Register and visualize dataset

Introduction

In this lab you will ingest and transform the customer product reviews dataset. Then you will use AWS data stack services such as AWS Glue and Amazon Athena for ingesting and querying the dataset. Finally you will use AWS Data Wrangler to analyze the dataset and plot some visuals extracting insights.

Table of Contents

Let's install the required modules first.

In [2]:

```
# please ignore warning messages during the installation
!pip install --disable-pip-version-check -q sagemaker==2.35.0
!pip install --disable-pip-version-check -q pandas==1.1.4
!pip install --disable-pip-version-check -q awswrangler==2.7.0
!pip install --disable-pip-version-check -q numpy==1.18.5
!pip install --disable-pip-version-check -q seaborn==0.11.0
!pip install --disable-pip-version-check -q matplotlib===3.3.3
```

```
Keyring is skipped due to an exception: 'keyring.backends'
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package ma
nager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
Keyring is skipped due to an exception: 'keyring.backends'
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is
the source of the following dependency conflicts.
sparkmagic 0.20.0 requires nest-asyncio==1.5.5, but you have nest-asyncio 1.5.6 which is incompatible.
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package ma
nager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
Keyring is skipped due to an exception: 'keyring.backends'
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is
the source of the following dependency conflicts.
sparkmagic 0.20.0 requires nest-asyncio==1.5.5, but you have nest-asyncio 1.5.6 which is incompatible.
sagemaker-data-insights 0.3.3 requires numpy>=1.21.6, but you have numpy 1.20.3 which is incompatible.
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package ma
nager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
Keyring is skipped due to an exception: 'keyring.backends'
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is
the source of the following dependency conflicts.
sparkmagic 0.20.0 requires nest-asyncio==1.5.5, but you have nest-asyncio 1.5.6 which is incompatible.
sagemaker-data-insights 0.3.3 requires numpy>=1.21.6, but you have numpy 1.18.5 which is incompatible.
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package ma
nager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
Keyring is skipped due to an exception: 'keyring.backends'
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package ma
nager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
Keyring is skipped due to an exception: 'keyring.backends'
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package ma
nager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
```

1. Ingest and transform the public dataset

The dataset Women's Clothing Reviews (https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews) has been chosen as the main dataset.

It is shared in a public Amazon S3 bucket, and is available as a comma-separated value (CSV) text format:

```
s3://dlai-practical-data-science/data/raw/womens_clothing_ecommerce_reviews.csv
```

1.1. List the dataset files in the public S3 bucket

The AWS Command Line Interface (CLI) (https://awscli.amazonaws.com/v2/documentation/api/latest/index.html) is a unified tool to manage your AWS services. With just one tool, you can control multiple AWS services from the command line and automate them through scripts. You will use it to list the dataset files.

View dataset files in CSV format

`aws s3 ls [bucket_name]` function lists all objects in the S3 bucket. Let's use it to view the reviews data files in CSV format:

Exercise 1

View the list of the files available in the public bucket `s3://dlai-practical-data-science/data/raw/` .

Instructions: Use `aws s3 ls [bucket_name]` function. To run the AWS CLI command from the notebook you will need to put an exclamation mark in front of it: `!aws` . You should see the data file `womens_clothing_ecommerce_reviews.csv` in the list.

In [5]:

```
### BEGIN SOLUTION - DO NOT delete this comment for grading purposes
None # Replace None
!aws s3 ls s3://dlai-practical-data-science/data/raw/
### END SOLUTION - DO NOT delete this comment for grading purposes

# EXPECTED OUTPUT
# ... womens_clothing_ecommerce_reviews.csv
```

```
2021-04-30 02:21:06    8457214 womens_clothing_ecommerce_reviews.csv
```

1.2. Copy the data locally to the notebook

`aws s3 cp [bucket_name/file_name] [file_name]` function copies the file from the S3 bucket into the local environment or into another S3 bucket. Let's use it to copy the file with the dataset locally.

In [6]:

```
!aws s3 cp s3://dlai-practical-data-science/data/raw/womens_clothing_ecommerce_reviews.csv ./womens_clothing_ecommerce_reviews.csv
```

```
download: s3://dlai-practical-data-science/data/raw/womens_clothing_ecommerce_reviews.csv to ./womens_clothing_ecommerce_revi
ews.csv
```

Now use the Pandas dataframe to load and preview the data.

In [7]:

```
import pandas as pd
import csv

df = pd.read_csv('./womens_clothing_ecommerce_reviews.csv',
                 index_col=0)

df.shape
```

Out[7]:

```
(23486, 10)
```

In [8]:

```
df
```

Out[8]:

	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
0	847	33	Cute, crisp shirt	If this product was in petite i would get the...	4	1	2	General	Tops	Blouses
1	1080	34	NaN	Love this dress! it's sooo pretty. i happene...	5	1	4	General	Dresses	Dresses
2	1077	60	Some major design flaws	I had such high hopes for this dress and reall...	3	0	0	General	Dresses	Dresses
3	1049	50	My favorite buy!	I love love love this jumpsuit. it's fun fl...	5	1	0	General Petite	Bottoms	Pants
4	847	47	Flattering shirt	This shirt is very flattering to all due to th...	5	1	6	General	Tops	Blouses
...	...	...	...	...	...	...	...	...	...	...
23481	1104	34	Great dress for many occasions	I was very happy to snag this dress at such a ...	5	1	0	General Petite	Dresses	Dresses
23482	862	48	Wish it was made of cotton	It reminds me of maternity clothes. soft stre...	3	1	0	General Petite	Tops	Knits
23483	1104	31	Cute, but see through	This fit well but the top was very see throug...	3	0	1	General Petite	Dresses	Dresses
23484	1084	28	Very cute dress, perfect for summer parties an...	I bought this dress for a wedding i have this ...	3	1	2	General	Dresses	Dresses
23485	1104	52	Please make more like this one!	This dress in a lovely platinum is feminine an...	5	1	22	General Petite	Dresses	Dresses

23486 rows × 10 columns

1.3. Transform the data

To simplify the task, you will transform the data into a comma-separated value (CSV) file that contains only a `review_body`, `product_category`, and `sentiment` derived from the original data.

In [11]:

```
df_transformed = df.rename(columns={'Review Text': 'review_body',
                                    'Rating': 'star_rating',
                                    'Class Name': 'product_category'})
df_transformed.drop(columns=['Clothing ID', 'Age', 'Title', 'Recommended IND', 'Positive Feedback Count', 'Division Name', 'Department Name'],
                    inplace=True)

df_transformed.dropna(inplace=True)

df_transformed.shape

df_transformed
```

Out[11]:

	review_body	star_rating	product_category
0	If this product was in petite i would get the...	4	Blouses
1	Love this dress! it's sooo pretty. i happene...	5	Dresses
2	I had such high hopes for this dress and reall...	3	Dresses
3	I love love love this jumpsuit. it's fun fl...	5	Pants
4	This shirt is very flattering to all due to th...	5	Blouses
...	...	...	...
23481	I was very happy to snag this dress at such a ...	5	Dresses
23482	It reminds me of maternity clothes. soft stre...	3	Knits
23483	This fit well but the top was very see throug...	3	Dresses
23484	I bought this dress for a wedding i have this ...	3	Dresses
23485	This dress in a lovely platinum is feminine an...	5	Dresses

22628 rows × 3 columns

Now convert the `star_rating` into the `sentiment` (positive, neutral, negative), which later on will be for the prediction.

In [16]:

```
def to_sentiment(star_rating):
    if star_rating in {1, 2}: # negative
        return -1
    if star_rating == 3:      # neutral
        return 0
    if star_rating in {4, 5}: # positive
        return 1

# transform star_rating into the sentiment
df_transformed['sentiment'] = df_transformed['star_rating'].apply(lambda star_rating:
    to_sentiment(star_rating=star_rating)
)

# drop the star rating column
df_transformed.drop(columns=['star_rating'],
                    inplace=True)

# remove reviews for product_categories with < 10 reviews
df_transformed = df_transformed.groupby('product_category').filter(lambda reviews : len(reviews) > 10)[['sentiment', 'review_body', 'produ
ct_category']]

df_transformed.shape
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
/opt/conda/lib/python3.7/site-packages/pandas/core/indexes/base.py in get_loc(self, key, method, tolerance)
   2894             try:
-> 2895                 return self._engine.get_loc(casted_key)
   2896             except KeyError as err:

pandas/_libs/index.pyx in pandas._libs.index.IndexEngine.get_loc()

pandas/_libs/index.pyx in pandas._libs.index.IndexEngine.get_loc()

pandas/_libs/hashtable_class_helper.pxi in pandas._libs.hashtable.PyObjectHashTable.get_item()

pandas/_libs/hashtable_class_helper.pxi in pandas._libs.hashtable.PyObjectHashTable.get_item()

KeyError: 'star_rating'

The above exception was the direct cause of the following exception:

KeyError                                  Traceback (most recent call last)
<ipython-input-16-7e92a1fbc9fb> in <module>
      8
      9 # transform star_rating into the sentiment
---> 10 df_transformed['sentiment'] = df_transformed['star_rating'].apply(lambda star_rating:
     11     to_sentiment(star_rating=star_rating)
     12 )

/opt/conda/lib/python3.7/site-packages/pandas/core/frame.py in __getitem__(self, key)
   2904             if self.columns.nlevels > 1:
   2905                 return self._getitem_multilevel(key)
-> 2906             indexer = self.columns.get_loc(key)
   2907             if is_integer(indexer):
   2908                 indexer = [indexer]

/opt/conda/lib/python3.7/site-packages/pandas/core/indexes/base.py in get_loc(self, key, method, tolerance)
   2895                 return self._engine.get_loc(casted_key)
   2896             except KeyError as err:
-> 2897                 raise KeyError(key) from err
   2898
   2899         if tolerance is not None:

KeyError: 'star_rating'
```

In [17]:

```
# preview the results
df_transformed
```

Out[17]:

	sentiment	review_body	product_category
0	1	If this product was in petite i would get the...	Blouses
1	1	Love this dress! it's sooo pretty. i happene...	Dresses
2	0	I had such high hopes for this dress and reall...	Dresses
3	1	I love love love this jumpsuit. it's fun fl...	Pants
4	1	This shirt is very flattering to all due to th...	Blouses
...	...	...	...
23481	1	I was very happy to snag this dress at such a ...	Dresses
23482	0	It reminds me of maternity clothes. soft stre...	Knits
23483	0	This fit well but the top was very see throug...	Dresses
23484	0	I bought this dress for a wedding i have this ...	Dresses
23485	1	This dress in a lovely platinum is feminine an...	Dresses

22626 rows × 3 columns

1.4 Write the data to a CSV file

In [18]:

```
df_transformed.to_csv('./womens_clothing_ecommerce_reviews_transformed.csv',
                      index=False)
```

In [19]:

```
!head -n 5 ./womens_clothing_ecommerce_reviews_transformed.csv
```

```
sentiment,review_body,product_category
1,If this product was in petite  i would get the petite. the regular is a little long on me but a tailor can do a simple fix
on that.     fits nicely! i'm 5'4  130lb and pregnant so i bough t medium to grow into.     the tie can be front or back so p
rovides for some nice flexibility on form fitting.,Blouses
1,"Love this dress!  it's sooo pretty.  i happened to find it in a store  and i'm glad i did bc i never would have ordered it
online bc it's petite.  i bought a petite and am 5'8"".  i love the length on me- hits just a little below the knee.  would d
efinitely be a true midi on someone who is truly petite.",Dresses
0,I had such high hopes for this dress and really wanted it to work for me. i initially ordered the petite small (my usual si
ze) but i found this to be outrageously small. so small in fact that i could not zip it up! i reordered it in petite medium
which was just ok. overall  the top half was comfortable and fit nicely  but the bottom half had a very tight under layer and
several somewhat cheap (net) over layers. imo  a major design flaw was the net over layer sewn directly into the zipper - it
c,Dresses
1,I love  love  love this jumpsuit. it's fun  flirty  and fabulous! every time i wear it  i get nothing but great compliment
s!,Pants
```

2. Register the public dataset for querying and visualizing

You will register the public dataset into an S3-backed database table so you can query and visualize our dataset at scale.

2.1. Register S3 dataset files as a table for querying

Let's import required modules.

`boto3` is the AWS SDK for Python to create, configure, and manage AWS services, such as Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Simple Storage Service (Amazon S3). The SDK provides an object-oriented API as well as low-level access to AWS services.

`sagemaker` is the SageMaker Python SDK which provides several high-level abstractions for working with the Amazon SageMaker.

In [20]:

```
import boto3
import sagemaker
import pandas as pd
import numpy as np
import botocore

config = botocore.config.Config(user_agent_extra='dlai-pds/c1/w1')

# low-level service client of the boto3 session
sm = boto3.client(service_name='sagemaker',
                  config=config)

sess = sagemaker.Session(sagemaker_client=sm)

bucket = sess.default_bucket()
role = sagemaker.get_execution_role()
region = sess.boto_region_name
account_id = sess.account_id

print('S3 Bucket: {}'.format(bucket))
print('Region: {}'.format(region))
print('Account ID: {}'.format(account_id))
```

```
S3 Bucket: sagemaker-us-east-1-392437363011
Region: us-east-1
Account ID: <bound method Session.account_id of <sagemaker.session.Session object at 0x7fc1fcecf810>>
```

Review the empty bucket which was created automatically for this account.

Instructions:

- open the link
- click on the S3 bucket name `sagemaker-us-east-1-ACCOUNT`
- check that it is empty at this stage

In [21]:

```
from IPython.core.display import display, HTML

display(HTML('<b>Review <a target="top" href="https://s3.console.aws.amazon.com/s3/home?region={}#">Amazon S3 buckets</a></b>'.format(region)))
```

Review Amazon S3 buckets (https://s3.console.aws.amazon.com/s3/home?region=us-east-1#)

Copy the file into the S3 bucket.

In [22]:

```
!aws s3 cp ./womens_clothing_ecommerce_reviews_transformed.csv s3://$bucket/data/transformed/womens_clothing_ecommerce_reviews_transformed.csv
```

```
upload: ./womens_clothing_ecommerce_reviews_transformed.csv to s3://sagemaker-us-east-1-392437363011/data/transformed/womens_clothing_ecommerce_reviews_transformed.csv
```

Review the bucket with the file we uploaded above.

Instructions:

- open the link
- check that the CSV file is located in the S3 bucket
- check the location directory structure is the same as in the CLI command above
- click on the file name and see the available information about the file (region, size, S3 URI, Amazon Resource Name (ARN))

In [23]:

```
from IPython.core.display import display, HTML

display(HTML('<b>Review <a target="top" href="https://s3.console.aws.amazon.com/s3/buckets/{}?region={}&prefix=data/transformed/#">Amazon S3 buckets</a></b>'.format(bucket, region)))
```

Review Amazon S3 buckets (https://s3.console.aws.amazon.com/s3/buckets/sagemaker-us-east-1-392437363011?region=us-east-1&prefix=data/transformed/#)

Import AWS Data Wrangler

AWS Data Wrangler (https://github.com/awslabs/aws-data-wrangler) is an AWS Professional Service open source python initiative that extends the power of Pandas library to AWS connecting dataframes and AWS data related services (Amazon Redshift, AWS Glue, Amazon Athena, Amazon EMR, Amazon QuickSight, etc).

Built on top of other open-source projects like Pandas, Apache Arrow, Boto3, SQLAlchemy, Psycopg2 and PyMySQL, it offers abstracted functions to execute usual ETL tasks like load/unload data from data lakes, data warehouses and databases.

Review the AWS Data Wrangler documentation: https://aws-data-wrangler.readthedocs.io/en/stable/ (https://aws-data-wrangler.readthedocs.io/en/stable/)

In [24]:

```
import awswrangler as wr
```

Create AWS Glue Catalog database

The data catalog features of **AWS Glue** and the inbuilt integration to Amazon S3 simplify the process of identifying data and deriving the schema definition out of the discovered data. Using AWS Glue crawlers within your data catalog, you can traverse your data stored in Amazon S3 and build out the metadata tables that are defined in your data catalog.

Here you will use `wr.catalog.create_database` function to create a database with the name `dsoaws_deep_learning` ("dsoaws" stands for "Data Science on AWS").

In [25]:

```
wr.catalog.create_database(
    name='dsoaws_deep_learning',
    exist_ok=True
)
```

In [26]:

```
dbs = wr.catalog.get_databases()

for db in dbs:
    print("Database name: " + db['Name'])
```

```
Database name: dsoaws_deep_learning
```

Review the created database in the AWS Glue Catalog.

Instructions:

- open the link
- on the left side panel notice that you are in the AWS Glue -> Data Catalog -> Databases
- check that the database `dsoaws_deep_learning` has been created
- click on the name of the database
- click on the `Tables in dsoaws_deep_learning` link to see that there are no tables

In [27]:

```
from IPython.core.display import display, HTML

display(HTML('<b>Review <a target="top" href="https://console.aws.amazon.com/glue/home?region={}#catalog:tab=databases">AWS Glue Databases</a></b>'.format(region)))
```

Review [AWS Glue Databases (https://console.aws.amazon.com/glue/home?region=us-east-1#catalog:tab=databases)](https://console.aws.amazon.com/glue/home?region=us-east-1#catalog:tab=databases)

Register CSV data with AWS Glue Catalog

Exercise 2

Register CSV data with AWS Glue Catalog.

Instructions: Use `wr.catalog.create_csv_table` function with the following parameters

```
res = wr.catalog.create_csv_table(
    database='AWS Glue Catalog', # AWS Glue Catalog database name
    path='s3://{}/data/transformed/'.format(bucket), # S3 object path for the data
    table='reviews', # registered table name
    columns_types={
        'sentiment': 'int',
        'review_body': 'string',
        'product_category': 'string'
    },
    mode='overwrite',
    skip_header_line_count=1,
    sep=','
)
```

In [29]:

```
wr.catalog.create_csv_table(
    ### BEGIN SOLUTION - DO NOT delete this comment for grading purposes
    database='dsoaws_deep_learning', # Replace None
    ### END SOLUTION - DO NOT delete this comment for grading purposes
    path='s3://{}/data/transformed/'.format(bucket),
    table="reviews",
    columns_types={
        'sentiment': 'int',
        'review_body': 'string',
        'product_category': 'string'
    },
    mode='overwrite',
    skip_header_line_count=1,
    sep=','
)
```

Review the registered table in the AWS Glue Catalog.

Instructions:

- open the link
- on the left side panel notice that you are in the AWS Glue -> Data Catalog -> Databases -> Tables
- check that you can see the table `reviews` from the database `dsoaws_deep_learning` in the list
- click on the name of the table
- explore the available information about the table (name, database, classification, location, schema etc.)

In [30]:

```
from IPython.core.display import display, HTML

display(HTML('<b>Review <a target="top" href="https://console.aws.amazon.com/glue/home?region={}#">AWS Glue Catalog</a></b>'.format(regio
n)))
```

Review AWS Glue Catalog (https://console.aws.amazon.com/glue/home?region=us-east-1#)

Review the table shape:

In [31]:

```
table = wr.catalog.table(database='dsoaws_deep_learning',
                         table='reviews')
table
```

Out[31]:

	Column Name	Type	Partition	Comment
0	sentiment	int	False	
1	review_body	string	False	
2	product_category	string	False	

2.2. Create default S3 bucket for Amazon Athena

Amazon Athena requires this S3 bucket to store temporary query results and improve performance of subsequent queries.

The contents of this bucket are mostly binary and human-unreadable.

In [32]:

```
# S3 bucket name
wr.athena.create_athena_bucket()

# EXPECTED OUTPUT
# 's3://aws-athena-query-results-ACCOUNT-REGION/'
```

Out[32]:

```
's3://aws-athena-query-results-392437363011-us-east-1/'
```

3. Visualize data

Reviews dataset - column descriptions

- `sentiment` : The review's sentiment (-1, 0, 1).
- `product_category` : Broad product category that can be used to group reviews (in this case digital videos).
- `review_body` : The text of the review.

3.1. Preparation for data visualization

Imports

In [33]:

```
import numpy as np
import seaborn as sns

import matplotlib.pyplot as plt
%matplotlib inline
%config InlineBackend.figure_format='retina'
```

Settings

Set AWS Glue database and table name.

In [34]:

```
# Do not change the database and table names - they are used for grading purposes!
database_name = 'dsoaws_deep_learning'
table_name = 'reviews'
```

Set seaborn parameters. You can review seaborn documentation following the link (https://seaborn.pydata.org/index.html).

In [35]:

```
sns.set_style = 'seaborn-whitegrid'

sns.set(rc={"font.style":"normal",
            "axes.facecolor":"white",
            'grid.color': '.8',
            'grid.linestyle': '-',
            "figure.facecolor":"white",
            "figure.titlesize":20,
            "text.color":"black",
            "xtick.color":"black",
            "ytick.color":"black",
            "axes.labelcolor":"black",
            "axes.grid":True,
            'axes.labelsize':10,
            'xtick.labelsize':10,
            'font.size':10,
            'ytick.labelsize':10})
```

Helper code to display values on barplots:

Run SQL queries using Amazon Athena

Amazon Athena lets you query data in Amazon S3 using a standard SQL interface. It reflects the databases and tables in the AWS Glue Catalog. You can create interactive queries and perform any data manipulations required for further downstream processing.

Standard SQL query can be saved as a string and then passed as a parameter into the Athena query. Run the following cells as an example to count the total number of reviews by sentiment. The SQL query here will take the following form:

```
SELECT column_name, COUNT(column_name) as new_column_name
FROM table_name
GROUP BY column_name
ORDER BY column_name
```

If you are not familiar with the SQL query statements, you can review some tutorials following the link (https://www.w3schools.com/sql/default.asp).

3.2. How many reviews per sentiment?

Set the SQL statement to find the count of sentiments:

In [36]:

```
statement_count_by_sentiment = """
SELECT sentiment, COUNT(sentiment) AS count_sentiment
FROM reviews
GROUP BY sentiment
ORDER BY sentiment
"""

print(statement_count_by_sentiment)
```

```
SELECT sentiment, COUNT(sentiment) AS count_sentiment
FROM reviews
GROUP BY sentiment
ORDER BY sentiment
```

Query data in Amazon Athena database cluster using the prepared SQL statement:

In [37]:

```
df_count_by_sentiment = wr.athena.read_sql_query(
    sql=statement_count_by_sentiment,
    database=database_name
)

print(df_count_by_sentiment)
```

```
   sentiment  count_sentiment
0         -1             2370
1          0             2823
2          1            17433
```

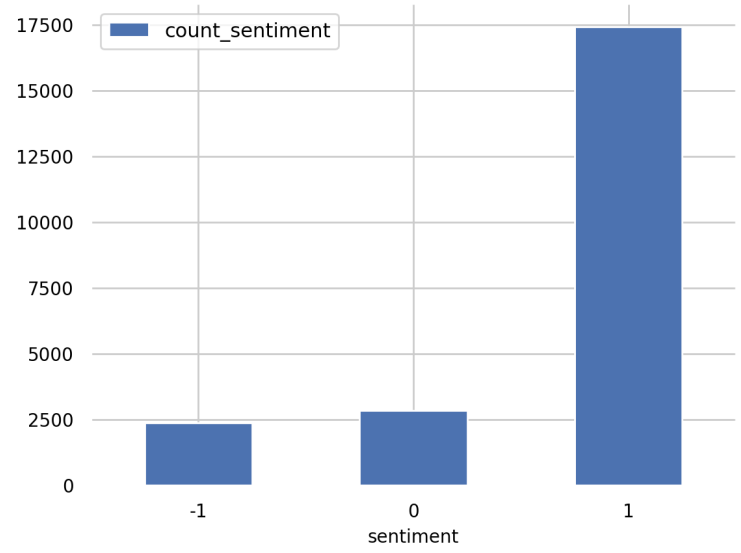
Preview the results of the query:

In [38]:

```
df_count_by_sentiment.plot(kind='bar', x='sentiment', y='count_sentiment', rot=0)
```

Out[38]:

```
<AxesSubplot:xlabel='sentiment'>
```



Exercise 3

Use Amazon Athena query with the standard SQL statement passed as a parameter, to calculate the total number of reviews per `product_category` in the table `reviews`.

Instructions: Pass the SQL statement of the form

```
SELECT category_column, COUNT(column_name) AS new_column_name
FROM table_name
GROUP BY category_column
ORDER BY new_column_name DESC
```

as a triple quote string into the variable `statement_count_by_category`. Please use the column `sentiment` in the `COUNT` function and give it a new name `count_sentiment`.

In [43]:

```
# Replace all None
### BEGIN SOLUTION - DO NOT delete this comment for grading purposes
statement_count_by_category = """
SELECT product_category, COUNT(sentiment) AS count_sentiment
FROM reviews
GROUP BY product_category
ORDER BY count_sentiment DESC
"""
### END SOLUTION - DO NOT delete this comment for grading purposes
print(statement_count_by_category)
```

```
SELECT product_category, COUNT(sentiment) AS count_sentiment
FROM reviews
GROUP BY product_category
ORDER BY count_sentiment DESC
```

Query data in Amazon Athena database passing the prepared SQL statement:

In [44]:

```
%%time
df_count_by_category = wr.athena.read_sql_query(
    sql=statement_count_by_category,
    database=database_name
)

df_count_by_category

# EXPECTED OUTPUT
# Dresses: 6145
# Knits: 4626
# Blouses: 2983
# Sweaters: 1380
# Pants: 1350
# ...
```

```
CPU times: user 295 ms, sys: 22.3 ms, total: 317 ms
Wall time: 2.89 s
```

Out[44]:

	product_category	count_sentiment
0	Dresses	6145
1	Knits	4626
2	Blouses	2983
3	Sweaters	1380
4	Pants	1350
5	Jeans	1104
6	Fine gauge	1059
7	Skirts	903
8	Jackets	683
9	Lounge	669
10	Swim	332
11	Outerwear	319
12	Shorts	304
13	Sleep	214
14	Legwear	158
15	Intimates	147
16	Layering	132
17	Trend	118

3.3. Which product categories are highest rated by average sentiment?

Set the SQL statement to find the average sentiment per product category, showing the results in the descending order:

In [45]:

```
statement_avg_by_category = """
SELECT product_category, AVG(sentiment) AS avg_sentiment
FROM {}
GROUP BY product_category
ORDER BY avg_sentiment DESC
""".format(table_name)

print(statement_avg_by_category)
```

```
SELECT product_category, AVG(sentiment) AS avg_sentiment
FROM reviews
GROUP BY product_category
ORDER BY avg_sentiment DESC
```

Query data in Amazon Athena database passing the prepared SQL statement:

In [46]:

```
%%time
df_avg_by_category = wr.athena.read_sql_query(
    sql=statement_avg_by_category,
    database=database_name
)
```

```
CPU times: user 415 ms, sys: 26 ms, total: 441 ms
Wall time: 2.97 s
```

Preview the query results in the temporary S3 bucket: `s3://aws-athena-query-results-ACCOUNT-REGION/`

Instructions:

- open the link
- check the name of the S3 bucket
- briefly check the content of it

In [47]:

```
from IPython.core.display import display, HTML

display(HTML('<b>Review <a target="top" href="https://s3.console.aws.amazon.com/s3/buckets/aws-athena-query-results-{}-{}?region={}">Amazon S3 buckets</a></b>'.format(account_id, region, region)))
```

Review [Amazon S3 buckets (https://s3.console.aws.amazon.com/s3/buckets/aws-athena-query-results-<bound method Session.account_id of <sagemaker.session.Session object at 0x7fc1fcecf810>>-us-east-1?region=us-east-1)](https://s3.console.aws.amazon.com/s3/buckets/aws-athena-query-results-<bound method Session.account_id of <sagemaker.session.Session object at 0x7fc1fcecf810>>-us-east-1?region=us-east-1)

Preview the results of the query:

In [48]:

```
df_avg_by_category
```

Out[48]:

	product_category	avg_sentiment
0	Layering	0.780303
1	Jeans	0.746377
2	Lounge	0.745889
3	Sleep	0.710280
4	Shorts	0.707237
5	Pants	0.705185
6	Intimates	0.700680
7	Jackets	0.699854
8	Skirts	0.696567
9	Legwear	0.696203
10	Fine gauge	0.692162
11	Outerwear	0.683386
12	Knits	0.653913
13	Swim	0.644578
14	Dresses	0.643287
15	Sweaters	0.641304
16	Blouses	0.641301
17	Trend	0.483051

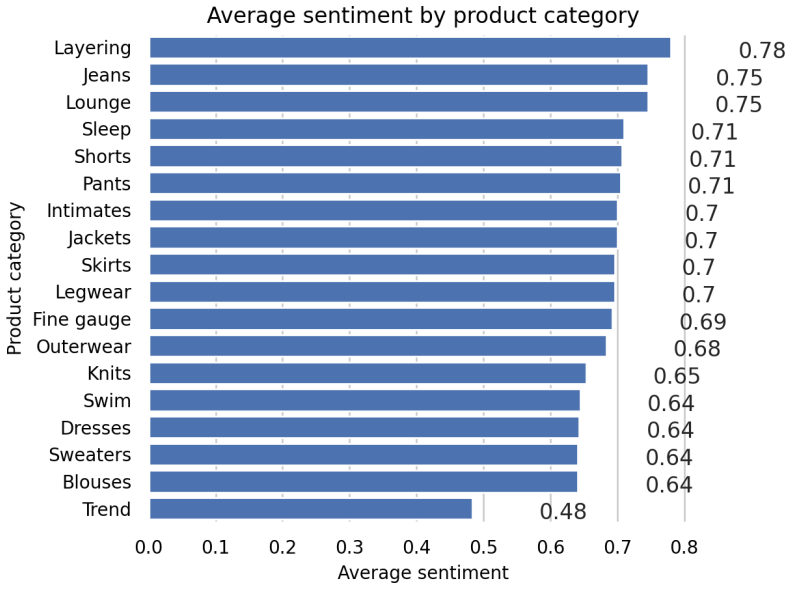
Visualization

In [49]:

```
def show_values_barplot(axs, space):
    def _show_on_plot(ax):
        for p in ax.patches:
            _x = p.get_x() + p.get_width() + float(space)
            _y = p.get_y() + p.get_height()
            value = round(float(p.get_width()),2)
            ax.text(_x, _y, value, ha="left")

    if isinstance(axs, np.ndarray):
        for idx, ax in np.ndenumerate(axs):
            _show_on_plot(ax)
    else:
        _show_on_plot(axs)
```

In [50]:

```
# Create plot
barplot = sns.barplot(
    data = df_avg_by_category,
    y='product_category',
    x='avg_sentiment',
    color="b",
    saturation=1
)

# Set the size of the figure
sns.set(rc={'figure.figsize':(15.0, 10.0)})

# Set title and x-axis ticks
plt.title('Average sentiment by product category')
#plt.xticks([-1, 0, 1], ['Negative', 'Neutral', 'Positive'])

# Helper code to show actual values afters bars
show_values_barplot(barplot, 0.1)

plt.xlabel("Average sentiment")
plt.ylabel("Product category")

plt.tight_layout()
# Do not change the figure name - it is used for grading purposes!
plt.savefig('avg_sentiment_per_category.png', dpi=300)

# Show graphic
plt.show(barplot)
```



In [51]:

```
# Upload image to S3 bucket
sess.upload_data(path='avg_sentiment_per_category.png', bucket=bucket, key_prefix="images")
```

Out[51]:

```
's3://sagemaker-us-east-1-392437363011/images/avg_sentiment_per_category.png'
```

Review the bucket on the account.

Instructions:

- open the link
- click on the S3 bucket name `sagemaker-us-east-1-ACCOUNT`
- open the images folder
- check the existence of the image `avg_sentiment_per_category.png`
- if you click on the image name, you can see the information about the image file. You can also download the file with the command on the top right Object Actions -> Download / Download as

In [52]:

```
from IPython.core.display import display, HTML

display(HTML('<b>Review <a target="top" href="https://s3.console.aws.amazon.com/s3/home?region={}">Amazon S3 buckets</a></b>'.format(regio
n)))
```

Review [Amazon S3 buckets (https://s3.console.aws.amazon.com/s3/home?region=us-east-1)]

3.4. Which product categories have the most reviews?

Set the SQL statement to find the count of sentiment per product category, showing the results in the descending order:

In [53]:

```
statement_count_by_category_desc = """
SELECT product_category, COUNT(*) AS count_reviews
FROM {}
GROUP BY product_category
ORDER BY count_reviews DESC
""".format(table_name)

print(statement_count_by_category_desc)
```

```
SELECT product_category, COUNT(*) AS count_reviews
FROM reviews
GROUP BY product_category
ORDER BY count_reviews DESC
```

Query data in Amazon Athena database passing the prepared SQL statement:

In [54]:

```
%%time
df_count_by_category_desc = wr.athena.read_sql_query(
    sql=statement_count_by_category_desc,
    database=database_name
)
```

```
CPU times: user 297 ms, sys: 19.8 ms, total: 316 ms
Wall time: 2.92 s
```

Store maximum number of sentiment for the visualization plot:

In [55]:

```
max_sentiment = df_count_by_category_desc['count_reviews'].max()
print('Highest number of reviews (in a single category): {}'.format(max_sentiment))
```

```
Highest number of reviews (in a single category): 6145
```

Visualization

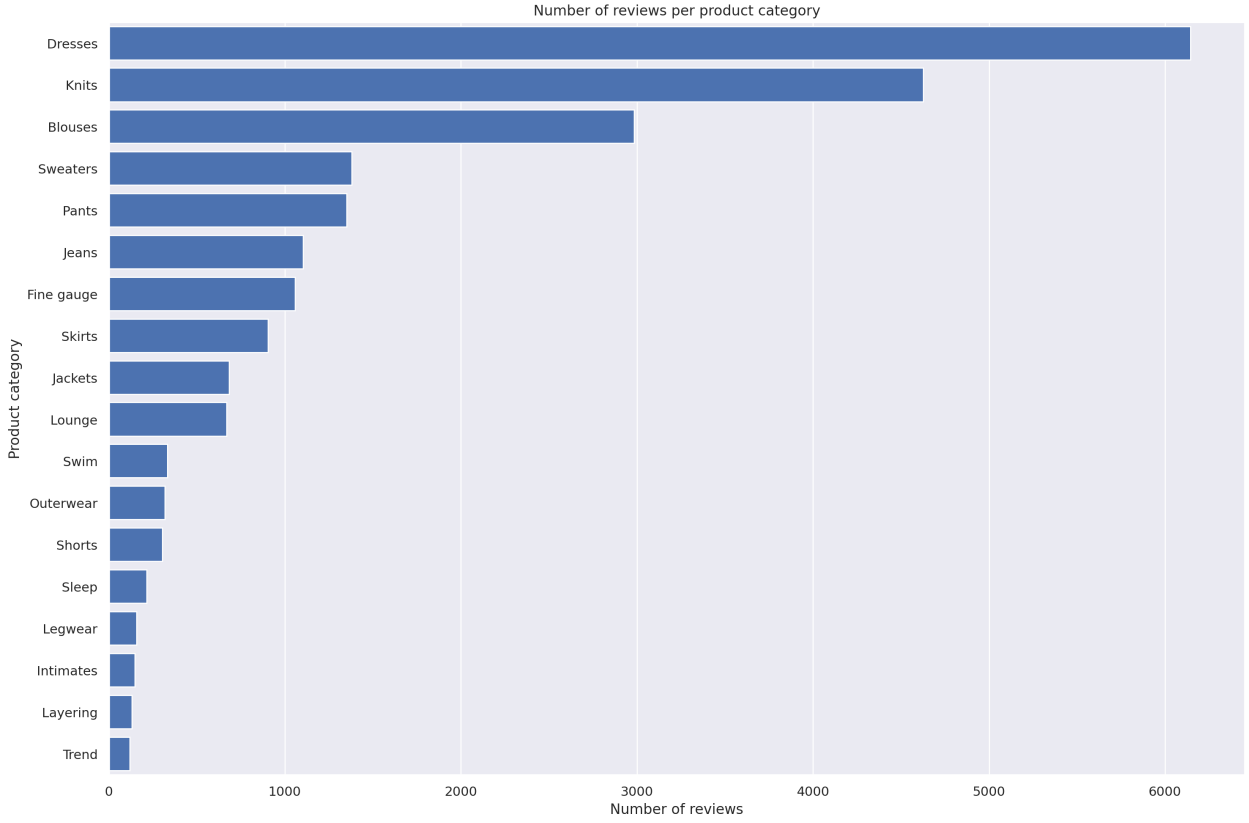
Exercise 4

Use `barplot` function to plot number of reviews per product category.

Instructions: Use the `barplot` chart example in the previous section, passing the newly defined dataframe `df_count_by_category_desc` with the count of reviews. Here, please put the `product_category` column into the `y` argument.

In [58]:

```
# Create seaborn barplot
barplot = sns.barplot(
    ### BEGIN SOLUTION - DO NOT delete this comment for grading purposes
    data=df_count_by_category_desc, # Replace None
    y="product_category", # Replace None
    x="count_reviews", # Replace None
    ### END SOLUTION - DO NOT delete this comment for grading purposes
    color="b",
    saturation=1
)

# Set the size of the figure
sns.set(rc={'figure.figsize':(15.0, 10.0)})

# Set title
plt.title("Number of reviews per product category")
plt.xlabel("Number of reviews")
plt.ylabel("Product category")

plt.tight_layout()

# Do not change the figure name - it is used for grading purposes!
plt.savefig('num_reviews_per_category.png', dpi=300)

# Show the barplot
plt.show(barplot)
```



In [59]:

```
# Upload image to S3 bucket
sess.upload_data(path='num_reviews_per_category.png', bucket=bucket, key_prefix="images")
```

Out[59]:

```
's3://sagemaker-us-east-1-392437363011/images/num_reviews_per_category.png'
```

3.5. What is the breakdown of sentiments per product category?

Set the SQL statement to find the count of sentiment per product category and sentiment:

In [60]:

```
statement_count_by_category_and_sentiment = """
SELECT product_category,
         sentiment,
         COUNT(*) AS count_reviews
FROM {}
GROUP BY  product_category, sentiment
ORDER BY  product_category ASC, sentiment DESC, count_reviews
""".format(table_name)

print(statement_count_by_category_and_sentiment)
```

```
SELECT product_category,
         sentiment,
         COUNT(*) AS count_reviews
FROM reviews
GROUP BY  product_category, sentiment
ORDER BY  product_category ASC, sentiment DESC, count_reviews
```

Query data in Amazon Athena database passing the prepared SQL statement:

In [61]:

```
%%time
df_count_by_category_and_sentiment = wr.athena.read_sql_query(
    sql=statement_count_by_category_and_sentiment,
    database=database_name
)
```

```
CPU times: user 474 ms, sys: 29.1 ms, total: 503 ms
Wall time: 3.7 s
```

Prepare for stacked percentage horizontal bar plot showing proportion of sentiments per product category.

In [62]:

```
# Create grouped dataframes by category and by sentiment
grouped_category = df_count_by_category_and_sentiment.groupby('product_category')
grouped_star = df_count_by_category_and_sentiment.groupby('sentiment')

# Create sum of sentiments per star sentiment
df_sum = df_count_by_category_and_sentiment.groupby(['sentiment']).sum()

# Calculate total number of sentiments
total = df_sum['count_reviews'].sum()
print('Total number of reviews: {}'.format(total))
```

```
Total number of reviews: 22626
```

Create dictionary of product categories and array of star rating distribution per category.

In [63]:

```
distribution = {}
count_reviews_per_star = []
i=0

for category, sentiments in grouped_category:
    count_reviews_per_star = []
    for star in sentiments['sentiment']:
        count_reviews_per_star.append(sentiments.at[i, 'count_reviews'])
        i=i+1;
    distribution[category] = count_reviews_per_star
```

Build array per star across all categories.

In [64]:

```
distribution
```

Out[64]:

```
{'Blouses': [2256, 384, 343],
 'Dresses': [4634, 830, 681],
 'Fine gauge': [837, 118, 104],
 'Intimates': [117, 16, 14],
 'Jackets': [550, 61, 72],
 'Jeans': [909, 110, 85],
 'Knits': [3523, 605, 498],
 'Layering': [113, 9, 10],
 'Legwear': [126, 16, 16],
 'Lounge': [545, 78, 46],
 'Outerwear': [254, 29, 36],
 'Pants': [1074, 154, 122],
 'Shorts': [240, 39, 25],
 'Skirts': [714, 104, 85],
 'Sleep': [175, 16, 23],
 'Sweaters': [1036, 193, 151],
 'Swim': [252, 42, 38],
 'Trend': [78, 19, 21]}
```

In [65]:

```
df_distribution_pct = pd.DataFrame(distribution).transpose().apply(
    lambda num_sentiments: num_sentiments/sum(num_sentiments)*100, axis=1
)
df_distribution_pct.columns=['1', '0', '-1']
df_distribution_pct
```

Out[65]:

	1	0	-1
Blouses	75.628562	12.872947	11.498491
Dresses	75.410903	13.506916	11.082181
Fine gauge	79.036827	11.142587	9.820585
Intimates	79.591837	10.884354	9.523810
Jackets	80.527086	8.931186	10.541728
Jeans	82.336957	9.963768	7.699275
Knits	76.156507	13.078253	10.765240
Layering	85.606061	6.818182	7.575758
Legwear	79.746835	10.126582	10.126582
Lounge	81.464873	11.659193	6.875934
Outerwear	79.623824	9.090909	11.285266
Pants	79.555556	11.407407	9.037037
Shorts	78.947368	12.828947	8.223684
Skirts	79.069767	11.517165	9.413068
Sleep	81.775701	7.476636	10.747664
Sweaters	75.072464	13.985507	10.942029
Swim	75.903614	12.650602	11.445783
Trend	66.101695	16.101695	17.796610

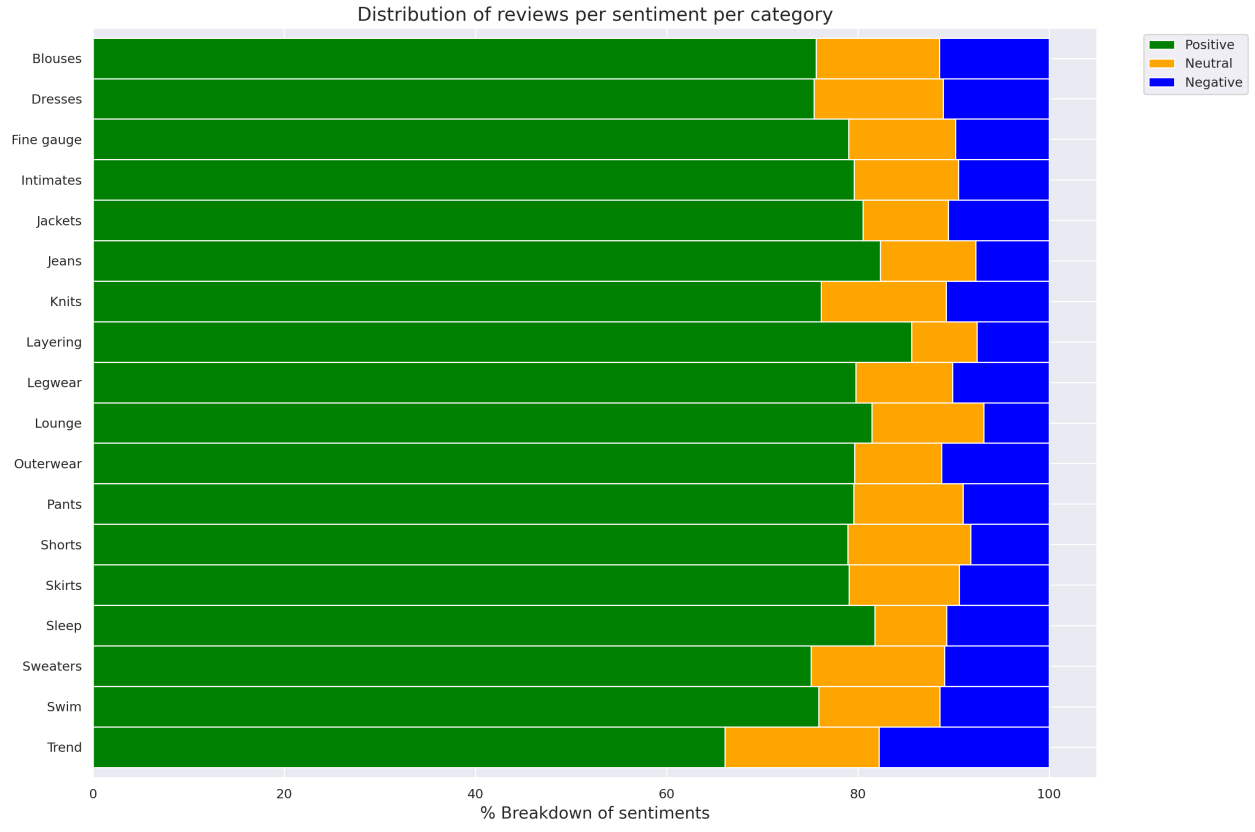
Visualization

Plot the distributions of sentiments per product category.

In [66]:

```
categories = df_distribution_pct.index

# Plot bars
plt.figure(figsize=(10,5))

df_distribution_pct.plot(kind="barh",
                         stacked=True,
                         edgecolor='white',
                         width=1.0,
                         color=['green',
                                'orange',
                                'blue'])

plt.title("Distribution of reviews per sentiment per category",
          fontsize='16')

plt.legend(bbox_to_anchor=(1.04,1),
           loc="upper left",
           labels=['Positive',
                   'Neutral',
                   'Negative'])

plt.xlabel("% Breakdown of sentiments", fontsize='14')
plt.gca().invert_yaxis()
plt.tight_layout()

# Do not change the figure name - it is used for grading purposes!
plt.savefig('distribution_sentiment_per_category.png', dpi=300)
plt.show()
```

```
<Figure size 1000x500 with 0 Axes>
```



In [67]:

```
# Upload image to S3 bucket
sess.upload_data(path='distribution_sentiment_per_category.png', bucket=bucket, key_prefix="images")
```

Out[67]:

```
's3://sagemaker-us-east-1-392437363011/images/distribution_sentiment_per_category.png'
```

3.6. Analyze the distribution of review word counts

Set the SQL statement to count the number of the words in each of the reviews:

In [68]:

```
statement_num_words = """
    SELECT CARDINALITY(SPLIT(review_body, ' ')) as num_words
    FROM {}
""".format(table_name)

print(statement_num_words)
```

```
    SELECT CARDINALITY(SPLIT(review_body, ' ')) as num_words
    FROM reviews
```

Query data in Amazon Athena database passing the SQL statement:

In [69]:

```
%%time
df_num_words = wr.athena.read_sql_query(
    sql=statement_num_words,
    database=database_name
)
```

```
CPU times: user 305 ms, sys: 30.9 ms, total: 336 ms
Wall time: 3.05 s
```

Print out and analyse some descriptive statistics:

In [70]:

```
summary = df_num_words["num_words"].describe(percentiles=[0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 1.00])
summary
```

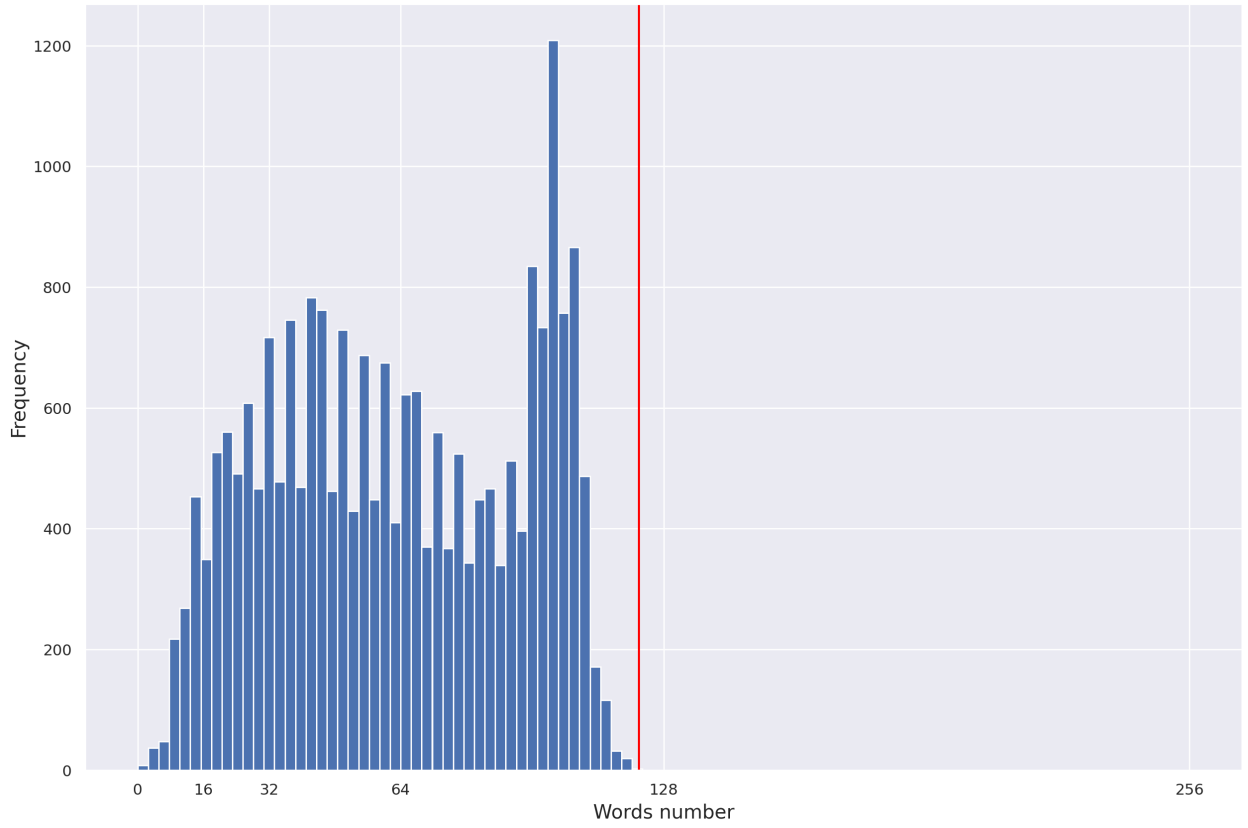
Out[70]:

```
count    22626.000000
mean        62.709847
std         29.993735
min          2.000000
10%         22.000000
20%         33.000000
30%         42.000000
40%         51.000000
50%         61.000000
60%         72.000000
70%         86.000000
80%         97.000000
90%        103.000000
100%       122.000000
max        122.000000
Name: num_words, dtype: float64
```

Plot the distribution of the words number per review:

In [71]:

```
df_num_words["num_words"].plot.hist(xticks=[0, 16, 32, 64, 128, 256], bins=100, range=[0, 256]).axvline(
    x=summary["100%"], c="red"
)

plt.xlabel("Words number", fontsize='14')
plt.ylabel("Frequency", fontsize='14')
plt.savefig('distribution_num_words_per_review.png', dpi=300)
plt.show()
```



In [72]:

```
# Upload image to S3 bucket
sess.upload_data(path='distribution_num_words_per_review.png', bucket=bucket, key_prefix="images")
```

Out[72]:

```
's3://sagemaker-us-east-1-392437363011/images/distribution_num_words_per_review.png'
```

Upload the notebook into S3 bucket for grading purposes.

Note: you may need to click on "Save" button before the upload.

In [73]:

```
!aws s3 cp ./C1_W1_Assignment.ipynb s3://$bucket/C1_W1_Assignment_Learner.ipynb
```

```
upload: ./C1_W1_Assignment.ipynb to s3://sagemaker-us-east-1-392437363011/C1_W1_Assignment_Learner.ipynb
```

Please go to the main lab window and click on `Submit` button (see the `Finish the lab` section of the instructions).

In [ ]:

```

```