Examen – Fouille de données SHS

Durée : 45min + 45min = 1h30

1. Le *big data*

(1) Qu'est-ce que les 3 Vs ? Donnez au moins 3 exemples de "V" et expliquez en quoi cela est du *big data*.
(2) Décrivez le fonctionnement de *map reduce* : vous pouvez illustrer votre pensée par un schéma/dessin.
(3) Que veut dire *lazy* dans la phrase "les *dataframe* de *spark* sont *lazy*" ?
(4) On exécute le code suivant :

```
df_ratings.show(5)
```

```
+-------+-------+------+---------+
|user_id|item_id|rating|timestamp|
+-------+-------+------+---------+
|    196|    242|     3|881250949|
|    186|    302|     3|891717742|
|     22|    377|     1|878887116|
|    244|     51|     2|880606923|
|    166|    346|     1|886397596|
+-------+-------+------+---------+
only showing top 5 rows
```

```
df_ratings.filter(df_ratings['user_id']==result_1[0][0])\
    .groupBy('user_id')\
    .count()\
    .show(1)
```

Qu'affiche-t-il (soyez précis) ?
(5) Quel est le principe de *hadoop file system* ?
(6) Quelle est la différence entre un RDD et un dataframe ?
(7) Pourquoi avons-nous besoin de la méthode `cache` des *dataframes* ? Donnez un exemple concret et précis d'utilisation.
(8) Décrivez le code suivant :

```
stream = ssc.textFileStream('data/output/')
stream.foreachRDD(process)
ssc.start()
```

(9) Complétez le code suivant :

```
from pyspark.ml import Pipeline
from pyspark.ml.classification import LogisticRegression
from pyspark.ml.feature import HashingTF, Tokenizer

tokenizer = Tokenizer(inputCol="sms", outputCol="words")
hashingTF = HashingTF(inputCol=tokenizer.getOutputCol(), outputCol="
    features")
lr = LogisticRegression(maxIter=10, regParam=0.001)

pipeline = Pipeline(stages=.........])
```

Justifiez l'intérêt de `Pipeline`.

(10) Après avoir entraîné un modèle, on exécute le code suivant :

```
predictions = model.transform(test)
predictions.printSchema()
```

```
root
 |-- label: float (nullable = true)
 |-- sms: string (nullable = true)
 |-- id: long (nullable = false)
 |-- words: array (nullable = true)
 |     |-- element: string (containsNull = true)
 |-- features: vector (nullable = true)
 |-- rawPrediction: vector (nullable = true)
 |-- probability: vector (nullable = true)
 |-- prediction: double (nullable = false)
```

Que sont les colonnes `features`, `rawPrediction`, `probability` et `prediction` ?