In [6]:
```
sc
```

Out[6]:

SparkContext

Spark UI

Version	v3.1.2
Master	yarn
AppName	pyspark-shell

In [1]:
```
from pyspark.sql import SQLContext
sqlContext = SQLContext(sc)
```

In [43]:
```
json_list = []
for month in range(6, 12):
        json_list.append("/data/ProjectDatasetFacebook/FBads-US-2020" + str(month)
print(json_list)
```

['/data/ProjectDatasetFacebook/FBads-US-20206*', '/data/ProjectDatasetFacebook/FBads-US-20207*', '/data/ProjectDatasetFacebook/FBads-US-20208*', '/data/ProjectDatasetFacebook/FBads-US-20209*', '/data/ProjectDatasetFacebook/FBads-US-202010*', '/data/ProjectDatasetFacebook/FBads-US-202011*']

[Stage 75:==========> (3154 + 149) / 17066]

In [44]:
```
df = sqlContext.read.json(json_list)
```

```
2022-05-21 16:41:59,141 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Req
uesting driver to remove executor 130 for reason Container marked as failed: contain
er_1652753568617_1390_01_000176 on host: ip-100-64-74-121.ap-southeast-2.compute.int
ernal. Exit status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:41:59,141 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Req
uesting driver to remove executor 157 for reason Container marked as failed: contain
er_1652753568617_1390_01_000203 on host: ip-100-64-74-121.ap-southeast-2.compute.int
ernal. Exit status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:41:59,141 ERROR cluster.YarnScheduler: Lost executor 130 on ip-100-64-
74-121.ap-southeast-2.compute.internal: Container marked as failed: container_165275
3568617_1390_01_000176 on host: ip-100-64-74-121.ap-southeast-2.compute.internal. Ex
it status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:41:59,145 ERROR cluster.YarnScheduler: Lost executor 157 on ip-100-64-
74-121.ap-southeast-2.compute.internal: Container marked as failed: container_165275
3568617_1390_01_000203 on host: ip-100-64-74-121.ap-southeast-2.compute.internal. Ex
it status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:41:59,149 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Req
uesting driver to remove executor 174 for reason Container marked as failed: contain
er_1652753568617_1390_01_000220 on host: ip-100-64-74-121.ap-southeast-2.compute.int
ernal. Exit status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:41:59,150 ERROR cluster.YarnScheduler: Lost executor 174 on ip-100-64-
74-121.ap-southeast-2.compute.internal: Container marked as failed: container_165275
3568617_1390_01_000220 on host: ip-100-64-74-121.ap-southeast-2.compute.internal. Ex
it status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:43:55,820 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Req
uesting driver to remove executor 214 for reason Container marked as failed: contain
er_1652753568617_1390_01_000260 on host: ip-100-64-74-25.ap-southeast-2.compute.inte
rnal. Exit status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:43:55,820 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Req
uesting driver to remove executor 159 for reason Container marked as failed: contain
er_1652753568617_1390_01_000205 on host: ip-100-64-74-25.ap-southeast-2.compute.inte
rnal. Exit status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:43:55,820 ERROR cluster.YarnScheduler: Lost executor 214 on ip-100-64-
74-25.ap-southeast-2.compute.internal: Container marked as failed: container_1652753
568617_1390_01_000260 on host: ip-100-64-74-25.ap-southeast-2.compute.internal. Exit
status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:43:55,824 ERROR cluster.YarnScheduler: Lost executor 159 on ip-100-64-
74-25.ap-southeast-2.compute.internal: Container marked as failed: container_1652753
568617_1390_01_000205 on host: ip-100-64-74-25.ap-southeast-2.compute.internal. Exit
status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:43:55,843 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Req
uesting driver to remove executor 169 for reason Container marked as failed: contain
er_1652753568617_1390_01_000215 on host: ip-100-64-74-25.ap-southeast-2.compute.inte
rnal. Exit status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:43:55,843 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Req
uesting driver to remove executor 204 for reason Container marked as failed: contain
er_1652753568617_1390_01_000250 on host: ip-100-64-74-25.ap-southeast-2.compute.inte
rnal. Exit status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:43:55,843 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Req
uesting driver to remove executor 212 for reason Container marked as failed: contain
er_1652753568617_1390_01_000258 on host: ip-100-64-74-25.ap-southeast-2.compute.inte
rnal. Exit status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:43:55,843 ERROR cluster.YarnScheduler: Lost executor 169 on ip-100-64-
74-25.ap-southeast-2.compute.internal: Container marked as failed: container_1652753
568617_1390_01_000215 on host: ip-100-64-74-25.ap-southeast-2.compute.internal. Exit
status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:43:55,844 ERROR cluster.YarnScheduler: Lost executor 204 on ip-100-64-
74-25.ap-southeast-2.compute.internal: Container marked as failed: container_1652753
568617_1390_01_000250 on host: ip-100-64-74-25.ap-southeast-2.compute.internal. Exit
status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:43:55,846 ERROR cluster.YarnScheduler: Lost executor 212 on ip-100-64-
74-25.ap-southeast-2.compute.internal: Container marked as failed: container_1652753
568617_1390_01_000258 on host: ip-100-64-74-25.ap-southeast-2.compute.internal. Exit
status: -100. Diagnostics: Container released on a *lost* node.
```

```
2022-05-21 16:50:16,871 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Req
uesting driver to remove executor 131 for reason Container marked as failed: contain
er_1652753568617_1390_01_000177 on host: ip-100-64-74-57.ap-southeast-2.compute.inte
rnal. Exit status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:50:16,871 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Req
uesting driver to remove executor 143 for reason Container marked as failed: contain
er_1652753568617_1390_01_000189 on host: ip-100-64-74-57.ap-southeast-2.compute.inte
rnal. Exit status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:50:16,871 ERROR cluster.YarnScheduler: Lost executor 131 on ip-100-64-
74-57.ap-southeast-2.compute.internal: Container marked as failed: container_1652753
568617_1390_01_000177 on host: ip-100-64-74-57.ap-southeast-2.compute.internal. Exit
status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:50:16,875 ERROR cluster.YarnScheduler: Lost executor 143 on ip-100-64-
74-57.ap-southeast-2.compute.internal: Container marked as failed: container_1652753
568617_1390_01_000189 on host: ip-100-64-74-57.ap-southeast-2.compute.internal. Exit
status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:50:16,880 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Req
uesting driver to remove executor 225 for reason Container marked as failed: contain
er_1652753568617_1390_01_000271 on host: ip-100-64-74-57.ap-southeast-2.compute.inte
rnal. Exit status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:50:16,881 ERROR cluster.YarnScheduler: Lost executor 225 on ip-100-64-
74-57.ap-southeast-2.compute.internal: Container marked as failed: container_1652753
568617_1390_01_000271 on host: ip-100-64-74-57.ap-southeast-2.compute.internal. Exit
status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:50:50,944 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Req
uesting driver to remove executor 137 for reason Container marked as failed: contain
er_1652753568617_1390_01_000183 on host: ip-100-64-74-23.ap-southeast-2.compute.inte
rnal. Exit status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:50:50,944 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Req
uesting driver to remove executor 201 for reason Container marked as failed: contain
er_1652753568617_1390_01_000247 on host: ip-100-64-74-23.ap-southeast-2.compute.inte
rnal. Exit status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:50:50,944 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Req
uesting driver to remove executor 176 for reason Container marked as failed: contain
er_1652753568617_1390_01_000222 on host: ip-100-64-74-23.ap-southeast-2.compute.inte
rnal. Exit status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:50:50,945 ERROR cluster.YarnScheduler: Lost executor 137 on ip-100-64-
74-23.ap-southeast-2.compute.internal: Container marked as failed: container_1652753
568617_1390_01_000183 on host: ip-100-64-74-23.ap-southeast-2.compute.internal. Exit
status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:50:50,948 ERROR cluster.YarnScheduler: Lost executor 201 on ip-100-64-
74-23.ap-southeast-2.compute.internal: Container marked as failed: container_1652753
568617_1390_01_000247 on host: ip-100-64-74-23.ap-southeast-2.compute.internal. Exit
status: -100. Diagnostics: Container released on a *lost* node.
2022-05-21 16:50:50,950 ERROR cluster.YarnScheduler: Lost executor 176 on ip-100-64-
74-23.ap-southeast-2.compute.internal: Container marked as failed: container_1652753
568617_1390_01_000222 on host: ip-100-64-74-23.ap-southeast-2.compute.internal. Exit
status: -100. Diagnostics: Container released on a *lost* node.
```

In [51]:

```
from pyspark.sql import Row
df = df.dropDuplicates(subset = ['page_id'])
df.count()
```

Out[51]:

```
40964
```

In [74]:

```
df.printSchema()
```

```
root
 |-- _corrupt_record: string (nullable = true)
 |-- ad_creation_time: string (nullable = true)
 |-- ad_creative_body: string (nullable = true)
 |-- ad_creative_link_caption: string (nullable = true)
 |-- ad_creative_link_description: string (nullable = true)
 |-- ad_creative_link_title: string (nullable = true)
 |-- ad_delivery_start_time: string (nullable = true)
 |-- ad_delivery_stop_time: string (nullable = true)
 |-- ad_snapshot_url: string (nullable = true)
 |-- currency: string (nullable = true)
 |-- demographic_distribution: array (nullable = true)
 |    |-- element: struct (containsNull = true)
 |    |    |-- age: string (nullable = true)
 |    |    |-- gender: string (nullable = true)
 |    |    |-- percentage: string (nullable = true)
 |-- funding_entity: string (nullable = true)
 |-- id: string (nullable = true)
 |-- impressions: struct (nullable = true)
 |    |-- lower_bound: string (nullable = true)
 |    |-- upper_bound: string (nullable = true)
 |-- page_id: string (nullable = true)
 |-- page_name: string (nullable = true)
 |-- region_distribution: array (nullable = true)
 |    |-- element: struct (containsNull = true)
 |    |    |-- percentage: string (nullable = true)
 |    |    |-- region: string (nullable = true)
 |-- spend: struct (nullable = true)
 |    |-- lower_bound: string (nullable = true)
 |    |-- upper_bound: string (nullable = true)
```

In [57]:

```
entity_count = df.groupby('funding_entity').count()
entity_count.orderBy(entity_count['count'].desc()).show()
```

```
[Stage 119:=================================================>(199 + 1) / 200]
+--------------------+-----+
|      funding_entity|count|
+--------------------+-----+
|                null|  642|
|   Real Voices Media|  222|
|    Metric Media LLC|   87|
|                AARP|   57|
|                DCCC|   51|
|AMERICANS FOR PRO...|   32|
|Local Government ...|   32|
|Realtors® Politic...|   30|
|National Associat...|   29|
|  TECH FOR CAMPAIGNS|   26|
| Protect Our Winters|   25|
|  Nature Conservancy|   24|
|IOWA DEMOCRATIC P...|   24|
|  BIDEN VICTORY FUND|   23|
|Robert William Caton|   20|
|Maine House Majority|   20|
|DONALD J. TRUMP F...|   20|
|      Resource Media|   20|
|Protect Our Winte...|   20|
|U.S. Term Limits ...|   20|
+--------------------+-----+
only showing top 20 rows
```

In [59]:
```
entity_count = entity_count.orderBy(entity_count['count'].desc())
entity_count.take(20)
```

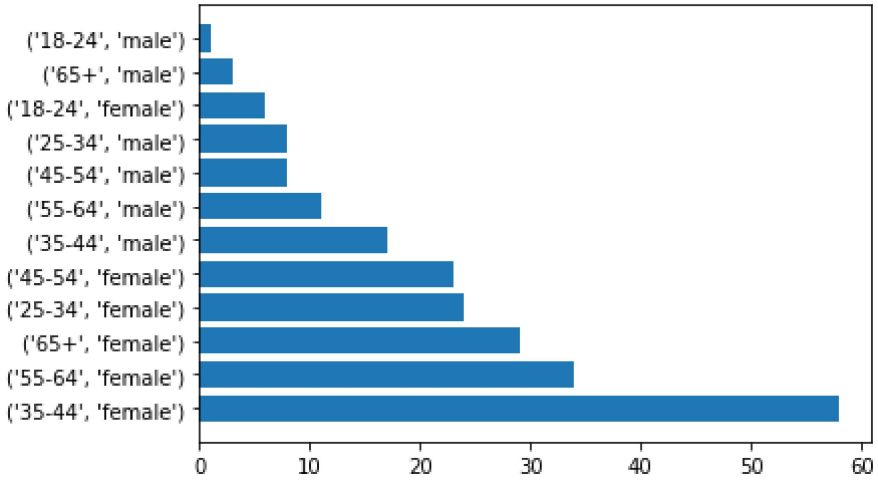
Out[59]:
```
[Row(funding_entity=None, count=644),
 Row(funding_entity='Real Voices Media', count=222),
 Row(funding_entity='Metric Media LLC', count=87),
 Row(funding_entity='AARP', count=58),
 Row(funding_entity='DCCC', count=51),
 Row(funding_entity='AMERICANS FOR PROSPERITY', count=32),
 Row(funding_entity='Local Government Information Services', count=32),
 Row(funding_entity='Realtors® Political Advocacy Committee', count=30),
 Row(funding_entity='National Association of REALTORS', count=29),
 Row(funding_entity='Protect Our Winters', count=28),
 Row(funding_entity='IOWA DEMOCRATIC PARTY', count=27),
 Row(funding_entity='TECH FOR CAMPAIGNS', count=26),
 Row(funding_entity='Nature Conservancy', count=24),
 Row(funding_entity='Maine House Majority', count=21),
 Row(funding_entity='BIDEN VICTORY FUND', count=21),
 Row(funding_entity='U.S. Term Limits Inc.', count=20),
 Row(funding_entity='Robert William Caton', count=20),
 Row(funding_entity='DONALD J. TRUMP FOR PRESIDENT, INC.', count=20),
 Row(funding_entity='American Heart Association, Inc.', count=18),
 Row(funding_entity='American Heart Association', count=18)]
```

In [60]:
```
# rvm = Real Voices Media
rvm_ad = df.filter(df['funding_entity'] == 'Real Voices Media')
```

In [61]:
```
from pyspark.sql.functions import col
rvm_dis = rvm_ad.select(col("demographic_distribution.percentage"), col("demographi
```

In [62]:
```
from operator import add
rvm_dis_rdd = rvm_dis.rdd.map(lambda x: ((x.age[x.percentage.index(max(x.percentag
rvm_dis_reduce = rvm_dis_rdd.reduceByKey(add)
rvm_dis_reduce = rvm_dis_reduce.sortBy(lambda x: x[1], False)
rvm_dis_reduce = rvm_dis_reduce.collect()
```

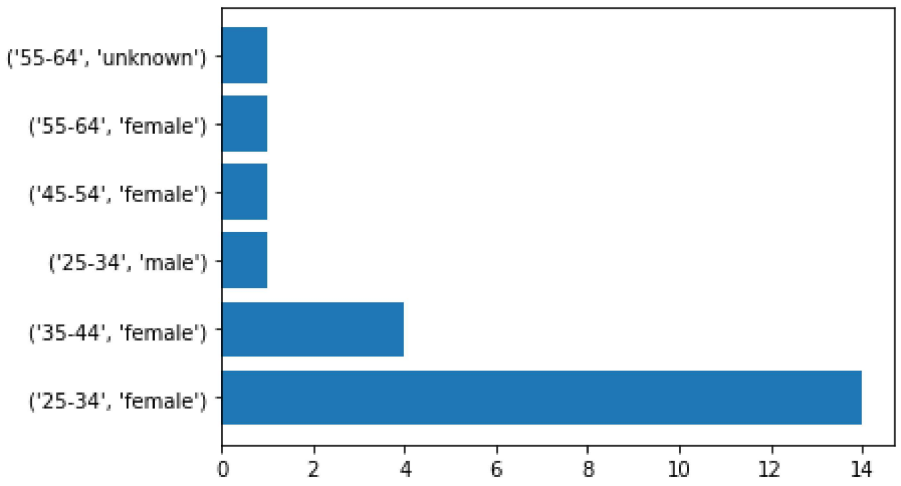
In [63]:
```
import matplotlib.pyplot as plt
import numpy as np
title = []
value = []
for item in rvm_dis_reduce:
    title.append(str(item[0]))
    value.append(item[1])
plt.barh(np.arange(len(title)), value)
plt.yticks([i for i in range(len(title))], tuple(title))
plt.show()
```



In [64]:
```
# bvf= BIDEN VICTORY FUND
from pyspark.sql.functions import col
bvf_ad = df.filter(df['funding_entity'] == 'BIDEN VICTORY FUND')
bvf_dis = bvf_ad.select(col("demographic_distribution.percentage"), col("demographi
```

In [66]:
```
from operator import add
bvf_dis_rdd = bvf_dis.rdd.map(lambda x: ((x.age[x.percentage.index(max(x.percentag
bvf_dis_reduce = bvf_dis_rdd.reduceByKey(add)
bvf_dis_reduce = bvf_dis_reduce.sortBy(lambda x: x[1], False)
bvf_dis_reduce = bvf_dis_reduce.collect()
# print(bvf_dis_reduce)
```

In [67]:
```
import matplotlib.pyplot as plt
import numpy as np
title = []
value = []
for item in bvf_dis_reduce:
    title.append(str(item[0]))
    value.append(item[1])
plt.barh(np.arange(len(title)), value)
plt.yticks([i for i in range(len(title))], tuple(title))
plt.show()
```



In [68]:
```
# dpi = DONALD J. TRUMP FOR PRESIDENT, INC.
from pyspark.sql.functions import col
```

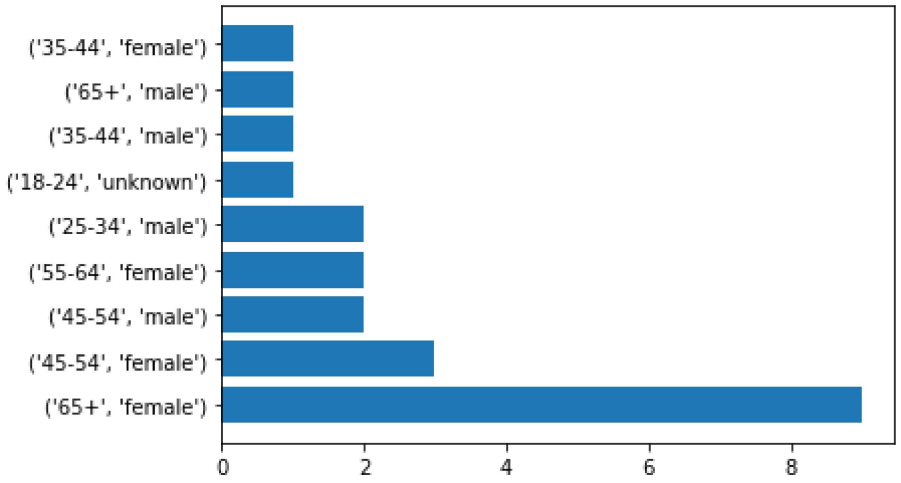
```
dpi_ad = df.filter(df['funding_entity'] == 'DONALD J. TRUMP FOR PRESIDENT, INC.')
dpi_dis = dpi_ad.select(col("demographic_distribution.percentage"), col("demographi
```

In [69]:

```
from operator import add
dpi_dis_rdd = dpi_dis.rdd.map(lambda x: ((x.age[x.percentage.index(max(x.percentag
dpi_dis_reduce = dpi_dis_rdd.reduceByKey(add)
dpi_dis_reduce = dpi_dis_reduce.sortBy(lambda x: x[1], False)
dpi_dis_reduce = dpi_dis_reduce.collect()
# print(sni_dis_reduce)
```

In [71]:

```
import matplotlib.pyplot as plt
import numpy as np
title = []
value = []
for item in dpi_dis_reduce:
    title.append(str(item[0]))
    value.append(item[1])
plt.barh(np.arange(len(title)), value)
plt.yticks([i for i in range(len(title))], tuple(title))
plt.show()
```



In [15]:

```
from pyspark.sql.functions import isnull
demo_dis = df.select("funding_entity", "spend")
demo_dis.filter(isnull("funding_entity")).count()
```

Out[15]: 128

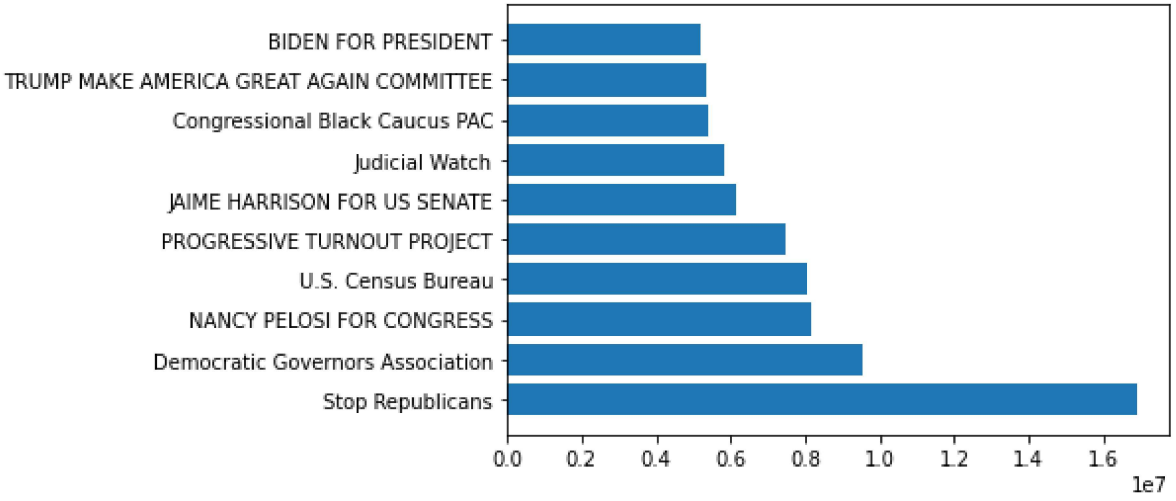
In [27]:

```
from operator import add
demo_dis = demo_dis.na.drop()
demo_dis_rdd = demo_dis.rdd.map(lambda x: (x.funding_entity, (int(x.spend[1]) + ir
demo_dis_reduce = demo_dis_rdd.reduceByKey(add)
demo_dis_reduce = demo_dis_reduce.sortBy(lambda x: x[1], False)
demo_dis_reduce.collect()
demo_dis_reduce = demo_dis_reduce.take(10)
```

In [28]:

```
import matplotlib.pyplot as plt
import numpy as np
```

```
entity = []
spend = []
for item in demo_dis_reduce:
    entity.append(item[0])
    spend.append(item[1])
plt.barh(np.arange(len(entity)), spend)
plt.yticks([i for i in range(len(entity))], tuple(entity))
plt.show()
```



In [26]:

```
print(demo_dis_reduce)
```

```
[('AMERICANS FOR PROSPERITY', 910428.5), ('Stanley R Jr Odell', 6020.5), ('Jose Luis Lopez', 26788.0), ('Christopher Anthony Vigliotti', 247.5), ('Jason H Beauregard', 497.0), ('Cardinale For Assembly', 594.0), ('Marcel Alonzo Santiz', 396.0), ('Katrina Gaar Altieri', 396.0), ('William Douglass Campaign Fund, Republician for District 1 County Commissioner', 1188.0), ('JoAnna for Wisconsin', 495.0)]
```