View-Controller (MVC)', 'asp.net']", course_summary='"[\'Master .NET full stack web dev: from .NET core basics to advanced web application creation using .NET technologies. \', ""Dive into React frontend development: from foundational HTML & CSS to dynamic UI design using JavaScript and React\'s core concepts. ""', course_description=" 'Explore backend with ASP.NET: Understand ASP.NET Core architecture", user_id=1)]

```python
print(f"Courses Dataset Data Types:\n{courses_df.dtypes}")
print(f"\nCourses Mappings Dataset Data Types:\n{courses_mapping_df.dtypes}")
print(f"\nReviews Dataset Data Types:\n{reviews_df.dtypes}")
```

Courses Dataset Data Types:
[('course_title', 'string'), ('course_organization', 'string'), ('course_certificate_type', 'string'), ('course_time', 'string'), ('course_rating', 'float'), ('course_reviews_num', 'string'), ('course_difficulty', 'string'), ('course_url', 'string'), ('course_students_enrolled', 'string'), ('course_skills', 'string'), ('course_summary', 'string'), ('course_description', 'string'), ('user_id', 'bigint')]

Courses Mappings Dataset Data Types:
[('name', 'string'), ('institution', 'string'), ('course_id', 'string')]

```
Reviews Dataset Data Types:
[('reviews', 'string'), ('reviewers', 'string'), ('date_reviews', 'string'),
('rating', 'float'), ('course_id', 'string')]
```

[5]: `reviews_df.show()`

```
+------------------+------------+------------+------+------------------+
|           reviews|   reviewers|date_reviews|rating|         course_id|
+------------------+------------+------------+------+------------------+
|Pretty dry, but I…|  By Robert S|Feb 12, 2020|   4.0|google-cbrs-cpi-t…|
|would be a better…|By Gabriel E R|Sep 28, 2020|   4.0|google-cbrs-cpi-t…|
|Information was p…|   By Jacob D|Apr 08, 2020|   4.0|google-cbrs-cpi-t…|
|A few grammatical…|    By Dale B|Feb 24, 2020|   4.0|google-cbrs-cpi-t…|
|Excellent course …|    By Sean G|Jun 18, 2020|   4.0|google-cbrs-cpi-t…|
|Some of the quizz…|  By Daniel F|Dec 23, 2019|   4.0|google-cbrs-cpi-t…|
|Solid presentatio…|   By Logan D|Sep 03, 2020|   5.0|google-cbrs-cpi-t…|
|Probably the best…|  By Luis M C|Nov 21, 2019|   5.0|google-cbrs-cpi-t…|
|The ProctorU.com …|   By scott w|Sep 28, 2020|   5.0|google-cbrs-cpi-t…|
|Covered all of th…|    By Ryan H|Aug 26, 2019|   5.0|google-cbrs-cpi-t…|
|Great course, lec…|  By Samuel D|Jan 24, 2020|   5.0|google-cbrs-cpi-t…|
|The course was st…|  By Scot A W|May 01, 2020|   5.0|google-cbrs-cpi-t…|
|Pretty well desig…|By Jonathan L|Jan 17, 2020|   5.0|google-cbrs-cpi-t…|
|Great course, i h…|   By Geoff T|Jul 21, 2019|   5.0|google-cbrs-cpi-t…|
|The course was fi…|By William J|Jun 07, 2019|   5.0|google-cbrs-cpi-t…|
|Well presented co…|   By Ariel R|Sep 27, 2019|   5.0|google-cbrs-cpi-t…|
|Course is easy en…|  By Austin N|Dec 27, 2019|   5.0|google-cbrs-cpi-t…|
|The instructions …|By Jayson P J|Nov 22, 2019|   5.0|google-cbrs-cpi-t…|
|Easy to follow. D…|  By Robert S|Nov 07, 2019|   5.0|google-cbrs-cpi-t…|
|Excellent Trainin…|   By Scott C|Sep 02, 2020|   5.0|google-cbrs-cpi-t…|
+------------------+------------+------------+------+------------------+
only showing top 20 rows
```

[6]:
```python
# ALS Collaborative Filtering Model
als = ALS(maxIter=10, regParam=0.01, userCol="user_id", itemCol="courseIndex",
          ratingCol="course_rating")
```

[7]:
```python
# Content-based recommendation pipeline
tokenizer = Tokenizer(inputCol="course_skills", outputCol="words")
vectorizer = CountVectorizer(inputCol="words", outputCol="features")
```

[8]:
```python
combined_data = reviews_df.join(courses_mapping_df, "course_id").join(
    courses_df, (courses_mapping_df["name"] == courses_df["course_title"])
)
```

[9]: `combined_data.dtypes`

```
[9]: [('course_id', 'string'),
     ('reviews', 'string'),
     ('reviewers', 'string'),
     ('date_reviews', 'string'),
     ('rating', 'float'),
     ('name', 'string'),
     ('institution', 'string'),
     ('course_title', 'string'),
     ('course_organization', 'string'),
     ('course_certificate_type', 'string'),
     ('course_time', 'string'),
     ('course_rating', 'float'),
     ('course_reviews_num', 'string'),
     ('course_difficulty', 'string'),
     ('course_url', 'string'),
     ('course_students_enrolled', 'string'),
     ('course_skills', 'string'),
     ('course_summary', 'string'),
     ('course_description', 'string'),
     ('user_id', 'bigint')]
```

```
[10]: combined_data.show()
```

```
+------------------+------------------+------------+----------+------+--
----------------+------------------+------------------+------------------
-+--------------------+-----------+-----------+------------------+-------
---------+------------------+--------------------+------------+----------
----------+------------------+-------+
|         course_id|           reviews|   reviewers|date_reviews|rating|
name|       institution|      course_title|
course_organization|course_certificate_type|
course_time|course_rating|course_reviews_num|course_difficulty|
course_url|course_students_enrolled|course_skills|     course_summary|
course_description|user_id|
+------------------+------------------+------------+----------+------+--
----------------+------------------+------------------+------------------
-+--------------------+-----------+-----------+------------------+-------
---------+------------------+--------------------+------------+----------
----------+------------------+-------+
|google-cbrs-cpi-t…|Pretty dry, but I…|  By Robert S|Feb 12, 2020|
4.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|          Course|1 - 3 Months|          4.6|
3.4k|           Mixed|https://www.cours…|                 9,008|
[]|['Bayesian Infere…|This Specializati…|     96|
|google-cbrs-cpi-t…|would be a better…|By Gabriel E R|Sep 28, 2020|
4.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|          Course|1 - 3 Months|          4.6|
3.4k|           Mixed|https://www.cours…|                 9,008|
```

```
[]|['Bayesian Infere…|This Specializati…|      96|
|google-cbrs-cpi-t…|Information was p…|    By Jacob D|Apr 08, 2020|
4.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|            Course|1 - 3 Months|        4.6|
3.4k|             Mixed|https://www.cours…|              9,008|
[]|['Bayesian Infere…|This Specializati…|      96|
|google-cbrs-cpi-t…|A few grammatical…|    By Dale B|Feb 24, 2020|
4.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|            Course|1 - 3 Months|        4.6|
3.4k|             Mixed|https://www.cours…|              9,008|
[]|['Bayesian Infere…|This Specializati…|      96|
|google-cbrs-cpi-t…|Excellent course …|    By Sean G|Jun 18, 2020|
4.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|            Course|1 - 3 Months|        4.6|
3.4k|             Mixed|https://www.cours…|              9,008|
[]|['Bayesian Infere…|This Specializati…|      96|
|google-cbrs-cpi-t…|Some of the quizz…|  By Daniel F|Dec 23, 2019|
4.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|            Course|1 - 3 Months|        4.6|
3.4k|             Mixed|https://www.cours…|              9,008|
[]|['Bayesian Infere…|This Specializati…|      96|
|google-cbrs-cpi-t…|Solid presentatio…|   By Logan D|Sep 03, 2020|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|            Course|1 - 3 Months|        4.6|
3.4k|             Mixed|https://www.cours…|              9,008|
[]|['Bayesian Infere…|This Specializati…|      96|
|google-cbrs-cpi-t…|Probably the best…|  By Luis M C|Nov 21, 2019|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|            Course|1 - 3 Months|        4.6|
3.4k|             Mixed|https://www.cours…|              9,008|
[]|['Bayesian Infere…|This Specializati…|      96|
|google-cbrs-cpi-t…|The ProctorU.com …|   By scott w|Sep 28, 2020|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|            Course|1 - 3 Months|        4.6|
3.4k|             Mixed|https://www.cours…|              9,008|
[]|['Bayesian Infere…|This Specializati…|      96|
|google-cbrs-cpi-t…|Covered all of th…|    By Ryan H|Aug 26, 2019|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|            Course|1 - 3 Months|        4.6|
3.4k|             Mixed|https://www.cours…|              9,008|
[]|['Bayesian Infere…|This Specializati…|      96|
|google-cbrs-cpi-t…|Great course, lec…|  By Samuel D|Jan 24, 2020|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|            Course|1 - 3 Months|        4.6|
3.4k|             Mixed|https://www.cours…|              9,008|
[]|['Bayesian Infere…|This Specializati…|      96|
|google-cbrs-cpi-t…|The course was st…|  By Scot A W|May 01, 2020|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
```

```
Spectrum…|                Course|1 - 3 Months|        4.6|
3.4k|            Mixed|https://www.cours…|            9,008|
[]|['Bayesian Infere…|This Specializati…|    96|
|google-cbrs-cpi-t…|Pretty well desig…| By Jonathan L|Jan 17, 2020|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|                Course|1 - 3 Months|        4.6|
3.4k|            Mixed|https://www.cours…|            9,008|
[]|['Bayesian Infere…|This Specializati…|    96|
|google-cbrs-cpi-t…|Great course, i h…|    By Geoff T|Jul 21, 2019|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|                Course|1 - 3 Months|        4.6|
3.4k|            Mixed|https://www.cours…|            9,008|
[]|['Bayesian Infere…|This Specializati…|    96|
|google-cbrs-cpi-t…|The course was fi…|  By William J|Jun 07, 2019|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|                Course|1 - 3 Months|        4.6|
3.4k|            Mixed|https://www.cours…|            9,008|
[]|['Bayesian Infere…|This Specializati…|    96|
|google-cbrs-cpi-t…|Well presented co…|    By Ariel R|Sep 27, 2019|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|                Course|1 - 3 Months|        4.6|
3.4k|            Mixed|https://www.cours…|            9,008|
[]|['Bayesian Infere…|This Specializati…|    96|
|google-cbrs-cpi-t…|Course is easy en…|   By Austin N|Dec 27, 2019|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|                Course|1 - 3 Months|        4.6|
3.4k|            Mixed|https://www.cours…|            9,008|
[]|['Bayesian Infere…|This Specializati…|    96|
|google-cbrs-cpi-t…|The instructions …| By Jayson P J|Nov 22, 2019|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|                Course|1 - 3 Months|        4.6|
3.4k|            Mixed|https://www.cours…|            9,008|
[]|['Bayesian Infere…|This Specializati…|    96|
|google-cbrs-cpi-t…|Easy to follow. D…|   By Robert S|Nov 07, 2019|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|                Course|1 - 3 Months|        4.6|
3.4k|            Mixed|https://www.cours…|            9,008|
[]|['Bayesian Infere…|This Specializati…|    96|
|google-cbrs-cpi-t…|Excellent Trainin…|    By Scott C|Sep 02, 2020|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|                Course|1 - 3 Months|        4.6|
3.4k|            Mixed|https://www.cours…|            9,008|
[]|['Bayesian Infere…|This Specializati…|    96|
+------------------+------------------+-------------+-----------+------+--
----------------+------------------+------------------+----------------
-+--------------------+-----------+-----------+----------------+--------
---------+------------------+--------------------+-----------+---------
----------+------------------+-------+
```

only showing top 20 rows

```
[11]:  # Create a pipeline for content-based recommendations
       pipeline_content = Pipeline(stages=[course_indexer, tokenizer, vectorizer])
```

```
[12]:  # Fit and transform the pipeline to get course features
       content_model = pipeline_content.fit(combined_data)
       data_content = content_model.transform(combined_data).alias("content_data")
```

```
[13]:  # Print the column names in the dataframes
       print(courses_df.printSchema())

       print(courses_mapping_df.printSchema())

       print(reviews_df.printSchema())

       print(data_content.printSchema())
```

```
root
 |-- course_title: string (nullable = true)
 |-- course_organization: string (nullable = true)
 |-- course_certificate_type: string (nullable = true)
 |-- course_time: string (nullable = true)
 |-- course_rating: float (nullable = true)
 |-- course_reviews_num: string (nullable = true)
 |-- course_difficulty: string (nullable = true)
 |-- course_url: string (nullable = true)
 |-- course_students_enrolled: string (nullable = true)
 |-- course_skills: string (nullable = true)
 |-- course_summary: string (nullable = true)
 |-- course_description: string (nullable = true)
 |-- user_id: long (nullable = false)

None
root
 |-- name: string (nullable = true)
 |-- institution: string (nullable = true)
 |-- course_id: string (nullable = true)

None
root
 |-- reviews: string (nullable = true)
 |-- reviewers: string (nullable = true)
 |-- date_reviews: string (nullable = true)
 |-- rating: float (nullable = true)
 |-- course_id: string (nullable = true)
```

```
None
root
 |-- course_id: string (nullable = true)
 |-- reviews: string (nullable = true)
 |-- reviewers: string (nullable = true)
 |-- date_reviews: string (nullable = true)
 |-- rating: float (nullable = true)
 |-- name: string (nullable = true)
 |-- institution: string (nullable = true)
 |-- course_title: string (nullable = true)
 |-- course_organization: string (nullable = true)
 |-- course_certificate_type: string (nullable = true)
 |-- course_time: string (nullable = true)
 |-- course_rating: float (nullable = true)
 |-- course_reviews_num: string (nullable = true)
 |-- course_difficulty: string (nullable = true)
 |-- course_url: string (nullable = true)
 |-- course_students_enrolled: string (nullable = true)
 |-- course_skills: string (nullable = true)
 |-- course_summary: string (nullable = true)
 |-- course_description: string (nullable = true)
 |-- user_id: long (nullable = false)
 |-- courseIndex: double (nullable = false)
 |-- words: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- features: vector (nullable = true)

None
```

```python
# Create temporary views for DataFrames
data_content.createOrReplaceTempView("data_content")
reviews_df.createOrReplaceTempView("reviews_df")
courses_mapping_df.createOrReplaceTempView("courses_mapping_df")
courses_df.createOrReplaceTempView("courses_df")

# SQL query to join the DataFrames
query = """
    SELECT
        data_content.*,
        reviews_df.rating AS review_rating,
        courses_mapping_df.name,
        courses_mapping_df.institution,
        courses_mapping_df.course_url,
        courses_df.course_title,
        courses_df.course_organization,
        courses_df.course_certificate_type,
        courses_df.course_time,
```

```
        courses_df.course_rating,
        courses_df.course_reviews_num,
        courses_df.course_difficulty,
        courses_df.course_url AS course_url_combined,
        courses_df.course_students_enrolled,
        courses_df.course_skills,
        courses_df.course_summary,
        courses_df.course_description,
        courses_df.user_id
    FROM
        data_content
    LEFT JOIN
        reviews_df
    ON
        data_content.course_id = reviews_df.course_id
    LEFT JOIN
        courses_mapping_df
    ON
        data_content.course_id = courses_mapping_df.course_id
    LEFT JOIN
        courses_df
    ON
        courses_mapping_df.name = courses_df.course_title
"""

# Execute the SQL query
data_content = spark.sql(query)
```

[15]:
```
data_content.show()
```

```
+------------------+------------------+------------+-----------+------+--
----------------+------------------+------------------+------------------
-+--------------------+-----------+-----------+-----------------+--------
---------+-----------------+--------------------+------------+----------
----------+------------------+------+----------+-----+---------------+
|         course_id|           reviews|   reviewers|date_reviews|rating|
name|       institution|      course_title|
course_organization|course_certificate_type|
course_time|course_rating|course_reviews_num|course_difficulty|
course_url|course_students_enrolled|course_skills|     course_summary|
course_description|user_id|courseIndex|words|       features|
+------------------+------------------+------------+-----------+------+--
----------------+------------------+------------------+------------------
-+--------------------+-----------+-----------+-----------------+--------
---------+-----------------+--------------------+------------+----------
----------+------------------+------+----------+-----+---------------+
|google-cbrs-cpi-t…|Pretty dry, but I…|  By Robert S|Feb 12, 2020|
4.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
```

```
Spectrum…|                    Course|1 - 3 Months|         4.6|
3.4k|              Mixed|https://www.cours…|                  9,008|
[]|['Bayesian Infere…|This Specializati…|     96|     198.0|
[[]]|(746,[11],[1.0])|
|google-cbrs-cpi-t…|would be a better…|By Gabriel E R|Sep 28, 2020|
4.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|                    Course|1 - 3 Months|         4.6|
3.4k|              Mixed|https://www.cours…|                  9,008|
[]|['Bayesian Infere…|This Specializati…|     96|     198.0|
[[]]|(746,[11],[1.0])|
|google-cbrs-cpi-t…|Information was p…|   By Jacob D|Apr 08, 2020|
4.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|                    Course|1 - 3 Months|         4.6|
3.4k|              Mixed|https://www.cours…|                  9,008|
[]|['Bayesian Infere…|This Specializati…|     96|     198.0|
[[]]|(746,[11],[1.0])|
|google-cbrs-cpi-t…|A few grammatical…|   By Dale B|Feb 24, 2020|
4.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|                    Course|1 - 3 Months|         4.6|
3.4k|              Mixed|https://www.cours…|                  9,008|
[]|['Bayesian Infere…|This Specializati…|     96|     198.0|
[[]]|(746,[11],[1.0])|
|google-cbrs-cpi-t…|Excellent course …|   By Sean G|Jun 18, 2020|
4.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|                    Course|1 - 3 Months|         4.6|
3.4k|              Mixed|https://www.cours…|                  9,008|
[]|['Bayesian Infere…|This Specializati…|     96|     198.0|
[[]]|(746,[11],[1.0])|
|google-cbrs-cpi-t…|Some of the quizz…|  By Daniel F|Dec 23, 2019|
4.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|                    Course|1 - 3 Months|         4.6|
3.4k|              Mixed|https://www.cours…|                  9,008|
[]|['Bayesian Infere…|This Specializati…|     96|     198.0|
[[]]|(746,[11],[1.0])|
|google-cbrs-cpi-t…|Solid presentatio…|   By Logan D|Sep 03, 2020|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|                    Course|1 - 3 Months|         4.6|
3.4k|              Mixed|https://www.cours…|                  9,008|
[]|['Bayesian Infere…|This Specializati…|     96|     198.0|
[[]]|(746,[11],[1.0])|
|google-cbrs-cpi-t…|Probably the best…|  By Luis M C|Nov 21, 2019|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|                    Course|1 - 3 Months|         4.6|
3.4k|              Mixed|https://www.cours…|                  9,008|
[]|['Bayesian Infere…|This Specializati…|     96|     198.0|
[[]]|(746,[11],[1.0])|
|google-cbrs-cpi-t…|The ProctorU.com …|   By scott w|Sep 28, 2020|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
```

```
Spectrum…|            Course|1 - 3 Months|       4.6|
3.4k|           Mixed|https://www.cours…|           9,008|
[]|['Bayesian Infere…|This Specializati…|    96|     198.0|
[[]]|(746,[11],[1.0])|
|google-cbrs-cpi-t…|Covered all of th…|    By Ryan H|Aug 26, 2019|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|            Course|1 - 3 Months|       4.6|
3.4k|           Mixed|https://www.cours…|           9,008|
[]|['Bayesian Infere…|This Specializati…|    96|     198.0|
[[]]|(746,[11],[1.0])|
|google-cbrs-cpi-t…|Great course, lec…|    By Samuel D|Jan 24, 2020|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|            Course|1 - 3 Months|       4.6|
3.4k|           Mixed|https://www.cours…|           9,008|
[]|['Bayesian Infere…|This Specializati…|    96|     198.0|
[[]]|(746,[11],[1.0])|
|google-cbrs-cpi-t…|The course was st…|    By Scot A W|May 01, 2020|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|            Course|1 - 3 Months|       4.6|
3.4k|           Mixed|https://www.cours…|           9,008|
[]|['Bayesian Infere…|This Specializati…|    96|     198.0|
[[]]|(746,[11],[1.0])|
|google-cbrs-cpi-t…|Pretty well desig…| By Jonathan L|Jan 17, 2020|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|            Course|1 - 3 Months|       4.6|
3.4k|           Mixed|https://www.cours…|           9,008|
[]|['Bayesian Infere…|This Specializati…|    96|     198.0|
[[]]|(746,[11],[1.0])|
|google-cbrs-cpi-t…|Great course, i h…|    By Geoff T|Jul 21, 2019|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|            Course|1 - 3 Months|       4.6|
3.4k|           Mixed|https://www.cours…|           9,008|
[]|['Bayesian Infere…|This Specializati…|    96|     198.0|
[[]]|(746,[11],[1.0])|
|google-cbrs-cpi-t…|The course was fi…|  By William J|Jun 07, 2019|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|            Course|1 - 3 Months|       4.6|
3.4k|           Mixed|https://www.cours…|           9,008|
[]|['Bayesian Infere…|This Specializati…|    96|     198.0|
[[]]|(746,[11],[1.0])|
|google-cbrs-cpi-t…|Well presented co…|    By Ariel R|Sep 27, 2019|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|            Course|1 - 3 Months|       4.6|
3.4k|           Mixed|https://www.cours…|           9,008|
[]|['Bayesian Infere…|This Specializati…|    96|     198.0|
[[]]|(746,[11],[1.0])|
|google-cbrs-cpi-t…|Course is easy en…|    By Austin N|Dec 27, 2019|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
```

```
Spectrum…|                Course|1 - 3 Months|            4.6|
3.4k|          Mixed|https://www.cours…|                  9,008|
[]|['Bayesian Infere…|This Specializati…|    96|     198.0|
[[]]|(746,[11],[1.0])|
|google-cbrs-cpi-t…|The instructions …| By Jayson P J|Nov 22, 2019|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|                Course|1 - 3 Months|            4.6|
3.4k|          Mixed|https://www.cours…|                  9,008|
[]|['Bayesian Infere…|This Specializati…|    96|     198.0|
[[]]|(746,[11],[1.0])|
|google-cbrs-cpi-t…|Easy to follow. D…|   By Robert S|Nov 07, 2019|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|                Course|1 - 3 Months|            4.6|
3.4k|          Mixed|https://www.cours…|                  9,008|
[]|['Bayesian Infere…|This Specializati…|    96|     198.0|
[[]]|(746,[11],[1.0])|
|google-cbrs-cpi-t…|Excellent Trainin…|   By Scott C|Sep 02, 2020|
5.0|Become a CBRS Cer…|Google - Spectrum…|Become a CBRS Cer…|Google -
Spectrum…|                Course|1 - 3 Months|            4.6|
3.4k|          Mixed|https://www.cours…|                  9,008|
[]|['Bayesian Infere…|This Specializati…|    96|     198.0|
[[]]|(746,[11],[1.0])|
+------------------+------------------+------------+-----------+------+--
----------------+------------------+------------------+------------------
-+------------------+-----------+-----------+------------------+--------
---------+-----------------+------------------+-----------+----------
----------+------------------+------+----------+-----+---------------+
only showing top 20 rows
```

[16]:
```python
(training_data, test_data) = data_content.randomSplit([0.8, 0.2])
```

[17]:
```python
# Train the collaborative filtering model on the combined data
model = als.fit(training_data)
```

[18]:
```python
# Moake Predictions on Test set
predictions = model.transform(test_data)
```

[19]:
```python
# Calculate Mean-Square-Error (MSE)
evaluator = RegressionEvaluator(labelCol="course_rating",
  ↪predictionCol="prediction")
mse = evaluator.evaluate(predictions, {evaluator.metricName: "mse"})
print(f"Mean Squared Error (MSE): {mse}")

# Calculate the Root Mean-Squeare-Error (RMSE)
rmse = evaluator.evaluate(predictions, {evaluator.metricName: "rmse"})
print(f"Root Mean Squared Error (RMSE): {rmse}")
```

```
# Calculate the Mean Absolute Error (MAE)
mae = evaluator.evaluate(predictions, {evaluator.metricName: "mae"})
print(f"Mean Absolute Error (RMSE): {mae}")

# Calculate the R^2 Coefficient
r2 = evaluator.evaluate(predictions, {evaluator.metricName: "r2"})
print(f"R^2 Score: {r2}")
```

```
Mean Squared Error (MSE): 6.522344981598273e-06
Root Mean Squared Error (RMSE): 0.0025538882085162366
Mean Absolute Error (RMSE): 0.0025287181603266566
R^2 Score: 0.9999425585168622
```

[20]:
```
target_skill = "Web Development"
skill_data = spark.createDataFrame([(target_skill,)], ["course_skills"])
skill_vector = content_model.transform(skill_data).select("features").
 ↪collect()[0][0]
course_recommendations = model.recommendForItemSubset(skill_vector, 14)  # Get
 ↪top recommendations
```

[21]:
```
# View predicted courses
course_recommendations.select("course_title", "course_rating").show()
```

```
+-----------------------------------------------------------+-------------+
|course_title                                               |course_rating|
+-----------------------------------------------------------+-------------+
|Introduction to Front-End Development                      |4.9          |
|Core Java                                                  |4.8          |
|Introduction to Web Development with HTML, CSS, JavaScript |4.7          |
|Meta Android Developer                                     |4.7          |
|Meta Back-End Developer                                    |4.7          |
|Meta iOS Developer                                         |4.7          |
|Programming with JavaScript                                |4.7          |
|React Basics                                               |4.7          |
|Web Design for Everybody: Basics of Web Development & Coding|4.7          |
|IBM Front-End Developer                                    |4.6          |
|Version Control                                            |4.6          |
|Desarrollador front-end de Meta                            |4.4          |
|Developing Front-End Apps with React                       |4.4          |
|.NET FullStack Developer                                   |4.3          |
+-----------------------------------------------------------+-------------+
```

[22]:
```
target_skill = "Talent Management"
skill_data = spark.createDataFrame([(target_skill,)], ["course_skills"])
skill_vector = content_model.transform(skill_data).select("features").
 ↪collect()[0][0]
```

```
course_recommendations = model.recommendForItemSubset(skill_vector, 5)  # Get␣
  ↪top 5 recommendations
```

[23]:
```
# View predicted courses
course_recommendations.select("course_title", "course_rating").show()
```

```
+-------------------------+-------------+
|course_title             |course_rating|
+-------------------------+-------------+
|Leading People and Teams |4.8          |
|Omnichannel Retail Strategy|4.8        |
|People Analytics         |4.8          |
+-------------------------+-------------+
```