Algorithms and Parallel Computing

Exercise 1 (6 points)

The following Apache Spark code processes the data coming from Stack Overflow.

```
import collections

Question = collections.namedtuple("Question", "id id_user title text keywords views votes")
Answer = collections.namedtuple("Answer", "id id_question id_user text")
User = collections.namedtuple("User", "id reputation profile")

q1= Question (1,1,"Cassandra Upsert not working on conditional writes",
            """I made a conditional insert (if not exists) \\\
            statement using DataStax java driver but it doesn't work""",
            "Java Cassandra DataStax", 1, 0)
q2= Question (2,1,"New Spark 2.2 Cassandra Connector",
            """ Tried to run the new connector to Spark 2.2 got error code 99129
            who can be of help?""",
            "Spark Cassandra", 2, 3)
u1= User(1, 1, "I'm an indipendent programmer, 8 years expertise in Java dev");
u2= User(2, 5, "I'm Matei, Spark creator");
u3= User(3, 5, "I'm Guido, Python benevolent dictator");

a1= Answer(1,1,2,"I think there is still a problem in DataStax connector, try to use the one at this link
     XXX")
a2= Answer(2,2,2,"Did you check server IP and Scala version?")
a3= Answer(3,2,3,"I think you are using Python 2.7, while the new API is for Python 3.0")
questionsRDD=sc.parallelize([q1,q2])
usersRDD=sc.parallelize([u1,u2,u3])
answersRDD=sc.parallelize([a1,a2,a3])
```

Provide the following queries:

- Select the power users (i.e., the users with the maximum reputation) providing also their profile
- Select the keywords of the questions answered by the power users

Exercise 2 (6 points)

The following Apache Spark code processes the data of well sites in California.

```
import collections

Perforation = collections.namedtuple("Perforation", "SITE_CODE PERFORATION_TOP_MSRMNT \
                                     PERFORATION_BOTTOM_MSRMNT LATITUDE LONGITUDE")

Site = collections.namedtuple("Site", "SITE_CODE BASIN_DESC COUNTY_NAME SITE_USE_DESC")

perforation1=Perforation("379583N1219669W001",46,110,38.559,-122.5215)
perforation2=Perforation("379632N1219700W001", 206, 223, 38.5292,-122.5015)
perforation3=Perforation("379178N1216700W001",50,150,35.6347,-117.7226)

site1=Site("379583N1219669W001","Enterprise","Napa","Residential")
site2=Site("379632N1219700W001","Millville","Kern","Irrigation")
site3=Site("379178N1216700W001","South Battle Creek", "Santa Barbara", "Unknown")

perforationRDD=sc.parallelize([perforation1, perforation2, perforation3])
siteRDD=sc.parallelize([site1, site2, site3])
```

Provide the following queries:

- Provide name, latitude and longitude of the site with deepest deep
- Provide name and county of sites devoted to Irrigation