Analysis on Sparks Funds Investment

18CN627 - Big Data Framework for Data Science

Submitted by,
Vyshnav M T (CB.EN.P2CEN18021)

Introduction:

Spark Funds is an asset management company. The company wants to make investments in few other companies. The CEO of Spark Funds wants to understand the global trends in investments so that she can take the investment decisions effectively. The company has two minor constraints for investments such as:

- a) To invest between 5 to 15 million USD per round of investment.
- b) To invest only in English-speaking countries because of the ease of communication with the companies it would invest.

In this work, an analysis on the previously available investment data is performed to identify best investment strategy for Sparks funds company. The main strategy is to invest where other companies are investing, implying that the best sectors and countries are the ones where most investors are investing.

Dataset Description:

The dataset is based on the real time investment data taken from crunchbase.com [1, 2, 3]. The dataset has three data files. The company details with basic data on companies are available in companies.txt. The funding round details are present in rounds2.csv. Finally the different sector based classification of data is available in mapping.csv, which maps various category names in company's data such as 3D printing, aerospace, agriculture, etc to eight broad sector names. Each of these data files are explained in detail in tables [1-3]

Attribute	Description
Permalink	Unique ID of company
name	Company name
homepage_url	Website URL
category_list	Category/categories to which a company belongs
status	Operational status
country_code	Country Code
state_code	State

Table 1: Company details in companies.txt file

Attributes	Description				
company_permalink	Unique ID of company				
funding_round_permalink	Unique ID of funding round				
funding_round_type	Type of funding – venture, angel, private equity etc.				
funding_round_code	Round of venture funding (round A, B etc.)				
funded_at	Date of funding				
raised_amount_usd	Money raised in funding (USD)				

Table 2: Funding rounds details in rounds.csv file

Attributes	Description					
Category_list	gory_list					
Main sector	Main sector Eight main sectors such as Automotive&Sports, Blanks,					
	Cleantech/Semiconductors, Entertainment, Health, Manufacturing,					
	News_Search_Messaging, Others, Social_Finance_Analytics_Advertising					

Table 3: Sector based classification of data in mapping.csv file.

Methodology:

The mail goal of this work is to identify best sectors, countries and a suitable investment type for making investments. The analysis is done in three different stages as follows:

1. Investment type analysis:

Comparison of the typical investment amounts in the funding types such as venture, seed, angel, private equity etc. is performed in this stage. This helps Spark Funds to choose the type that is best suited for their strategy. The funding types such as seed, venture, angel, etc. depend on the type of the company (start-up, corporate, etc.), its stage (early stage start-up, funded start-up, etc.), the amount of funding (a few million USD to a billion USD), and so on. For instance seed, angel and venture are three common stages of start-up funding. Seed/angel funding refer to early stage start-ups whereas venture funding occurs after seed or angel stages and involves a relatively higher amount of investment. Private equity type investments are associated with much larger companies and involve much higher investments than venture type. Start-ups which have grown in scale may also receive private equity funding. This means that if a company has reached the venture stage, it would have already passed through the angel or seed stages. In order to decide the funding type which is most suitable for sparks funds to invest, the total investment amount for each of the funding types is calculated. From this calculation, the funding type which has highest investment amounts is chosen.

2. Country analysis:

This stage identifies the countries which have been most heavily invested in the past. From the above stage the best investment type for sparks funds is found, which is to be

narrowed down to countries. According to the spark funds constraints, the top English speaking countries with highest amount of funding needs to be found. To achieve this, first the top nine countries having highest amount of funding is identified and from these top three English speaking countries are chosen. Now having known the three most investment-friendly countries and the most suited funding type for Spark Funds, next stage is to find the best sectors in these countries.

3. Sector analysis:

Understanding the distribution of investments across the eight main sectors. Even though the companies and rounds data have various sub-sectors, only eight main sectors provided in the mapping file is considered by mapping various sub-sector (primary sector) names in companies and rounds2 to its main sector. For instance, category lists such as '3D', '3D Printing', '3D Technology', etc. are mapped to the main sector 'Manufacturing'. Now having each company's main sector mapped to it, the analysis on these main sectors is to be performed. One of the other constraint of spark funds is to invest between 5 to 15 million USD per round of investment. Therefore, the final aim is to find the most heavily invested main sectors in each of the three best English speaking countries for best identified funding types in the range of 5 to 15 million USD.

Experiments and Results:

The experiments are performed as per the three stages mentioned in the methodology. Figure 1 and 2 shows the sample data of the companies.txt and rounds.csv respectively. The companies data has total of 66368 rows and rounds data has only 66370 rows, which means two data points are missing in companies data. Now in order to perform the first stage of analysis, the companies data needs to be merged with the rounds.csv data. The two data frames are joined on the permalink column of rounds dataframe, by lowercasing all permalink data in both companies and rounds dataframes. The join operation is performed by intersection of keys from both dataframes. The resulting merged dataframe, master_frame has total of 114942 rows. Figure 3 shows the merged dataframe master_frame. The master_frame has many attributes which are not useful for the analysis such as funding_round_code,funding_round_permalink,funded_at,permalink,homepage_url,state_c ode,region, city, founded_at,status. The master_frame id cleaned by dropping all of these non-contributing attributes and null values. The cleaned master_frame has total of 88529 rows. Figure 3 shows the sample data of master frame.

1	+	+				+				+
	permalink	name	homepage_url	category_list	status	country_code	state_code	region	city	founded_at
	/Organization/-Fame				operating operating	IND USA	16 DE DE	Mumbai - Other	Mumbai Delaware City	
	only showing top 2 rows	-				+				

Figure 1: company details in companies.txt.

+	 funding_round_permalink	funding_round_type	funding_round_code	funded_at	 raised_amount_usd
/ /organization/-fame /ORGANIZATION/-QO				05-01-2015 14-10-2014	'

Figure 2: Funding rounds details in rounds.csv.

 funding_round_type raise	ed_amount_usd	name	category_list	country_code
venture seed	100000000 700000		Media Application Platf	

Figure3: Master frame sample data.

The cleaned master_frame is now analysed to obtain the funding type with highest investment amounts. The number of investment and the sum of investment for each funding type is calculated from the master_frame. From figure 4, it is observed that the funding type Venture has the highest number of investments as well as highest investment amount.

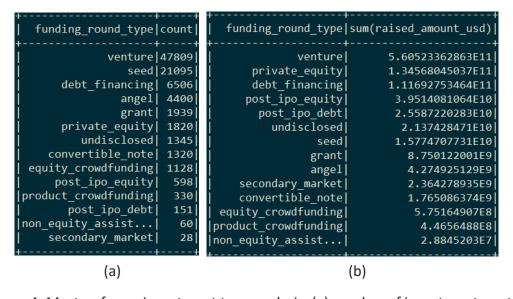


Figure 4: Master_frame investment type analysis: (a) number of investment per type, (b) investment amount for each type.

Now after filtering out the master_frame with investment type as venture, the country analysis is performed. As per spark funds constraint, the investment need to be done in top English speaking countries. The english speaking countries are short listed as USA, GBR, IND, by manual analysis of the list provided for english as official language list. Figure 5 shows the filtered out master_frame with investment type as venture and top 3 english speaking

countries as United States of America (USA), Britain (GBR), India (IND). The country analysis result in figure 6 shows that USA has highest number of investment as well as highest investment amount.

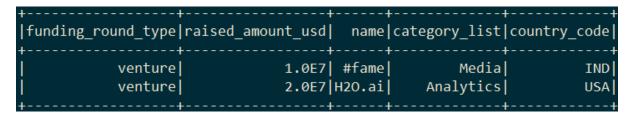


Figure 5: Filtered master_frame with investment type venture and english speaking countries USA, GBR, IND.

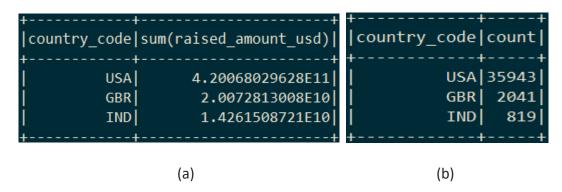


Figure 6: Master_frame country analysis: (a) investment amount for each country, (b) number of investment per country.

The third stage of analysis is performed sector wise on the eight main sectors available in mapping.csv file. Figure 7 shows the sample data of mapping.csv file. The mapping data has the eight main sectors in a pivoted format, which is unpivoted and used to map each primary sector in master_frame to one of the eight main sectors. The mapped master frame is filtered out to investment amount between 5-15 million USD, as per the constraint of spark funds. The master_frame is further filtered to obtain three dataframes D1, D2, D3 based on the country code. The sector based analysis is performed on these country filtered dataframe.

	category_lis	: Automotive_	+- Sports B	lanks	Cleantech_Semiconductors	Entertainment	 Health	Manufacturing	News_Search_Messaging	Others	 Social_Finance_Analytics_Advertising
	nul. 31		0 0	1 0		0 0	0	0 1	0	0 0	
4		.+	+-	+						+	++

Figure 7: Sector details in mapping.csv.

The results of sector based analysis is shown in figure 8 & 9 respectively. Figure 8 shows the investment amounts of each of the eight main sector for each countries USA, GBR and IND. Similarly, figure 9 shows the number of investment in each main sectors for these three countries. From the results obtained, it is inferred that the best investment strategy for spark funds is to invest in Others main sector in USA with investment type as venture. Further investment can be made on Britain and India under main sectors Cleantech/semiconductors and Others respectively.



Figure 8: Total investment amounts of each main sectors in countries: (a) USA, (b) GBR, (c) IND.

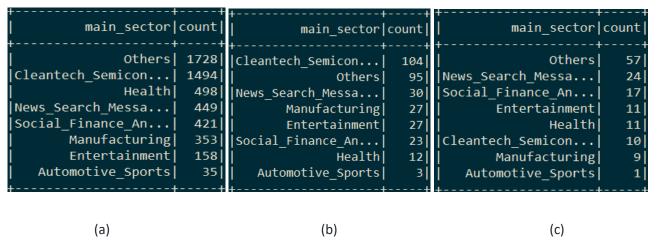


Figure 9: Total number of investments in each main sectors in countries: (a) USA, (b) GBR, (c) IND.

Conclusion:

Based on the data analysis performed, Sparks Funds should invest in:

- Funding type Venture.
- Countries USA, Britain and India.
- Top two sectors to invest in are Others and Cleantech/semiconductors.

References:

- [1] https://www.crunchbase.com/, accessed on October 2019.
- [2] https://www.kaggle.com/goyalshalini93/data, accessed on October 2019.
- [3] https://github.com/santhoshpkumar/Spark-Funds-Investment-CaseStudy/blob/master/Spark%20Funds%20Presentation.pdf , accessed on October 2019.

Codes:

```
======Task 1: Reading, merging and Data cleaning================================
var companies = spark.read.format("csv").option("sep","\t").option("header","True").
option("encoding","ISO-8859-1").load("companies.txt")
companies = companies.withColumn("permalink",lower(col("permalink")))
companies.select("permalink").distinct.count()
var rounds = spark.read.format("csv").option("header","True").option("encoding","ISO-
8859-1").load("rounds2.csv")
rounds = rounds.withColumn("company permalink",lower(col("company permalink")))
rounds.select("company permalink").distinct.count()
rounds = rounds.withColumnRenamed("company permalink","permalink")
var master_frame = rounds.join(companies, Seq("permalink"), "inner")
var master = master_frame.drop("funding_round_code", "funding_round_permalink",
"funded at", "permalink", "homepage url", "state code", "region", "city",
"founded_at","status")
master.count()
master = master.na.drop()
master.count()
======= Observing the unique funding_round_type ==================
master.groupBy("funding round type").count().orderBy(desc("count")).show
master = master.withColumn("raised amount usd",$"raised amount usd".cast("float"))
master.groupBy("funding_round_type").agg(sum("raised_amount_usd")).orderBy(desc("su
m(raised amount usd)")).show
master.groupBy("funding round type").agg(avg("raised amount usd")).orderBy(desc("avg(
raised amount usd)")).show
```

```
val fil = List("venture", "angel", "seed", "private_equity")
master.filter($"funding round type".isin(fil: *)).show(3)
master.groupBy("funding_round_type").agg(avg("raised_amount_usd")).orderBy(desc("avg(
raised amount usd)")).show
master.groupBy("funding_round_type").count().show
master = master.filter($"funding_round_type"==="venture")
=======creating new dataframe with highest funding countries========
master.groupBy("country_code").agg(avg("raised_amount_usd")).orderBy(desc("avg(raised
_amount_usd)")).show
master.groupBy("country code").count().orderBy(desc("count")).show
master.groupBy("country code").agg(sum("raised amount usd")).orderBy(desc("sum(raise
d amount usd)")).show
val top9 country = List("USA", "CHN", "GBR", "IND", "CAN", "FRA", "ISR", "DEU", "JPN")
var top9 = master.filter($"country code".isin(top9 country: *))
top9.groupBy("country_code").agg(sum("raised_amount_usd")).orderBy(desc("sum(raised_
amount usd)")).show
=======Identify the top three English-speaking countries in the data frame top9.
            The countires has been short listed by manual analysis of the list provided
            for english as offical lanaguage list===========
val english = List("USA", "GBR", "IND")
var top3 english = top9.filter($"country code".isin(english: *))
top3 english.show(2)
top3 english.groupBy("country code").agg(sum("raised amount usd")).orderBy(desc("sum
(raised_amount_usd)")).show
top3 english.groupBy("country code").count().orderBy(desc("count")).show
```

```
var mapping = spark.read.format("csv").option("header","True").option("encoding","ISO-
8859-1").load("mapping.csv")
```

mapping =

mapping.selectExpr("category_list","stack(9,'Automotive_Sports',Automotive_Sports,'Blank s',Blanks,'Cleantech_Semiconductors',Cleantech_Semiconductors,'Entertainment',Entertainment,'Health,'Manufacturing',Manufacturing,'News_Search_Messaging',News_Search_Messaging,'Social_Finance_Analytics_Advertising,'Others',Others)").withColumnRenamed("col0","main_sector").withColumnRenamed("col1", "value").filter(\$"value"=!=0).filter(\$"category_list".isNotNull)

```
mapping = mapping.drop("value")
var top3 = top3_english.join(mapping, Seq("category_list"), "left")
```

======Drop all rows whose investment is not between 5 and 15 million=======

```
top3 = top3.filter($"raised_amount_usd" > 5000000)

top3 = top3.filter($"raised_amount_usd" < 15000000)

top3 = top3.na.drop()

var d1 = top3.filter($"country_code"==="USA")

var d2 = top3.filter($"country_code"==="GBR")

var d3 = top3.filter($"country_code"==="IND")

d1.groupBy("main_sector").agg(sum("raised_amount_usd")).orderBy(desc("sum(raised_amount_usd)")).show

d2.groupBy("main_sector").agg(sum("raised_amount_usd")).orderBy(desc("sum(raised_amount_usd)")).show

d3.groupBy("main_sector").agg(sum("raised_amount_usd")).orderBy(desc("sum(raised_amount_usd)")).show

d1.groupBy("main_sector").count().orderBy(desc("count")).show

d2.groupBy("main_sector").count().orderBy(desc("count")).show

d3.groupBy("main_sector").count().orderBy(desc("count")).show
```