First look at the data, comparing feature importance from RandomForest classifier

In [13]:

```
import warnings
warnings.filterwarnings('ignore')
```

In [14]:

```
from utils_all import *

from sklearn.cross_validation import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.preprocessing import OneHotEncoder, LabelEncoder

import pandas as pd
import numpy as np
import os

import matplotlib.pyplot as plt
```

In [22]:

```
%store -r data
```

In [16]:

```
data = df_1_final
```

In [23]:

```
css_prop = data.iloc[:,9:]
```

In [24]:

```
data_cl = clean_df(data)
```

In [25]:

```
data_cl.shape
```

Out[25]:

```
(81618, 297)
```

In [26]:

```
data_cl.url.unique().shape
```

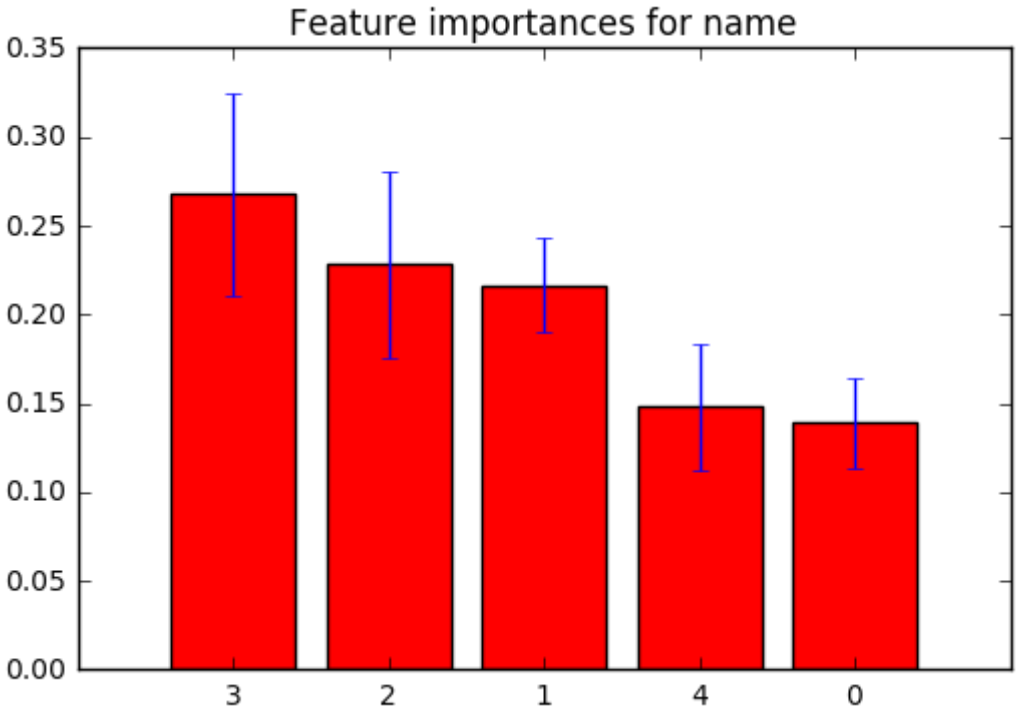
Out[26]:

```
(41550,)
```

In [27]:

```
forest = RandomForestClassifier(n_estimators=100)
perform_analysis_of_field('name', forest, data)
```

```
train: 0.999975518397924, test: 0.9876236393458919
Feature ranking:
3. feature 'block_width' (0.267927)
2. feature 'block_height' (0.228328)
1. feature 'y_coords' (0.216552)
4. feature 'num_siblings' (0.147920)
0. feature 'x_coords' (0.139274)
```



In [ ]:

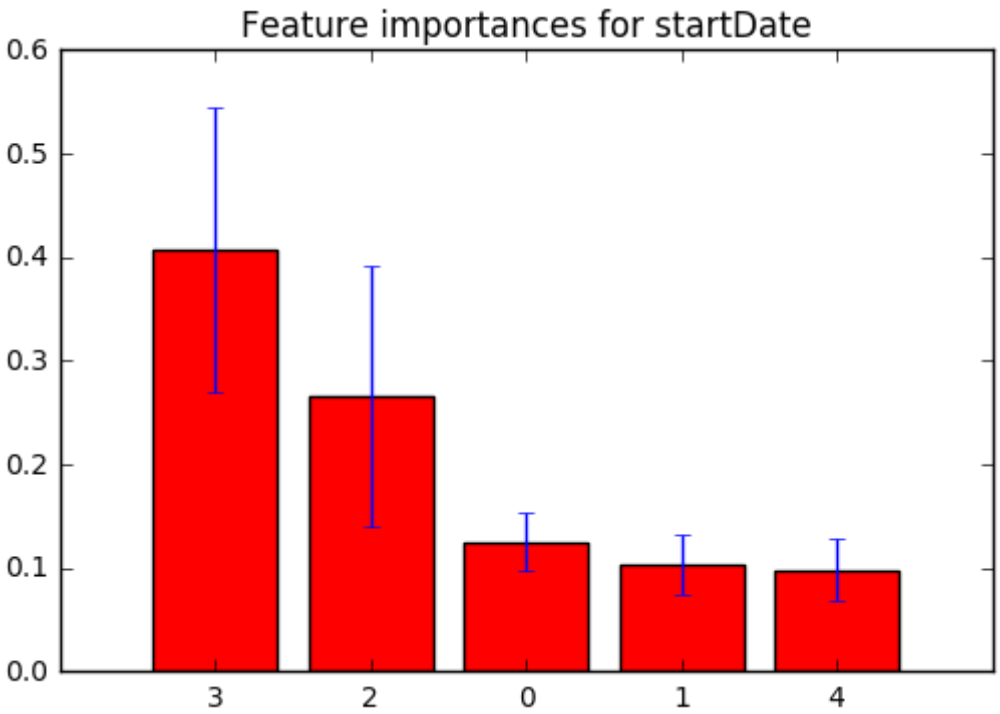
```

```

In [13]:

```
forest = RandomForestClassifier(n_estimators=100)
perform_analysis_of_field('startDate', forest, data)
```

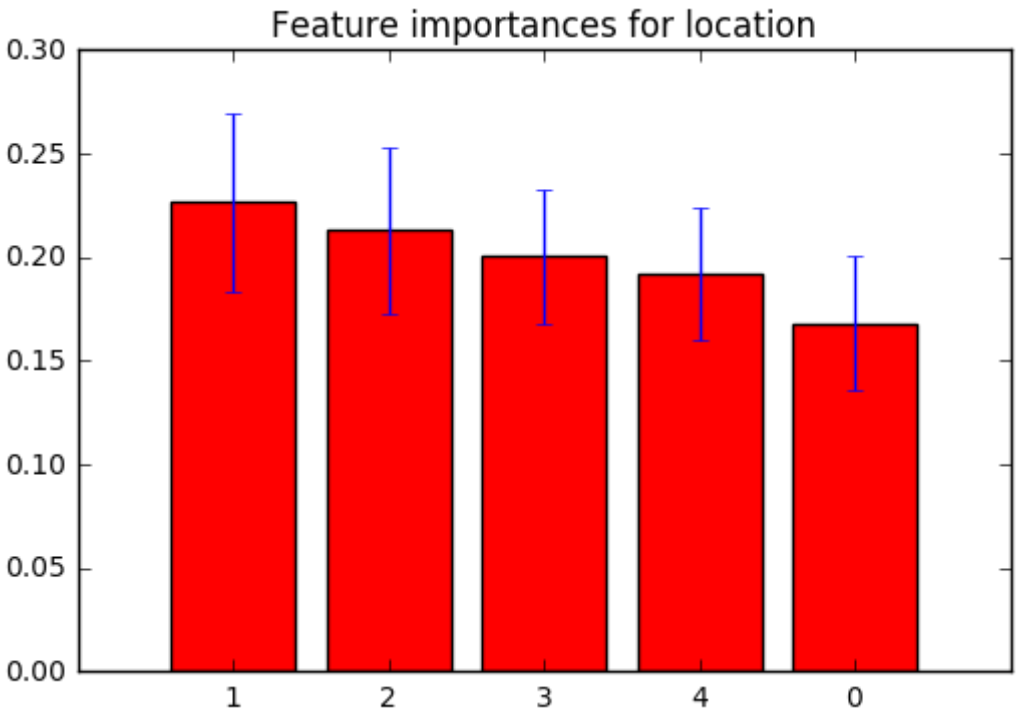
```
train: 0.9999506026476981, test: 0.98676293622142
Feature ranking:
3. feature 'block_width' (0.407187)
2. feature 'block_height' (0.265663)
0. feature 'x_coords' (0.125087)
1. feature 'y_coords' (0.103793)
4. feature 'num_siblings' (0.098270)
```



In [14]:

```
forest = RandomForestClassifier(n_estimators=100)
perform_analysis_of_field('location', forest, data)
```

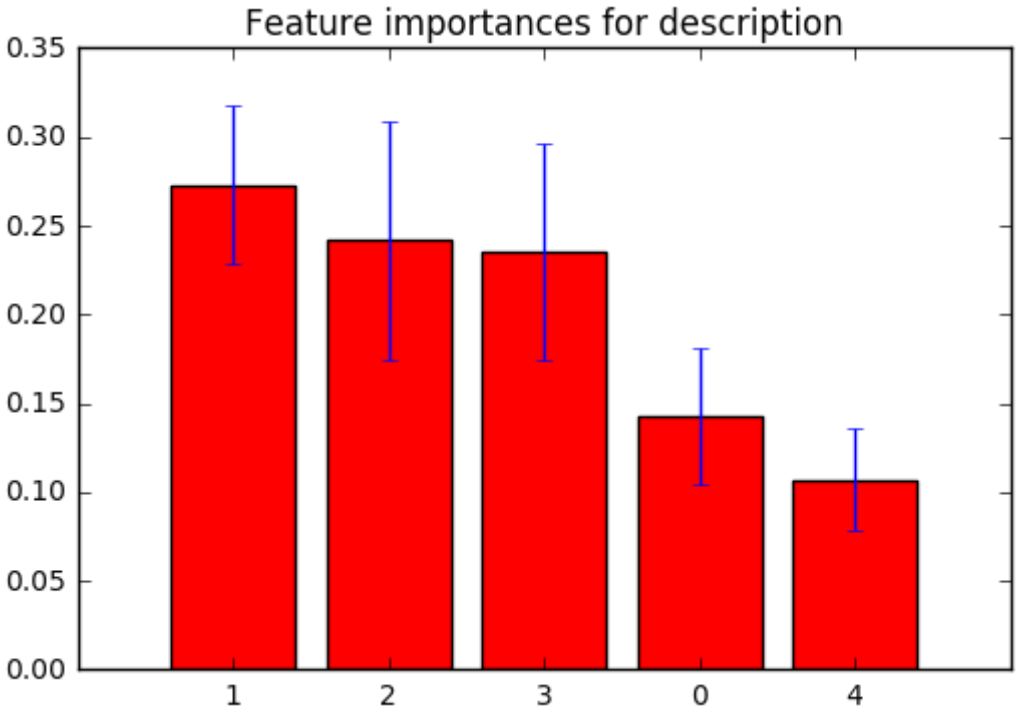
```
train: 0.9999167152494378, test: 0.987826184974356
Feature ranking:
1. feature 'y_coords' (0.226604)
2. feature 'block_height' (0.212874)
3. feature 'block_width' (0.200560)
4. feature 'num_siblings' (0.191961)
0. feature 'x_coords' (0.168001)
```



In [15]:

```
forest = RandomForestClassifier(n_estimators=100)
perform_analysis_of_field('description', forest, data)
```

```
train: 1.0, test: 0.9822730284956925
Feature ranking:
1. feature 'y_coords' (0.273154)
2. feature 'block_height' (0.241604)
3. feature 'block_width' (0.235432)
0. feature 'x_coords' (0.142900)
4. feature 'num_siblings' (0.106911)
```



In [ ]:

In [ ]:

In [ ]: