In [1]:

```
from utils_all import *
from sklearn.preprocessing import OneHotEncoder, LabelEncoder
from sklearn.manifold import TSNE
import seaborn as sns
%matplotlib inline
import inspect
from sklearn.decomposition import IncrementalPCA
from sklearn import manifold
```

```
/Users/jetbrains/miniconda3/envs/py35/lib/python3.5/site-packages/sk
learn/cross_validation.py:44: DeprecationWarning: This module was de
precated in version 0.18 in favor of the model_selection module into
which all the refactored classes and functions are moved. Also note
 that the interface of the new CV iterators are different from that
 of this module. This module will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)
```

In [2]:

```
from time import time
from matplotlib import offsetbox
from sklearn import (manifold, datasets, decomposition, ensemble,
                     discriminant_analysis, random_projection)
```

In [287]:

```
%store -r data
%store -r data_t
```

In [288]:

```
data_t
```

Out[288]:

	url	meta_name	text
0	http://www.therapeutenfinder.com/veranstaltung...	description	Ein Sonntag in jedem 2. Monat ab 18 Uhr\r\rFür...
1	http://www.therapeutenfinder.com/veranstaltung...	startDate	12.07.2015
2	http://www.therapeutenfinder.com/veranstaltung...	location	Ort:\rWürzstr. 1\r \r\r81371 \rMünchen
3	http://www.ticketbis.com/jp/morrisey-%E3%83%8B...	name	Morrissey (モリッシー) Dallas
4	http://www.ticketbis.com/jp/morrisey-%E3%83%8B...	startDate	2017. 4. 15 - 土 20:00
5	http://www.ticketbis.com/jp/morrisey-%E3%83%8B...	location	Majestic Theatre Dallas, Dallas, ...
7	http://www.volksfeste-in-deutschland.de/doelke...	name	Dölker Mühlenfest in Viersen OT Dülken 2017
8	http://www.volksfeste-in-deutschland.de/doelke...	description	Das Mühlenfest in Viersen OT Dülken hält hält ...
10	http://www.volksfeste-in-deutschland.de/doelke...	location	Alter Markt om Ortsteil Dülken\r\r41751\r Viersen
12	https://bxl.demosphere.eu/rv/3412\n	location	Lieu :BruxellesIHECSRue de l'Etuve 58-601000 B...
14	https://bxl.demosphere.eu/rv/3412\n	description	Ciné-Débat : Louise Michel - L'autogestion ici...
15	http://www.chicagoparkdistrict.com/parks/Galew...	name	Movie Inside the Park at Mayfair Park
16	http://www.chicagoparkdistrict.com/parks/Galew...	startDate	NaN
17	http://www.chicagoparkdistrict.com/parks/Galew...	location	Mayfair Park\r\r \r\r 4...
18	http://www.chicagoparkdistrict.com/parks/Galew...	description	We've moved the movies inside for our spring m...
19	http://www.chicagoparkdistrict.com/parks/Bosle...	name	45th Annual Special Olympics Spring Games & Op...

	url	meta_name	text
20	http://www.chicagoparkdistrict.com/parks/Bosle...	startDate	NaN
23	http://www.chicagoparkdistrict.com/parks/Bosle...	location	Eckersall Playground Park\r\r \r\r ...
24	http://www.elbe-wochenblatt.de/eissendorf/loka...	name	VAN WOLFEN – ERDIGER BLUESROCK ! SUPPORT : WIR...
25	http://www.elbe-wochenblatt.de/eissendorf/loka...	location	Marias Ballroom, Lassallestraße 11, 21073 Hamb...
26	http://www.elbe-wochenblatt.de/eissendorf/loka...	description	VAN WOLFEN :Und der hat im Laufe der Zeit reic...
27	http://www.elbe-wochenblatt.de/kirchdorf/lokal...	name	Puppen entern Heimatmuseum
29	http://www.elbe-wochenblatt.de/kirchdorf/lokal...	location	Museum der Elbinsel, Kirchdorfer Straße 163, 2...
30	http://www.elbe-wochenblatt.de/kirchdorf/lokal...	description	Wenn Erika Harenkamp mit ihren wunderschönen, ...
31	http://metroo.es/e/torrejon/luis-mariano-el-pr...	description	Género/Subgénero: Teatro/Musical. Compañía: FE...
32	http://metroo.es/e/torrejon/luis-mariano-el-pr...	startDate	NaN
34	http://metroo.es/e/torrejon/luis-mariano-el-pr...	location	Dirección : Torrejón de ArdozMadrid
37	https://www.finesettimana.it/Scoprire/Dettagli...	location	NaN
38	https://www.finesettimana.it/Scoprire/Dettagli...	description	VENDREDI 20 JUIN 2014 dès 23H30le Moth Club pr...
41	https://www.finesettimana.it/Scoprire/Dettagli...	location	NaN
...	...	...	...
545	http://www.southwestorlandobulletin.com/?optio...	location	Oakland Nature Preserve747 Machete Trail, Oakl...
546	http://www.southwestorlandobulletin.com/?optio...	description	Oakland Nature Preserve sponsors a

			two- to thr...

	url	meta_name	text
547	http://www.southwestorlandobulletin.com/?optio...	name	Kayak Tour
548	http://www.southwestorlandobulletin.com/?optio...	startDate	8:30 AM-3:00 PM
549	http://www.southwestorlandobulletin.com/?optio...	location	Oakland Nature Preserve747 Machete Trail, Oakl...
550	http://www.southwestorlandobulletin.com/?optio...	description	Oakland Nature Preserve sponsors a two- to thr...
551	http://www.reservix.de/tickets-cuba-insel-im-a...	name	Tickets - Karten Cuba - Insel im Aufbruch
552	http://www.reservix.de/tickets-cuba-insel-im-a...	startDate	So, 19.03.2017 11:00 Uhr
553	http://www.reservix.de/tickets-cuba-insel-im-a...	location	Linden-Museum StuttgartHegelplatz 1, 70174 Stu...
554	http://www.reservix.de/tickets-cuba-insel-im-a...	description	Das sozialistische Kuba befindet sich im Wande...
1	http://www.shreveportla.gov/Calendar.aspx?EID=...	location	Location:\r\r\rView Facility\r\r\rSouthern Hil...
2	http://www.shreveportla.gov/Calendar.aspx?EID=...	name	318-673-7818
3	http://www.shreveportla.gov/Calendar.aspx?EID=...	name	Southern Hills Community Center Park Advisory ...
4	http://www.shreveportla.gov/Calendar.aspx?EID=...	description	NaN
5	http://www.southwestorlandobulletin.com/?optio...	name	Kayak Tour
6	http://www.southwestorlandobulletin.com/?optio...	startDate	8:30 AM-3:00 PM
7	http://www.southwestorlandobulletin.com/?optio...	location	Oakland Nature Preserve747 Machete Trail, Oakl...
8	http://www.southwestorlandobulletin.com/?optio...	description	Oakland Nature Preserve sponsors a two- to thr...
9	http://www.southwestorlandobulletin.com/?optio...	name	Kayak Tour

	url	meta_name	text
10	http://www.southwestorlandobulletin.com/?optio...	startDate	8:30 AM-3:00 PM
11	http://www.southwestorlandobulletin.com/?optio...	location	Oakland Nature Preserve747 Machete Trail, Oakl...
12	http://www.southwestorlandobulletin.com/?optio...	description	Oakland Nature Preserve sponsors a two- to thr...
13	http://www.southwestorlandobulletin.com/?optio...	name	Kayak Tour
14	http://www.southwestorlandobulletin.com/?optio...	startDate	8:30 AM-3:00 PM
15	http://www.southwestorlandobulletin.com/?optio...	location	Oakland Nature Preserve747 Machete Trail, Oakl...
16	http://www.southwestorlandobulletin.com/?optio...	description	Oakland Nature Preserve sponsors a two- to thr...
17	http://www.therapeutenfinder.com/veranstaltung...	name	Neu! Psycho-Holistischer Stammtisch
18	http://www.therapeutenfinder.com/veranstaltung...	description	Ein Sonntag in jedem 2. Monat ab 18 Uhr\r\rFür...
19	http://www.therapeutenfinder.com/veranstaltung...	startDate	12.07.2015
20	http://www.therapeutenfinder.com/veranstaltung...	location	Ort:\rWürzstr. 1\r \r\r81371 \rMünchen

6140 rows × 304 columns

In [243]:

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

In [284]:

```
def prepare_num_XY(data_cl, num_features):
    data_cl = data_cl.dropna(axis=0)
    data_num = data_cl[num_features]
    data_num = data_num[data_num.applymap(lambda x: isinstance(x, (int,
float)))]
    data_cl[num_features] = data_num
    data_cl = data_cl.dropna(axis=0)
    y = data_cl['meta_name']
    X = data_cl.drop('meta_name', 1)
    le = LabelEncoder()
    le_tag = le.fit_transform(X.tag.values)
    X.loc[:,'tag'] = le_tag
    return X, y, le_tag
```

In [285]:

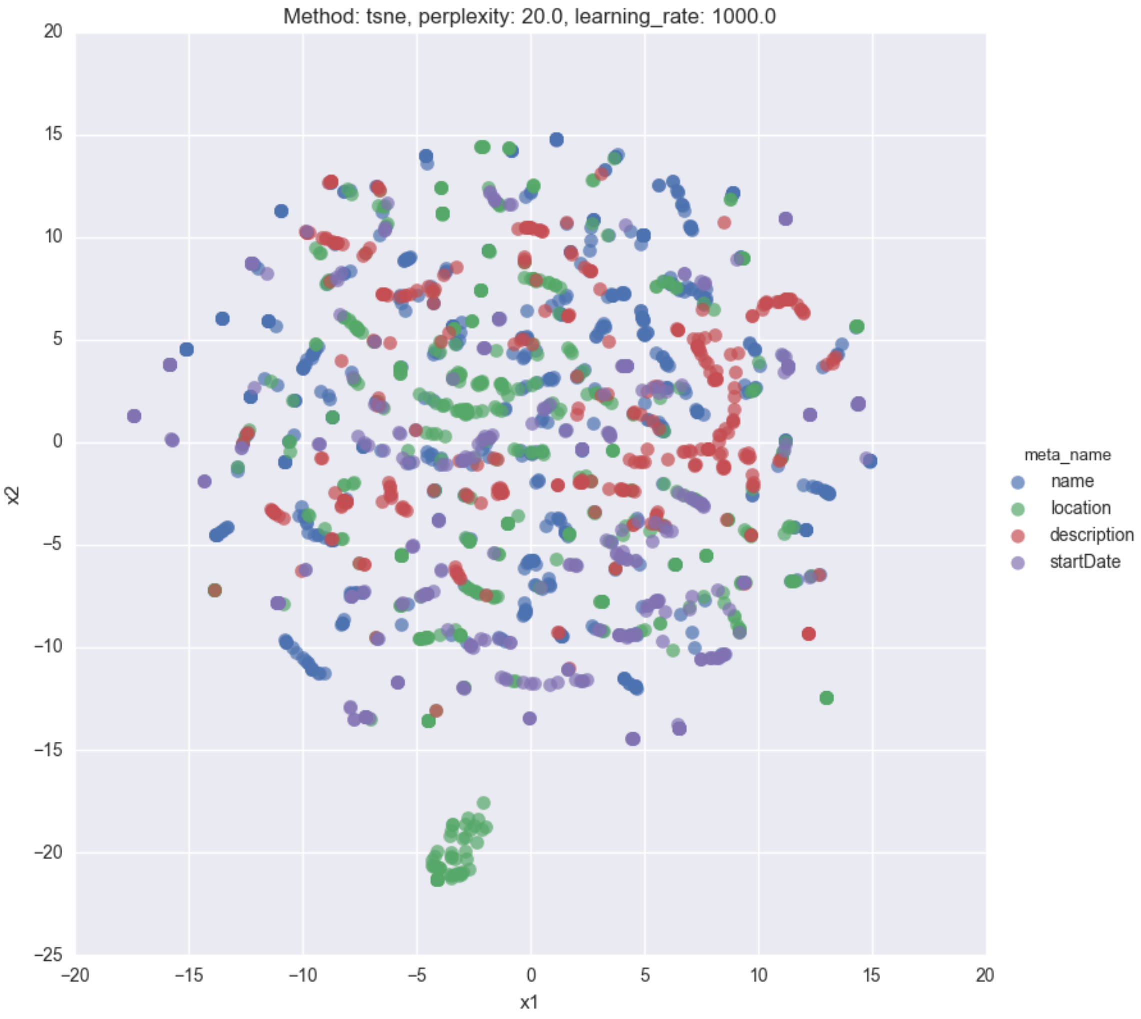
```
data_cl = data_cl[['meta_name','x_coords','y_coords','block_height','block_widt
h','num_siblings', 'tag']]
num_features = ['x_coords','y_coords','block_height','block_width','num_sibling
s']
X, y, le_tag = prepare_num_XY(data_cl, num_features)
```

In [368]:

```
def run_dim_reduction_and_draw(data, labels, model, title):
    X_model = model.fit_transform(data)
    tsne = pd.DataFrame({'x1': X_model[:, 0], 'x2': X_model[:, 1], 'meta_name':
labels, })
    tsne_lim = tsne[tsne['meta_name'].isin(['location', 'name', 'description',
'startDate'])]
    tsne_smpl = tsne_lim.sample(n=4000)
    g = sns.lmplot('x1', 'x2', tsne_smpl, hue='meta_name', fit_reg=False,
size=8,scatter_kws={'alpha':0.7,'s':60})
    g.axes.flat[0].set_title(title)
    return X_model
```

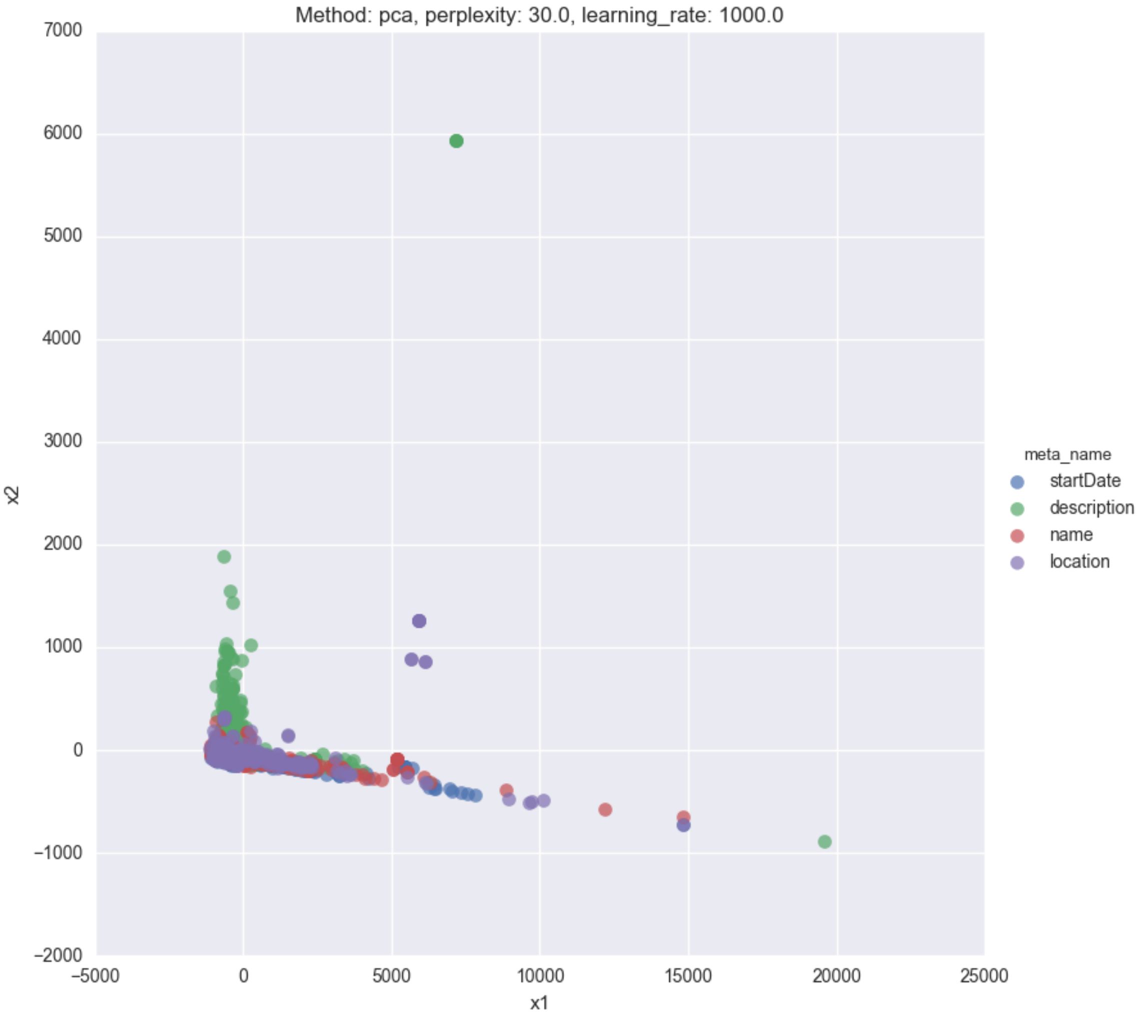
In [272]:

```
perplexity=20
learning_rate=1000.0
model = TSNE(n_components=2,
             random_state=0,
             perplexity=perplexity,
             learning_rate=learning_rate)
title = 'Method: {}, perplexity: {}, learning_rate: {}'.format(method, perplexit
y, learning_rate)
run_dim_reduction_and_draw(X, y, model, title)
```



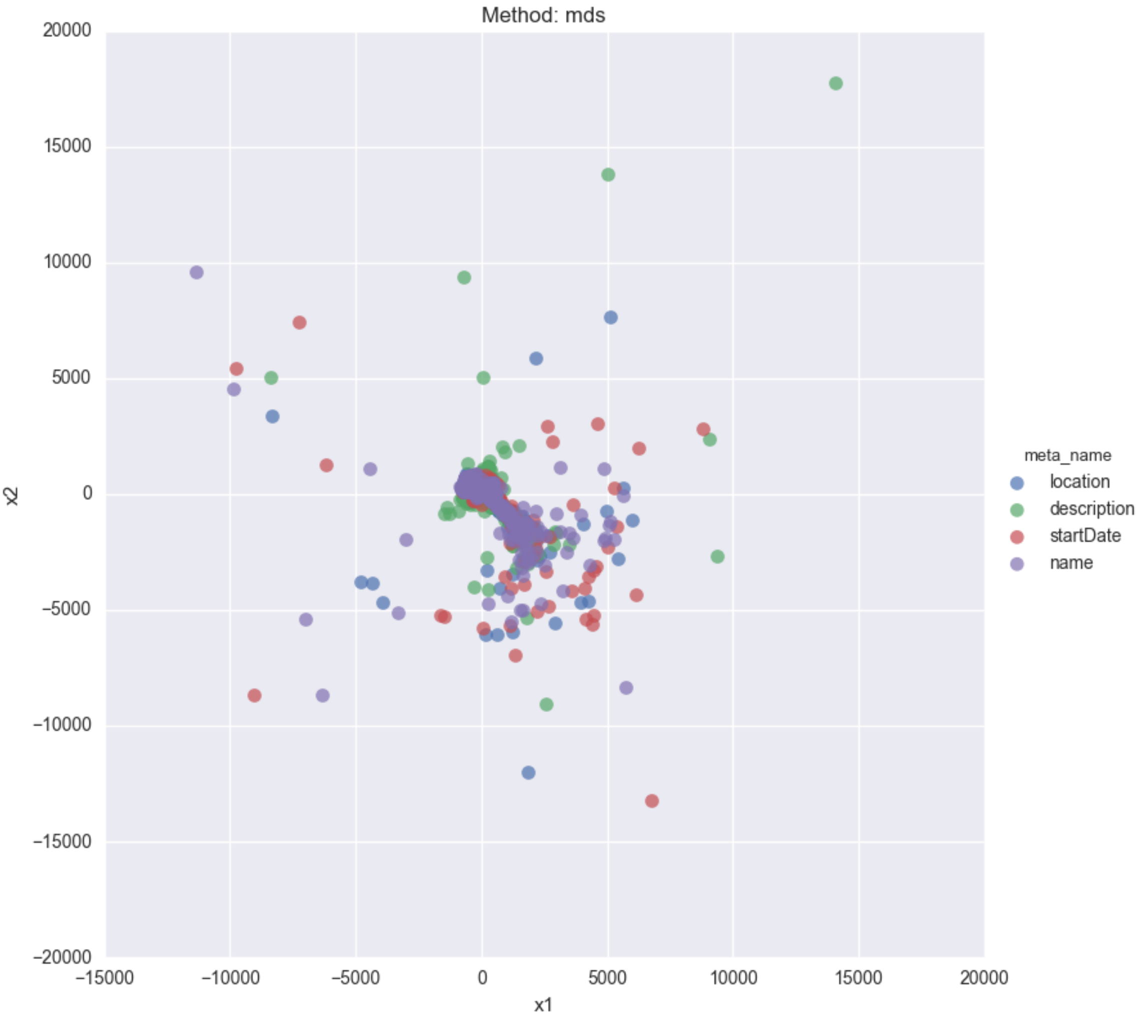
In [277]:

```
model = IncrementalPCA(n_components=2, batch_size=3)
title = 'Method: PCA'
run_dim_reduction_and_draw(X, y, model, title)
```



In [281]:

```
model = manifold.MDS(n_components=2, max_iter=100, n_init=1)
title = 'Method: MDS'
run_dim_reduction_and_draw(data=X, labels=y, model, title)
```



Dimensionality reduction did not show clear clusters, that means that either the fieatures are not well saparable or these methods don't shoe the structure and we need more complicated methods and classifiers for the data. Good exmplanations are here:

http://distill.pub/2016/misread-tsne/ (http://distill.pub/2016/misread-tsne/) https://www.quora.com/Is-it-indistinguishable-if-t-SNE-method-does-not-show-clear-two-clusters-for-2-class-classification-problem-1 (https://www.quora.com/Is-it-indistinguishable-if-t-SNE-method-does-not-show-clear-two-clusters-for-2-class-classification-problem-1)

In [107]:

```
def get_XY_tsne(field_name, data):
    df_pos = data[data['meta_name'] == field_name]
    df_neg = data[data['meta_name'] != field_name]

    no_field_smpl = df_neg.sample(n = df_pos.shape[0], random_state=0)
    real_meta = no_field_smpl.meta_name
    no_field_smpl = no_field_smpl.drop('meta_name', 1)
    no_field_smpl.insert(1, 'meta_name', 'no_'+ field_name)
    X_field = pd.concat((df_pos, no_field_smpl), axis=0)
    y = X_field['meta_name']
    X_field = X_field.drop('meta_name', 1)
    X_field = X_field.replace('none', 'NA')

    X = X_field

    X['y'] = y
    X = X.convert_objects(convert_numeric=True).dropna()
    y = X['y']
    X = X.drop('y', 1)
    return X, y, real_meta
```

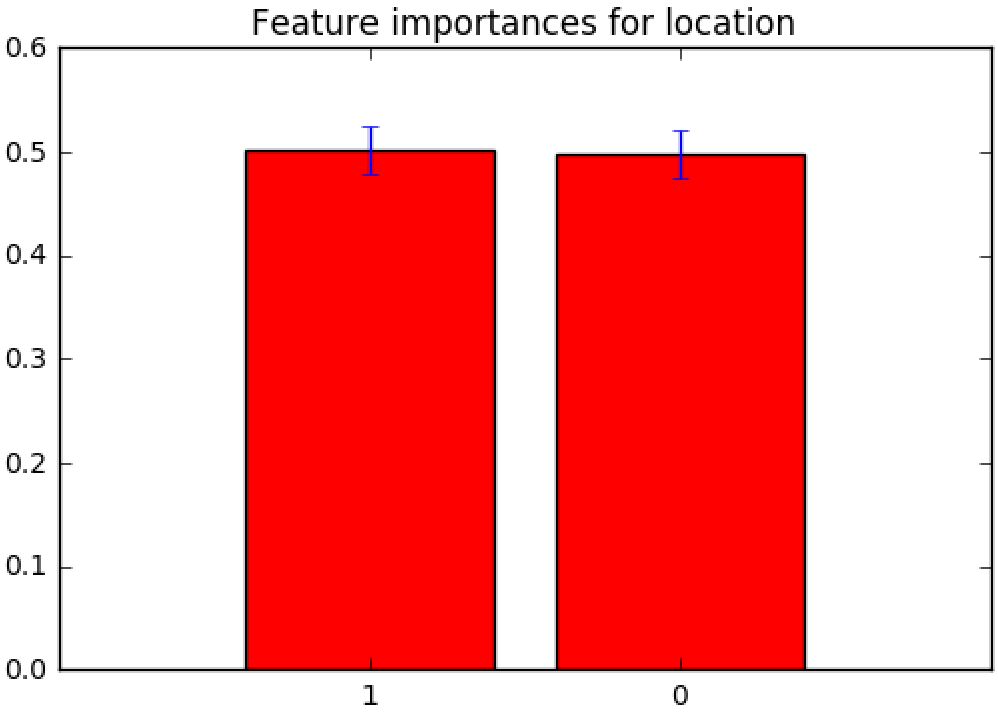
In [119]:

```
def perform_analysis_tsne(X, y, clf, field_name):
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=7)
    clf.fit(X_train, y_train)
    print('{} prediction'.format(field_name))
    print("train: {}, test: {}".format(clf.score(X_train, y_train), clf.score(X_test, y_test)))
    draw_feature_importance(clf, X_train, field_name)
```

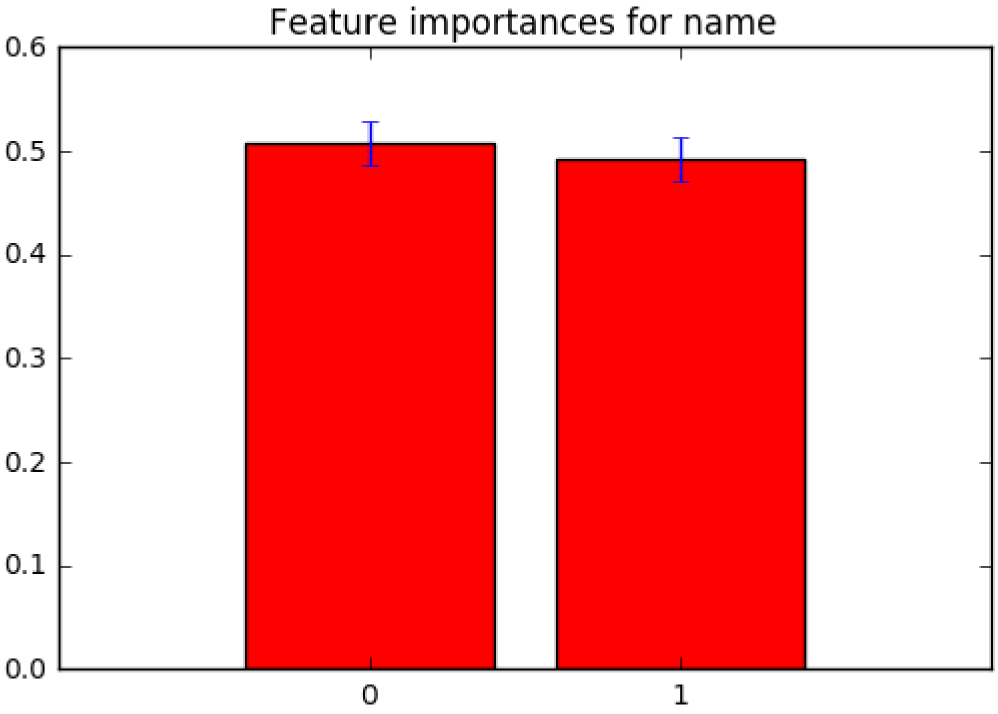
In [120]:

```
# fields = ['location', 'name', 'description', 'startDate']
# for field_name in fields:
#     forest = RandomForestClassifier(n_estimators=100)
#     X, y, real_meta = get_XY_tsne(field_name, tsne)
#     perform_analysis_tsne(X, y, forest, field_name)
```

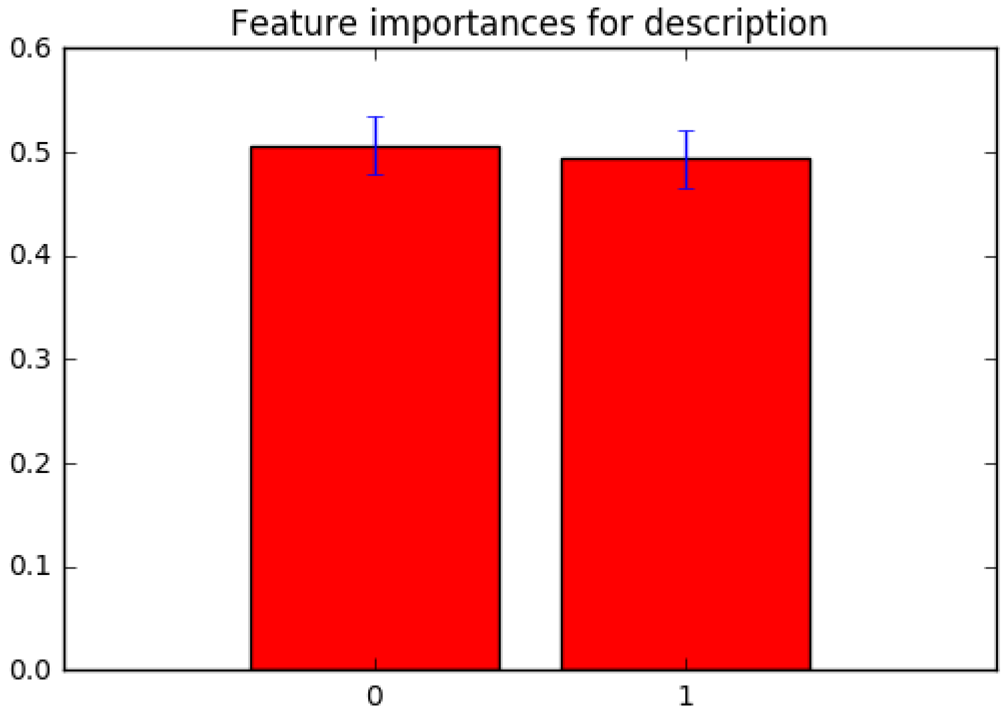
```
location prediction
train: 0.9847586790855207, test: 0.49828178694158076
Feature ranking:
1. feature 'x2' (0.502560)
0. feature 'x1' (0.497440)
```



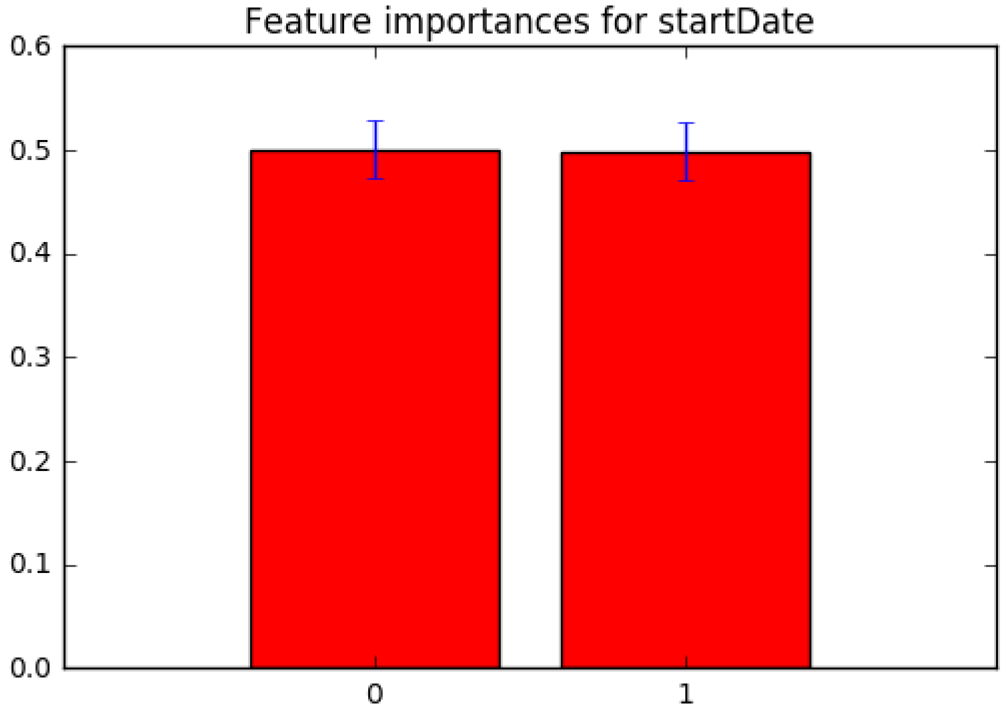
```
name prediction
train: 0.9887231720625682, test: 0.4966789667896679
Feature ranking:
0. feature 'x1' (0.507443)
1. feature 'x2' (0.492557)
```



```
description prediction
train: 0.9927849927849928, test: 0.4926900584795322
Feature ranking:
0. feature 'x1' (0.506556)
1. feature 'x2' (0.493444)
```



```
startDate prediction
train: 0.9915458937198067, test: 0.5343137254901961
Feature ranking:
0. feature 'x1' (0.501098)
1. feature 'x2' (0.498902)
```



Dim reduction for X with additional text features

Let's do the same dim reduction but with additional text features from notebook "Analysis 3". We will see if the result differs from the first try.

In [289]:

```
data_cl = clean_df(data_t)
```

In [290]:

```
data_cl
```

Out[290]:

	url	meta_name	text
0	http://www.therapeutenfinder.com/veranstaltung...	description	Ein Sonntag in jedem 2. Monat ab 18 Uhr\r\rFür...
1	http://www.therapeutenfinder.com/veranstaltung...	startDate	12.07.2015
2	http://www.therapeutenfinder.com/veranstaltung...	location	Ort:\rWürzstr. 1\r \r\r81371 \rMünchen
3	http://www.ticketbis.com/jp/morrisey-%E3%83%8B...	name	Morrissey (モリッシー) Dallas
4	http://www.ticketbis.com/jp/morrisey-%E3%83%8B...	startDate	2017. 4. 15 - 土 20:00
5	http://www.ticketbis.com/jp/morrisey-%E3%83%8B...	location	Majestic Theatre Dallas, Dallas, ...
7	http://www.volksfeste-in-deutschland.de/doelke...	name	Dölker Mühlenfest in Viersen OT Dülken 2017
8	http://www.volksfeste-in-deutschland.de/doelke...	description	Das Mühlenfest in Viersen OT Dülken hält hält ...
10	http://www.volksfeste-in-deutschland.de/doelke...	location	Alter Markt om Ortsteil Dülken\r\r41751\r Viersen
12	https://bxl.demosphere.eu/rv/3412\n	location	Lieu :BruxellesIHECSRue de l'Etuve 58-601000 B...
14	https://bxl.demosphere.eu/rv/3412\n	description	Ciné-Débat : Louise Michel - L'autogestion ici...
15	http://www.chicagoparkdistrict.com/parks/Galew...	name	Movie Inside the Park at Mayfair Park
16	http://www.chicagoparkdistrict.com/parks/Galew...	startDate	NaN
17	http://www.chicagoparkdistrict.com/parks/Galew...	location	Mayfair Park\r\r \r\r 4...
18	http://www.chicagoparkdistrict.com/parks/Galew...	description	We've moved the movies inside for our spring m...
19	http://www.chicagoparkdistrict.com/parks/Bosle...	name	45th Annual Special Olympics Spring Games & Op...

	url	meta_name	text
20	http://www.chicagoparkdistrict.com/parks/Bosle...	startDate	NaN
23	http://www.chicagoparkdistrict.com/parks/Bosle...	location	Eckersall Playground Park\r\r \r\r ...
24	http://www.elbe-wochenblatt.de/eissendorf/loka...	name	VAN WOLFEN – ERDIGER BLUESROCK ! SUPPORT : WIR...
25	http://www.elbe-wochenblatt.de/eissendorf/loka...	location	Marias Ballroom, Lassallestraße 11, 21073 Hamb...
26	http://www.elbe-wochenblatt.de/eissendorf/loka...	description	VAN WOLFEN :Und der hat im Laufe der Zeit reic...
27	http://www.elbe-wochenblatt.de/kirchdorf/lokal...	name	Puppen entern Heimatmuseum
29	http://www.elbe-wochenblatt.de/kirchdorf/lokal...	location	Museum der Elbinsel, Kirchdorfer Straße 163, 2...
30	http://www.elbe-wochenblatt.de/kirchdorf/lokal...	description	Wenn Erika Harenkamp mit ihren wunderschönen, ...
31	http://metroo.es/e/torrejon/luis-mariano-el-pr...	description	Género/Subgénero: Teatro/Musical. Compañía: FE...
32	http://metroo.es/e/torrejon/luis-mariano-el-pr...	startDate	NaN
34	http://metroo.es/e/torrejon/luis-mariano-el-pr...	location	Dirección : Torrejón de ArdozMadrid
37	https://www.finesettimana.it/Scoprire/Dettagli...	location	NaN
38	https://www.finesettimana.it/Scoprire/Dettagli...	description	VENDREDI 20 JUIN 2014 dès 23H30le Moth Club pr...
41	https://www.finesettimana.it/Scoprire/Dettagli...	location	NaN
...	...	...	...
545	http://www.southwestorlandobulletin.com/?optio...	location	Oakland Nature Preserve747 Machete Trail, Oakl...
546	http://www.southwestorlandobulletin.com/?optio...	description	Oakland Nature Preserve sponsors a

			two- to thr...

	url	meta_name	text
547	http://www.southwestorlandobulletin.com/?optio...	name	Kayak Tour
548	http://www.southwestorlandobulletin.com/?optio...	startDate	8:30 AM-3:00 PM
549	http://www.southwestorlandobulletin.com/?optio...	location	Oakland Nature Preserve747 Machete Trail, Oakl...
550	http://www.southwestorlandobulletin.com/?optio...	description	Oakland Nature Preserve sponsors a two- to thr...
551	http://www.reservix.de/tickets-cuba-insel-im-a...	name	Tickets - Karten Cuba - Insel im Aufbruch
552	http://www.reservix.de/tickets-cuba-insel-im-a...	startDate	So, 19.03.2017 11:00 Uhr
553	http://www.reservix.de/tickets-cuba-insel-im-a...	location	Linden-Museum StuttgartHegelplatz 1, 70174 Stu...
554	http://www.reservix.de/tickets-cuba-insel-im-a...	description	Das sozialistische Kuba befindet sich im Wande...
1	http://www.shreveportla.gov/Calendar.aspx?EID=...	location	Location:\r\r\rView Facility\r\r\rSouthern Hil...
2	http://www.shreveportla.gov/Calendar.aspx?EID=...	name	318-673-7818
3	http://www.shreveportla.gov/Calendar.aspx?EID=...	name	Southern Hills Community Center Park Advisory ...
4	http://www.shreveportla.gov/Calendar.aspx?EID=...	description	NaN
5	http://www.southwestorlandobulletin.com/?optio...	name	Kayak Tour
6	http://www.southwestorlandobulletin.com/?optio...	startDate	8:30 AM-3:00 PM
7	http://www.southwestorlandobulletin.com/?optio...	location	Oakland Nature Preserve747 Machete Trail, Oakl...
8	http://www.southwestorlandobulletin.com/?optio...	description	Oakland Nature Preserve sponsors a two- to thr...
9	http://www.southwestorlandobulletin.com/?optio...	name	Kayak Tour

	url	meta_name	text
10	http://www.southwestorlandobulletin.com/?optio...	startDate	8:30 AM-3:00 PM
11	http://www.southwestorlandobulletin.com/?optio...	location	Oakland Nature Preserve747 Machete Trail, Oakl...
12	http://www.southwestorlandobulletin.com/?optio...	description	Oakland Nature Preserve sponsors a two- to thr...
13	http://www.southwestorlandobulletin.com/?optio...	name	Kayak Tour
14	http://www.southwestorlandobulletin.com/?optio...	startDate	8:30 AM-3:00 PM
15	http://www.southwestorlandobulletin.com/?optio...	location	Oakland Nature Preserve747 Machete Trail, Oakl...
16	http://www.southwestorlandobulletin.com/?optio...	description	Oakland Nature Preserve sponsors a two- to thr...
17	http://www.therapeutenfinder.com/veranstaltung...	name	Neu! Psycho-Holistischer Stammtisch
18	http://www.therapeutenfinder.com/veranstaltung...	description	Ein Sonntag in jedem 2. Monat ab 18 Uhr\r\rFür...
19	http://www.therapeutenfinder.com/veranstaltung...	startDate	12.07.2015
20	http://www.therapeutenfinder.com/veranstaltung...	location	Ort:\rWürzstr. 1\r \r\r81371 \rMünchen

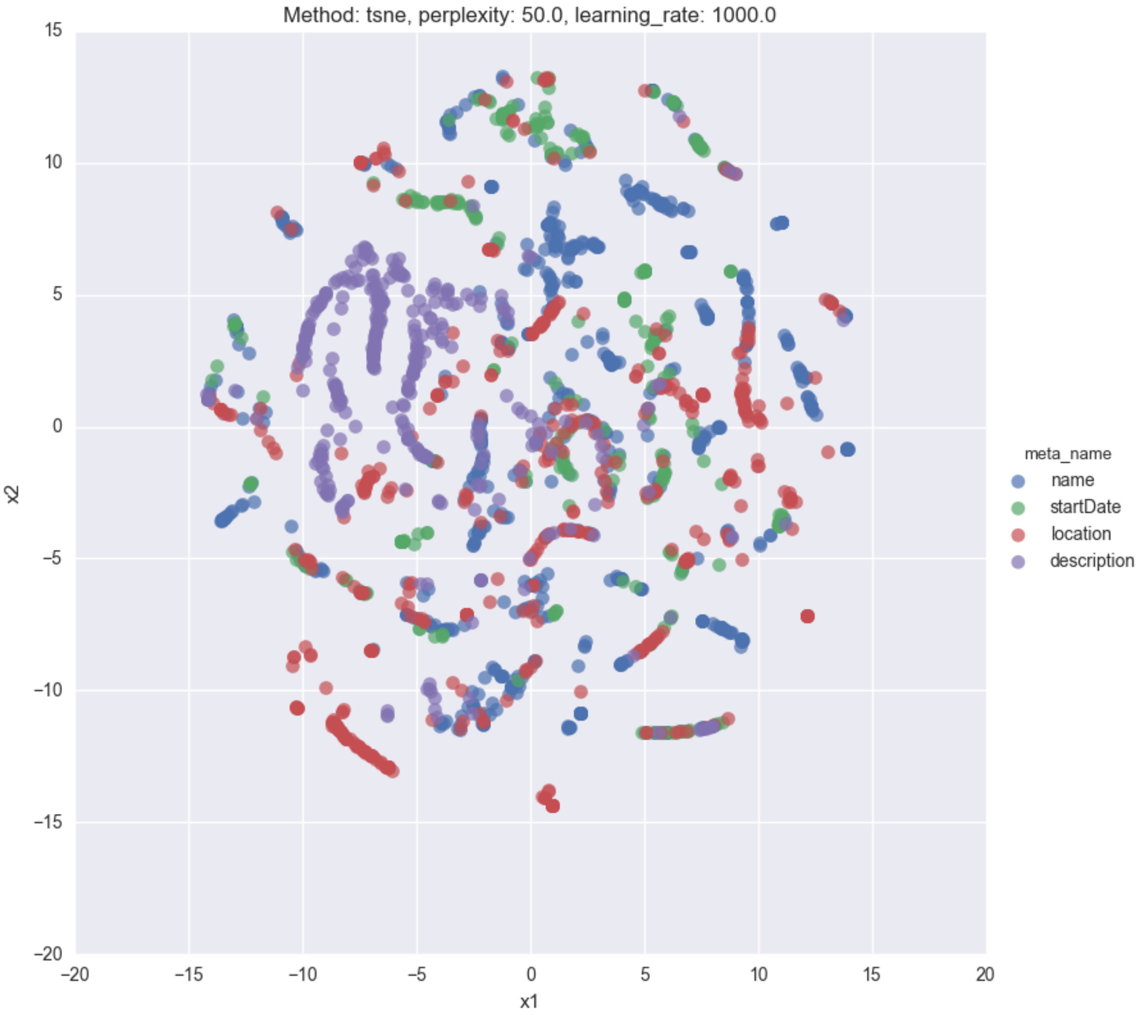
6140 rows × 304 columns

In [291]:

```
data_cl = data_cl[['meta_name','x_coords','y_coords','block_height','block_width',
                   'num_siblings', 'tag', 'text_len', 'num_punctuation','num_digits',
                   'digits_share','num_upper','num_space']]
num_features = ['x_coords','y_coords','block_height','block_width',
                   'num_siblings', 'text_len', 'num_punctuation','num_digits',
                   'digits_share','num_upper','num_space']
X, y, le_tag = prepare_num_XY(data_cl, num_features)
```

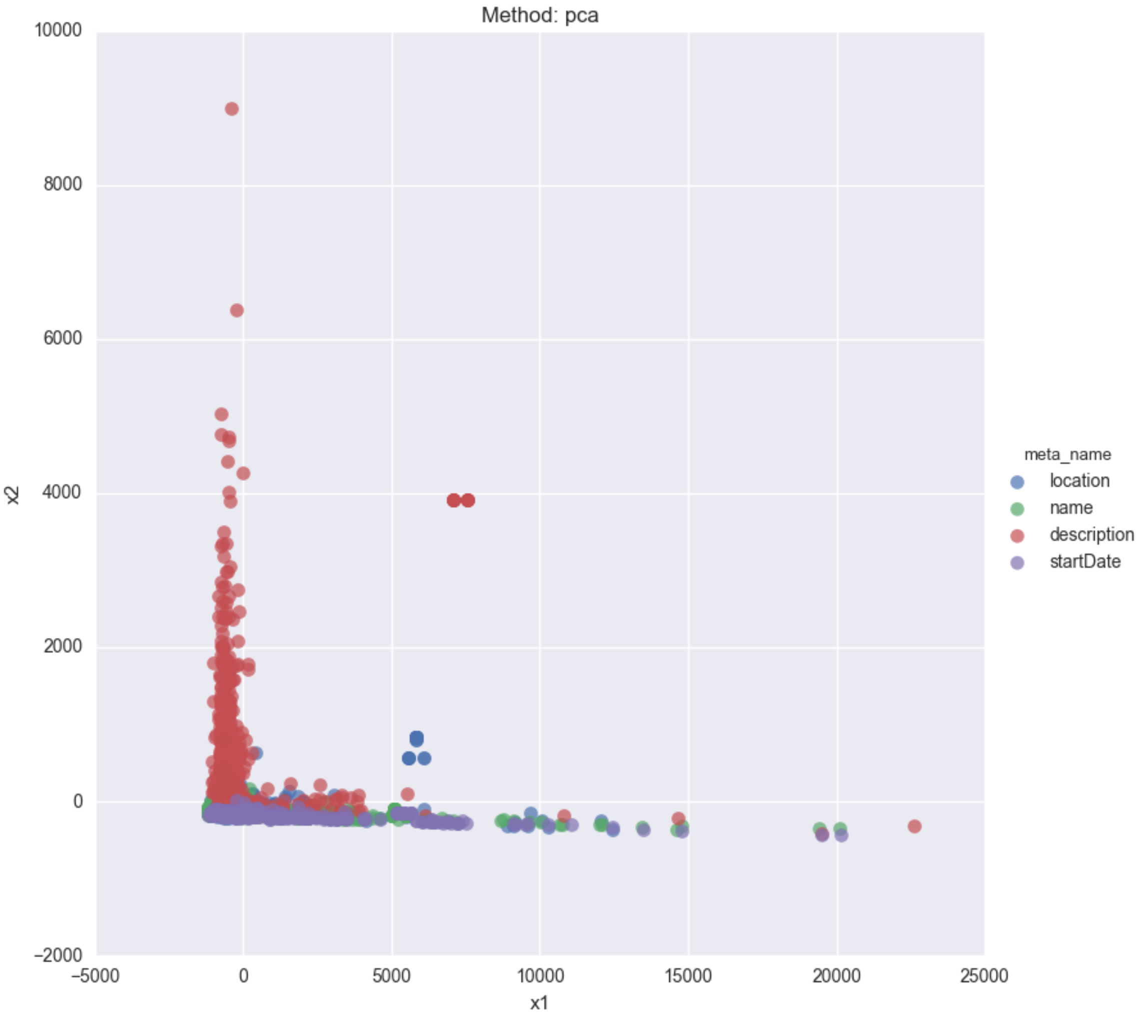
In [295]:

```
run_dim_reduction_and_draw(X, y, 50.0, 1000.0)
```



In [297]:

```
run_dim_reduction_and_draw(X, y, method='pca')
```

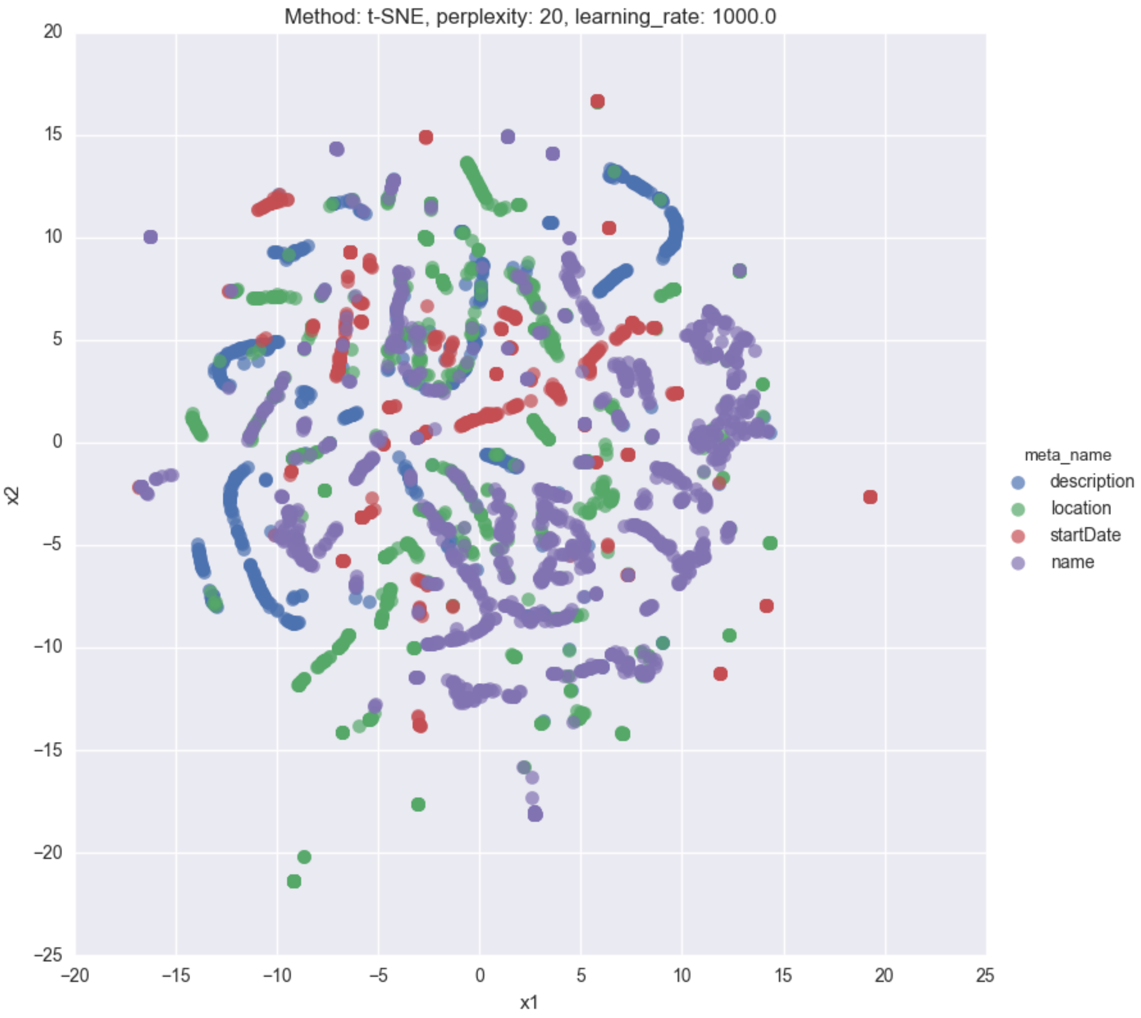


We see that description is already looks more or less like a cluster. Probably the important field was a text length.

t-sne only for text-related fields

```
In [305]:
```

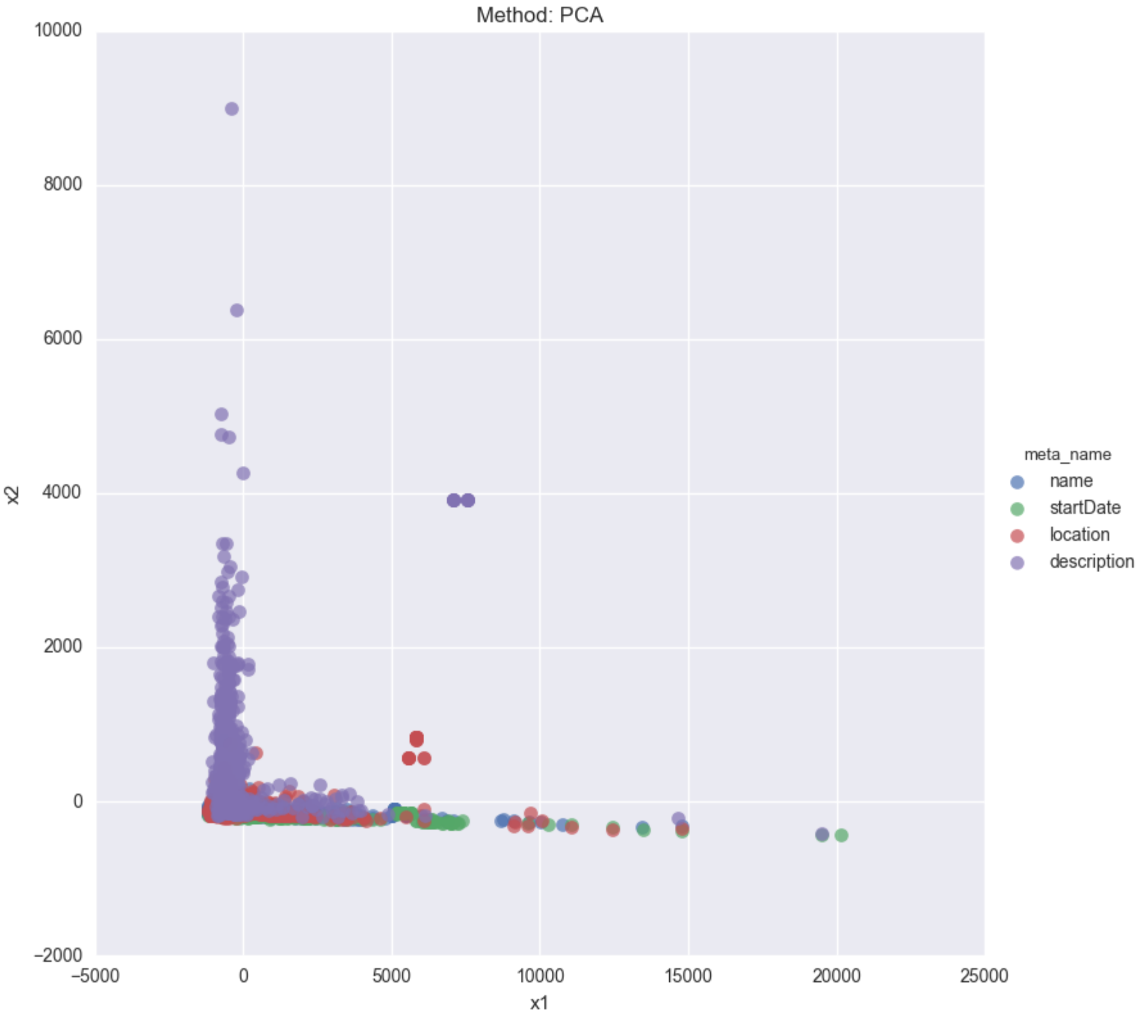
```
perplexity=20
learning_rate=1000.0

model = TSNE(n_components=2,
             random_state=0,
             perplexity=perplexity,
             learning_rate=learning_rate, n_iter=200)
title = 'Method: t-SNE, perplexity: {}, learning_rate: {}'.format(perplexity, le
arning_rate)
run_dim_reduction_and_draw(X.iloc[:,5:], y, model, title)
```



PCA

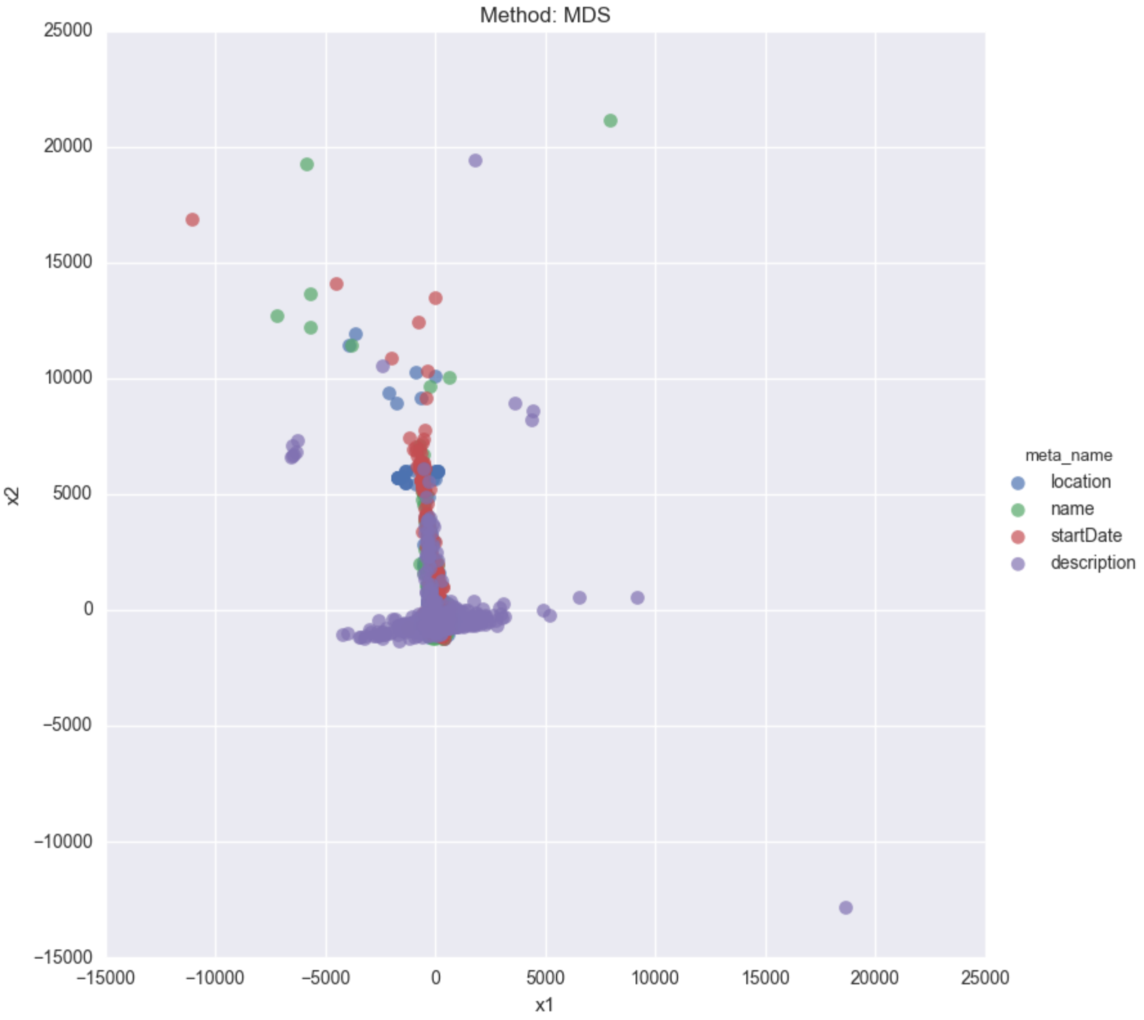
In [306]:

```
model = IncrementalPCA(n_components=2, batch_size=3)
title = 'Method: PCA'
run_dim_reduction_and_draw(X, y, model, title)
```



In [308]:

```
model = manifold.MDS(n_components=2, max_iter=300, n_init=1)
title = 'Method: MDS'
run_dim_reduction_and_draw(X, y, model, title)
```



In [344]:

```
%store -r X_t
```

Dimensionality reduction for TF-IDF feature (sparse matrix)

In [345]:

```
X_t_df = pd.DataFrame(X_t.toarray(), columns=range(28738))
```

In [346]:

```
X_t_df.shape
```

Out[346]:

```
(6140, 28738)
```

In [347]:

```
data_t.shape
```

Out[347]:

```
(6140, 304)
```

In [355]:

```
data_t.index = range(6140)
X_t_df.index = range(6140)
```

In [356]:

```
X_text = pd.concat([X_t_df, data_t], axis=1)
```

In [359]:

```
X_text_cl = clean_df(X_text)
```

In [364]:

```
X_t = X_text_cl.iloc[:,:28738]
```

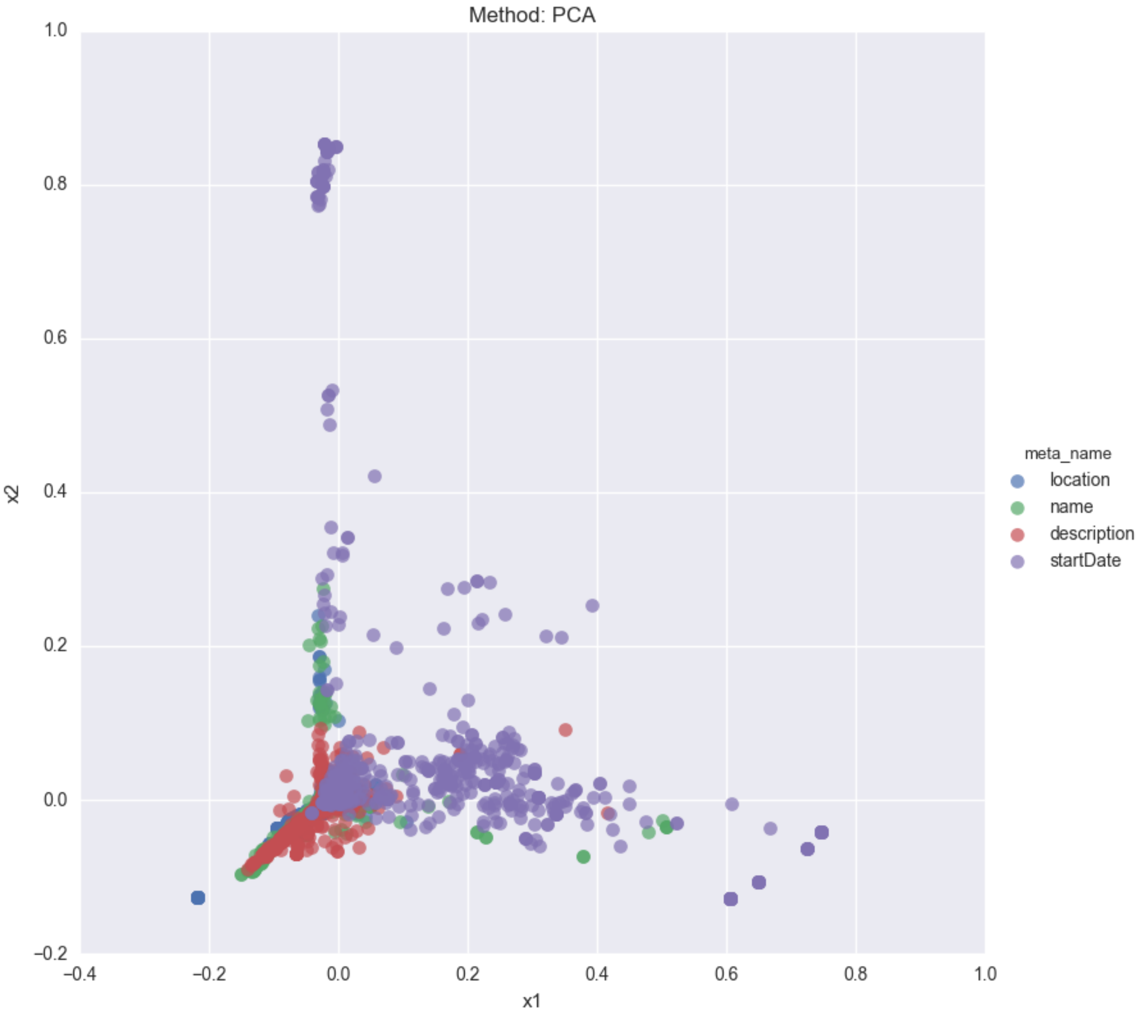
In [365]:

```
y_t = X_text['meta_name']
```

PCA for tf-idf

In [366]:

```
model = IncrementalPCA(n_components=2, batch_size=3)
title = 'Method: PCA'
run_dim_reduction_and_draw(X_t, y_t, model, title)
```



t-SNE for tf-idf