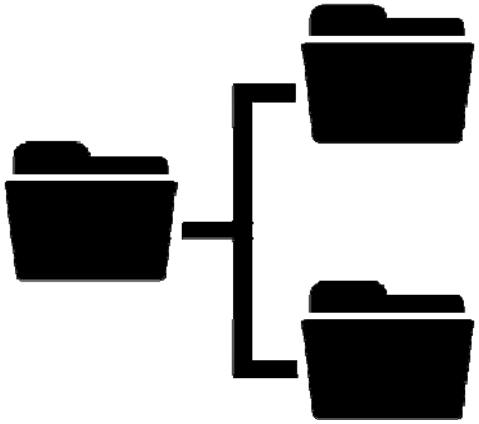
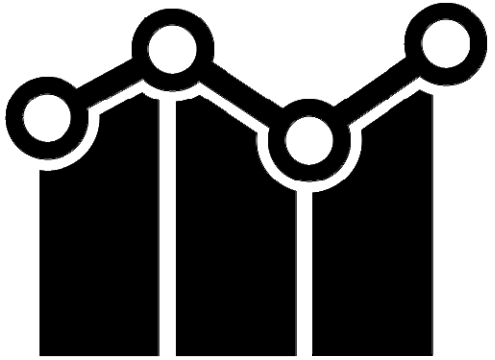
Intern Presentation

Ken Oung Yong Quan



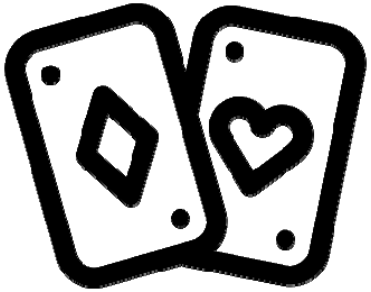
Website Classification

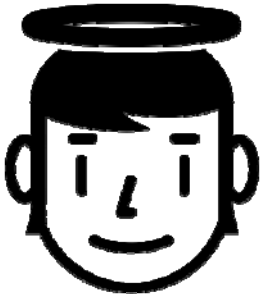


Trending Threads

Website Classification

To develop a classification system to distinguish online gambling websites from non-gambling websites





Website Classification

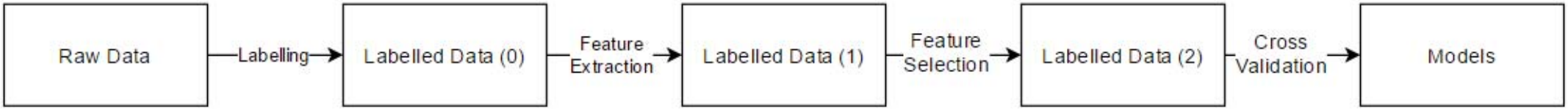
Remote Gambling Act 2014 (RGA)

"Websites which provide unauthorised remote gambling services, that is or may be used by individuals present in Singapore to gamble or contain remote gambling service advertisement or promotion accessible in Singapore will be blocked."

- Ministry of Home Affairs (2015)

Website Classification

Steps



Libraries Used

Scrapy

BeautifulSoup

NLTK

jieba

Scraping for data

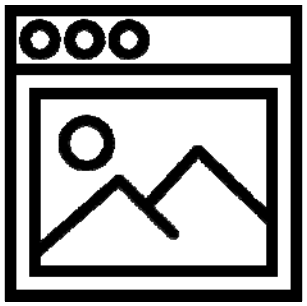
Getting Links

First 150 results on Google for:

"博彩", "赌注", "彩票", "竞猜", "赌博平台", "彩票", "开奖", "炸金花", "赌马", "六合彩", "在线赌博"

```
"https://www.google.com.sg/search?q={}&start={}".format(query, start)
```

Scraping for data

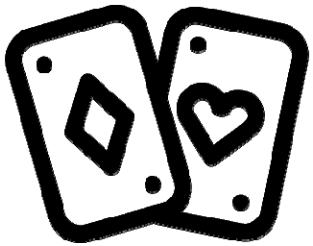


Page Content

Fails to capt

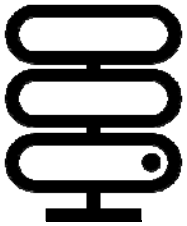
```
def better_get_page_text(response):
    if not response:
        raise ValueError("Error: No response“)
    else:
        driver = webdriver.PhantomJS()
        driver.get(response.url)
        content = driver.page_source
        page_text= [content]

        if 'iframe' in content:
            iframe = driver.find_elements_by_tag_name('iframe')
            if iframe_list:
                driver.switch_to_frame(iframe_list[0])

                # Add new content
                content2 = driver.page_source
                page_text.append(content2)

        driver.close()

        return u"".join(page_text).encode('utf-8').strip()
```

Labelling Data



Group 2

Online gambling sites



Group 1

Aggregators that link to a host of other gambling sites (Online gambling not supported on this site)

Group 0

Remaining sites that do not belong to 1 or 2

Labelling Data

URL	Content	Class
http://www.88879.com	澳门新濠天地上网导航设为...	2
http://9699.com	欧洲娱乐场官网欧洲娱乐...	1
http://da55555.com	大家旺娱乐城会员资讯站...	2

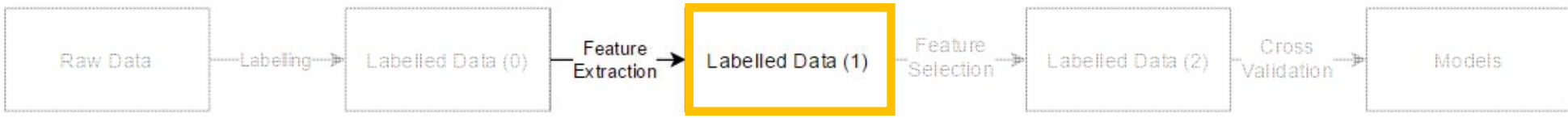
Feature Extraction

Text Segmentation

- Word and Char ngrams
- *Jieba* word segmentation

Convert to vector

- CountVectorizer
- **TfidfVectorizer**



Feature Extraction

Term Frequency-
Inverse Document Frequency

$$w_{x,y} = tf_{x,y} \times \ln \frac{N}{df_x}$$

For a term **x** in a document **y**

Feature Selection

Chi-squared

```
select = SelectPercentile(score_func=chi2,
percentile=28)
```

Feature Selection

Labelled Data (2)

URL	澳门	新濠	…	天地	上网	Class
http://www.88879.com	0.006155	0.009172	…	0.002818	0.088933	2
http://9699.com	0	0.135856	…	0	0.006978	1
http://da55555.com	0.003228	0	…	0.163493	0	2

Model Tuning

Split into training and test set

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=12)
```

Train using GridSearchCV

```
grid = GridSearchCV(model, cv=5, param_grid=param_grid, scoring="f1_macro", n_jobs=-1,
verbose=4)

grid.fit(X_train, y_train)
```

Check against test set

```
y_true, y_pred = y_test, clf.predict(X_test)
print metrics.classification_report(y_true, y_pred)
```

Model Selection (Chinese)

Model	Cross Validated Score	
	Char Ngram	Jieba
Logistic Regression	0.82 (+/- 0.17)	0.82 (+/- 0.18)
MultinomialNB	0.77 (+/- 0.17)	0.77 (+/- 0.12)
RandomForestClassifier	0.78 (+/- 0.16)	0.78 (+/- 0.16)
KNeighborsClassifier	0.79 (+/- 0.15)	0.79 (+/- 0.14)
SVC	0.81 (+/- 0.19)	0.83 (+/- 0.16)

```
cross_validation.cross_val_score(model_clf, X, y, cv = 5, scoring = 'f1_macro')
```

Improving Classification Accuracy

- More representative data set
  - Depends on how the crawler is designed
- New Features
  - Position in page (above/below the fold)
  - HTML tags (e.g. presence of form)
  - URL (.gov/.edu vs .com; numeric)
  - Image
  - Page authority

Other Cool Stuff I Tried

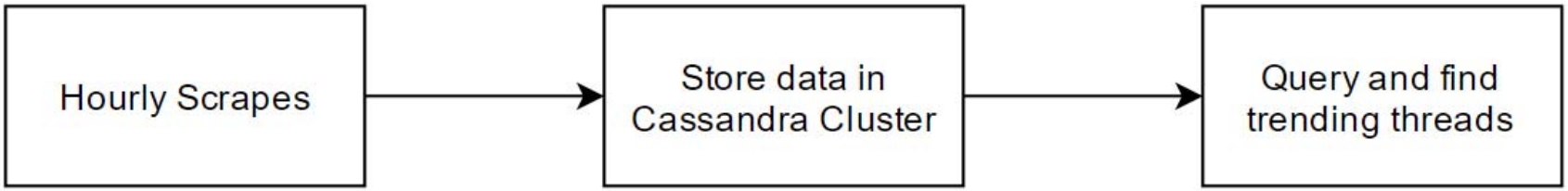
- pyLDAvis for topic modelling
- LIME for model interpretability
- Outsourcing with MicroWorkers

HWZ Trending Threads

To find currently trending threads on HWZ

HWZ Trending Threads

Steps



Libraries Used

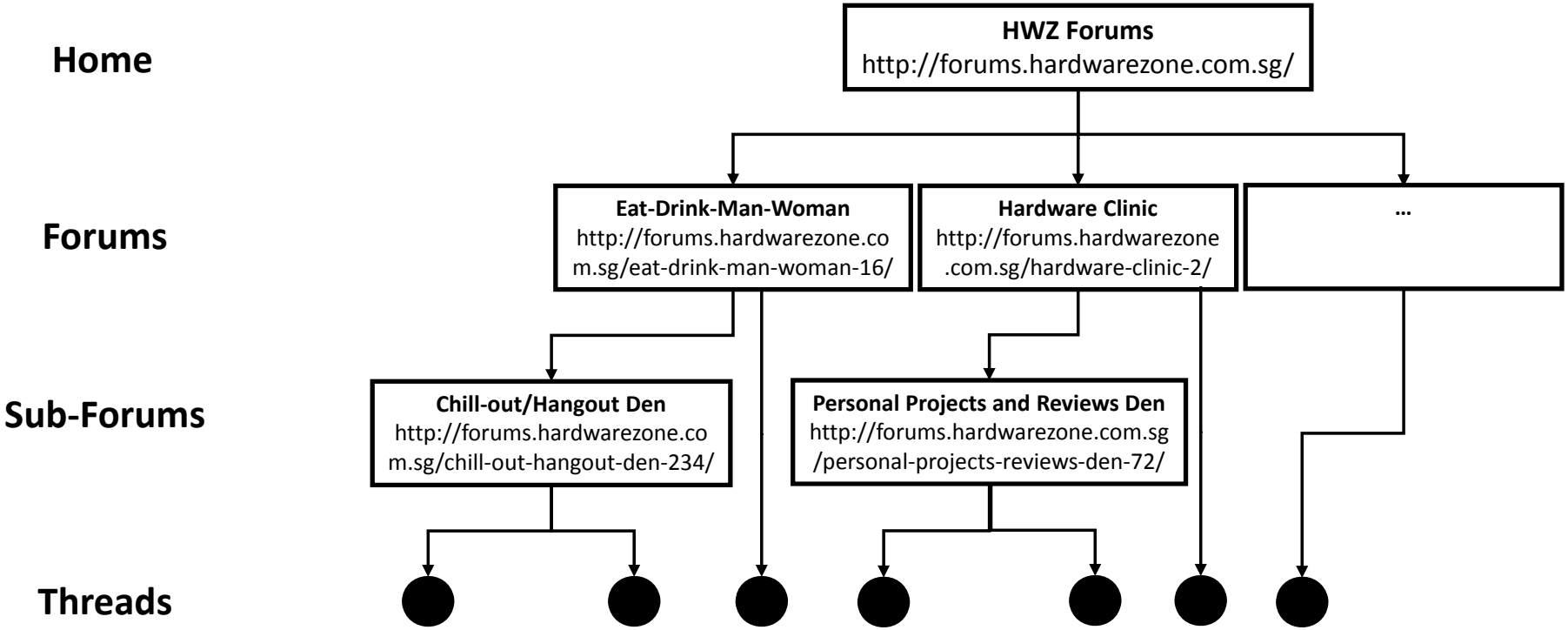
Scrapy

Cassandra

Flask
web development, one drop at a time

MetricsGraphics.js

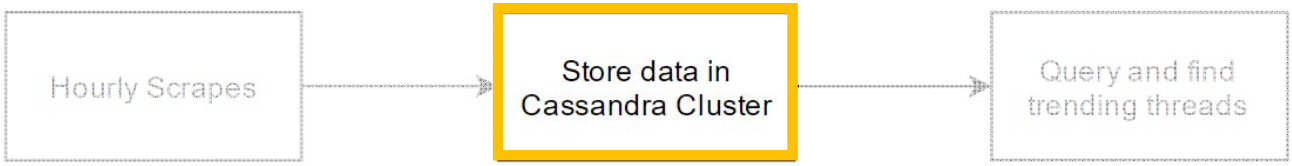
Scraping for Data



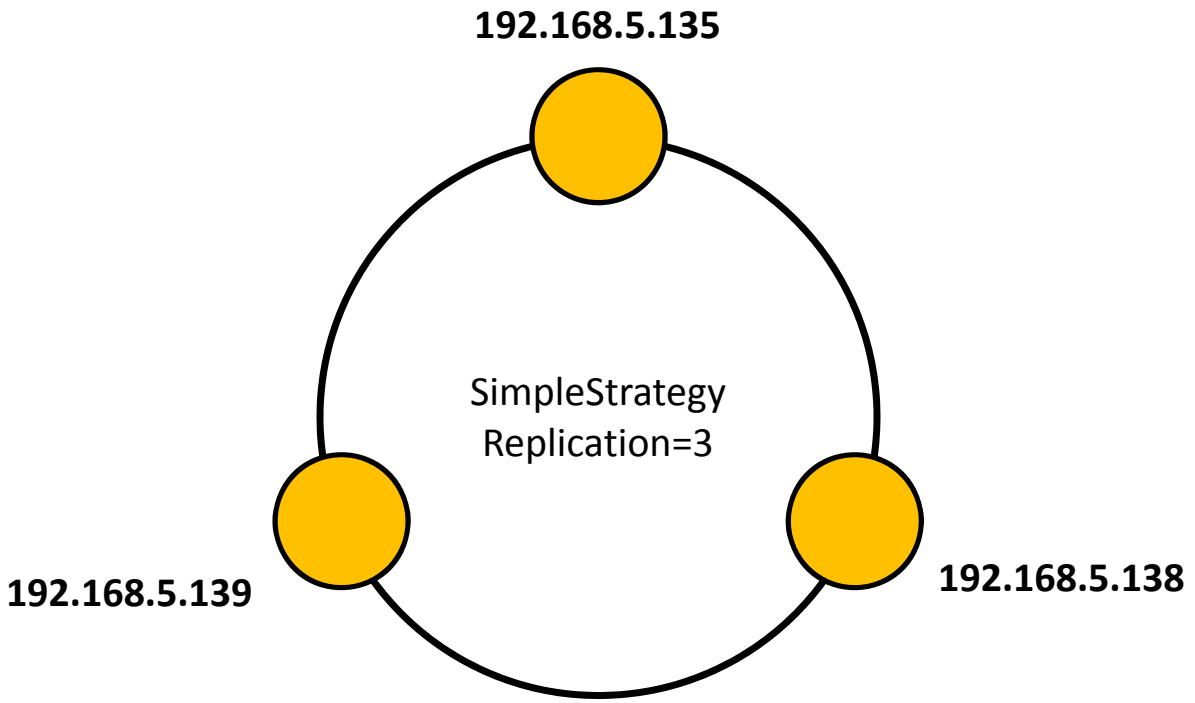
Scraping for Data

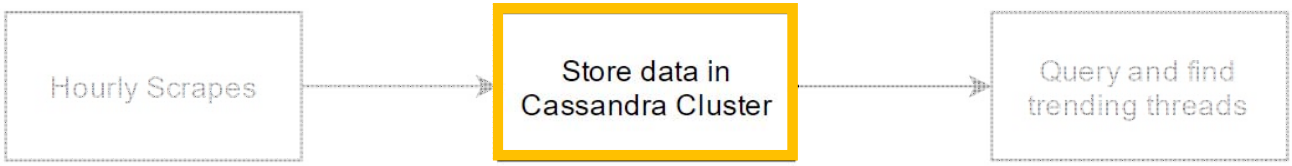
Item	Type
thread_id	int
thread_title	text
forum_name	text
thread_url	text
starter_name	text
starter_id	int
replies_count	int
views_count	int
scrape_time	timestamp
last_post_time	timestamp

Cassandra Cluster



Cassandra Cluster



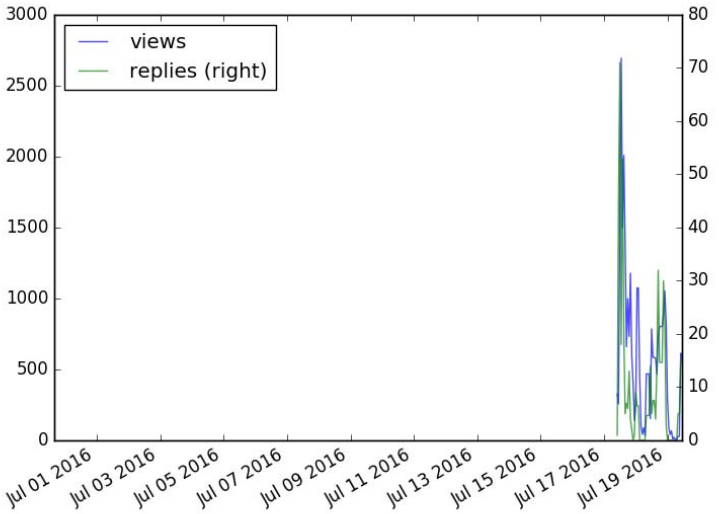


Trending Threads

Scoring Metrics

Max

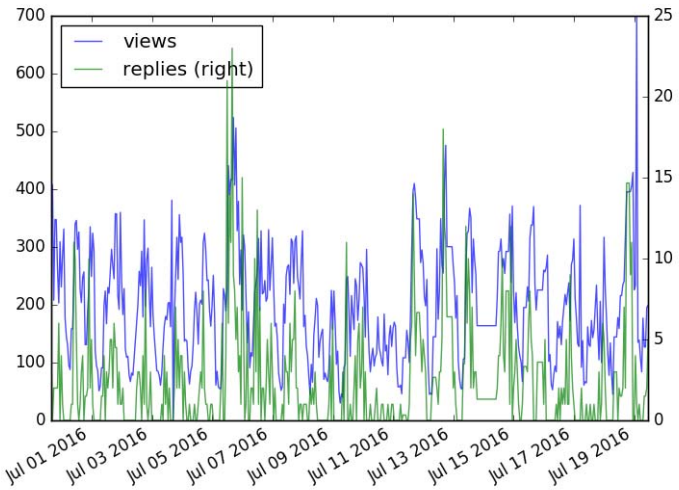
Life as an insurance agent soliciting customers in malls

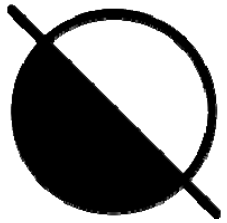


Trending Threads

Scoring Metrics

[Official] Geforce GTX 1080/GTX 1070 discussion



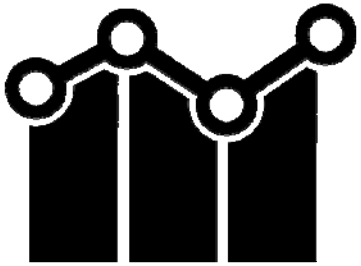


MaxDiff

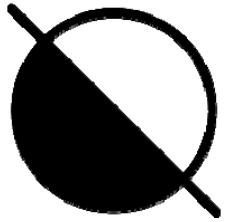
Trending Threads

Scoring Metrics

Max



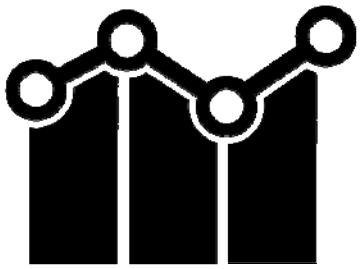
Adj-Zscore



MaxDiff

Trending Threads

Scoring Metrics



Adj-Zscore

```
zscore(row)[-1] *
np.sqrt(
       np.average(
              row,
              weights=2**np.arange( 0,
                                    10,
                                    10.0/len(row))
                    )
         )
```

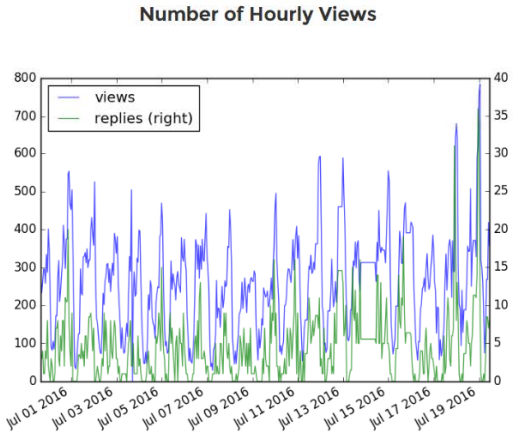
Interface

HWZ TRENDS METRIC EXAMPLES

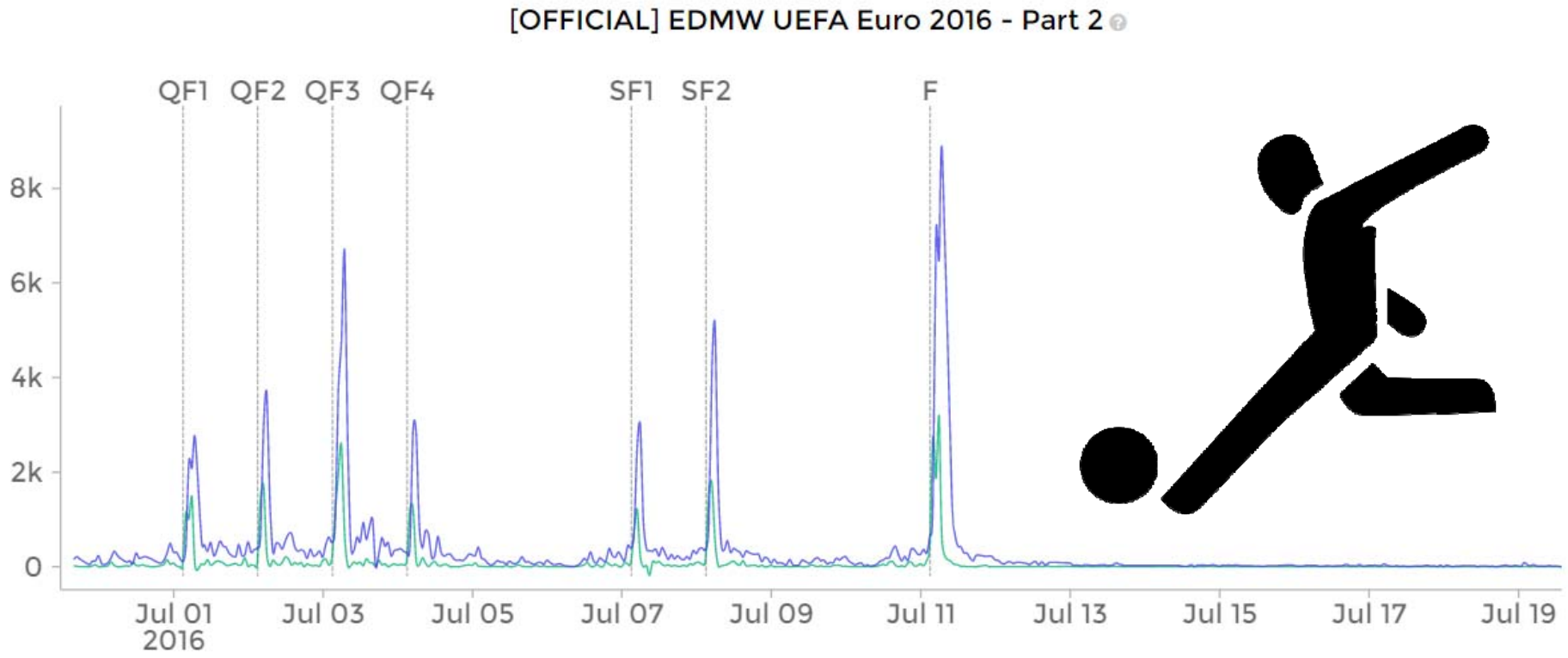
Today's Trending Threads

Scored using the faz metric

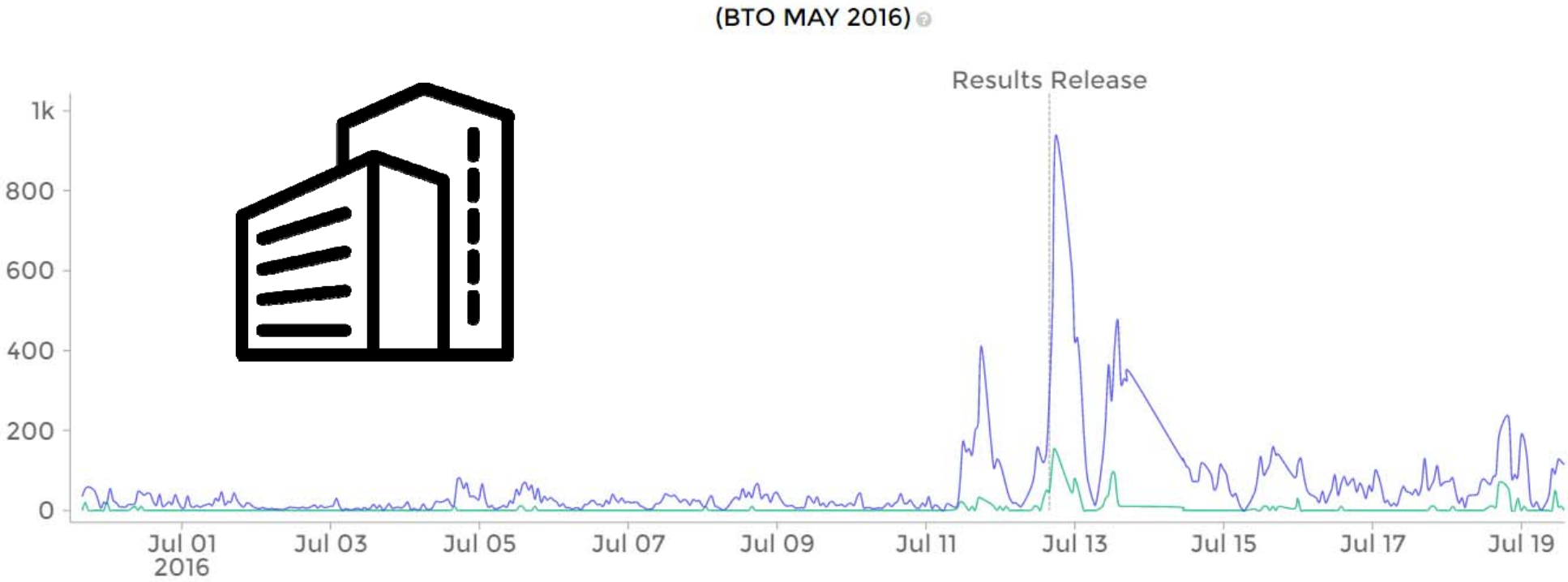
Rank	Thread Title	Forum Name
1	The Watch Thread - Part 10	Fashion & Grooming
2	[GPGT] Bt Batok is farking jinxed, Another new lift accident, old auntie fell and broke her wrists	Eat-Drink-Man-Woman
3	[GPGT] ANTS problem? Sic most effective ant bait i ever used...	Eat-Drink-Man-Woman
4	[Official] Geforce GTX 1080/GTX 1070 discussion	Hardware Clinic
5	[BREAKING NEWS!] Vehicle Entry Permit in effect for Singapore cars entering Johor	Eat-Drink-Man-Woman
6	30s-40s chitchat club - Part 22	Eat-Drink-Man-



Event Detection?



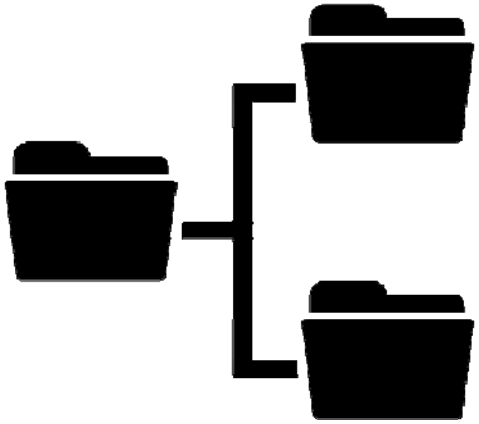
Event Detection?



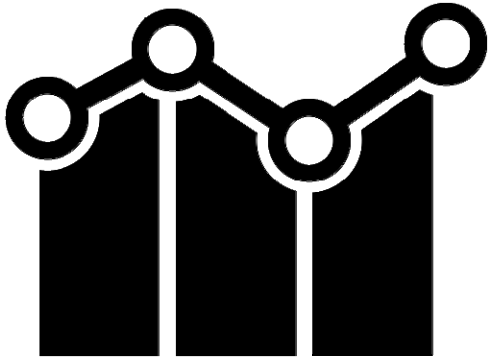
Attribution

- Icons
  - Folder Tree by To Uyen from the Noun Project
  - analytics by Hyhyhehe from the Noun Project
  - link by Icon Fair from the Noun Project
  - Picture Content by Oliviu Stoian from the Noun Project
  - Folder by Creative Stall from the Noun Project
  - Playing Cards by Raymond Felix from the Noun Project
  - database by Paweł Wypych from the Noun Project
  - Precent by ✦ Shmidt Sergey ✦ from the Noun Project
  - Dialect by Guillaume Beaulieu from the Noun Project
  - text by allenwang from the Noun Project
  - Contrast by Musket from the Noun Project
  - crest by TAKAHASHI YOSHIOMI from the Noun Project
  - soccer by David Padrosa from the Noun Project
  - buildings by Creative Stall from the Noun Project

Questions?



Website Classification



Trending Threads