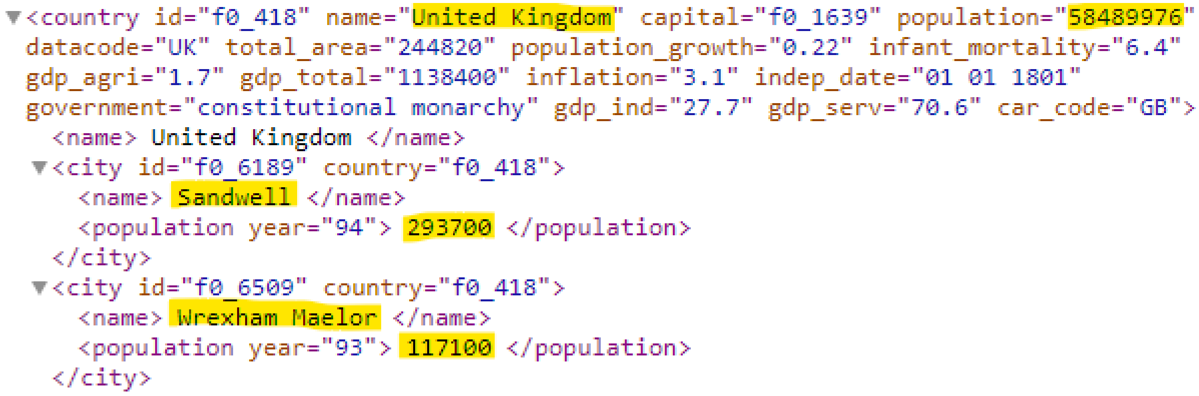
Spark SQL XML World Fact

1. Download the following XML file which contains facts about countries and cities.

   https://gitlab.com/opstar/share20/-/raw/master/worldfact.xml

   In the file you will find there is a hierarchy of cities within countries with the population of each city and country.



```
▼<country id="f0_418" name="United Kingdom" capital="f0_1639" population="58489976"
 datacode="UK" total_area="244820" population_growth="0.22" infant_mortality="6.4"
 gdp_agri="1.7" gdp_total="1138400" inflation="3.1" indep_date="01 01 1801"
 government="constitutional monarchy" gdp_ind="27.7" gdp_serv="70.6" car_code="GB">
   <name> United Kingdom </name>
  ▼<city id="f0_6189" country="f0_418">
     <name> Sandwell </name>
     <population year="94"> 293700 </population>
   </city>
  ▼<city id="f0_6509" country="f0_418">
     <name> Wrexham Maelor </name>
     <population year="93"> 117100 </population>
   </city>
```

2. Put the file into HDFS and open it as a dataframe using Spark

3. Extract the country names, country populations, city names and city populations.

   Hint: you will probably find it easier to extract the country and city parts of the XML separately and then join them together.

4. Calculate each city's population as a percentage of the national population and display the results in descending order of the percentage.

   Hint: the city names contain newline characters and spaces. To remove those characters you can use the trim and regexp_replace functions. You will need to import the functions first from pyspark.sql.functions. For example (assuming you have named the city name column as cityname):

```
.withColumn('cityname',trim(regexp_replace('cityname','[\n]+','')))
```

Expected result:

```
+--------------------+----------+-------------+-------+-------+
|         countryname|countrypop|     cityname|citypop|citypct|
+--------------------+----------+-------------+-------+-------+
|       Liechtenstein|     31122|        Vaduz|  27714|     89|
|           Singapore|   3396924|    Singapore|2558000|     75|
| Antigua and Barbuda|     65647|  Saint Johns|  36000|     54|
|             Bahamas|    259367|       Nassau| 140000|     53|
|            Holy See|       840| Vatican City|    392|     46|
|               Palau|     16952|        Koror|   7685|     45|
|               Qatar|    547761|         Doha| 217294|     39|
|             Uruguay|   3238952|   Montevideo|1247000|     38|
|              Latvia|   2468982|         Riga| 900000|     36|
```