In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
names = ['acquisitions', 'investments']
dfs = [pd.read_csv(name + '.csv') for name in names]
```

In [3]:
```
for n, df in enumerate(dfs):
    #Drop NaNs in company category column and redo the index
    df = df[df['company_category_list'].notnull()]
    df.index = range(len(df))

    #Assign each unique, single tag a column number
    unique = df['company_category_list'].astype(str).unique()
    singles = []
    for tag in unique:
        singles.extend(tag.split('|'))
    singles = pd.unique(singles)
    tagInds = pd.Series(data=range(len(singles)), index=singles)

    #Make a dataframe where each row has a 1 for each of its tags, 0 otherwise
    lists = df['company_category_list'].astype(str).map(lambda x: x.split('|'))
    arr = np.zeros((len(lists), len(tagInds)))
    for i, l in enumerate(lists):
        for tag in l:
            arr[i][tagInds[tag]] = 1
    dummies = pd.DataFrame(arr, columns=singles)

    #Isolate the top 50 tags and reassign tag numbers
    top50 = dummies.sum().sort_values(ascending=False).head(50)
    tagInds = pd.Series(range(50), index=top50.index)

    #Make another dataframe but with an 'Other' column for tags not in the top 50
    arr = np.zeros((len(lists), 51))
    for i, l in enumerate(lists):
        for tag in l:
            if tag in tagInds:
                arr[i][tagInds[tag]] = 1
            else:
                arr[i][50] = 1
    dummies = pd.DataFrame(arr, columns=list(tagInds.index) + ['Other'])

    #Add the dummy columns to the data
    dfs[n] = df.join(dummies)
```

In [5]:
```
#Dummy dataframe can be retrieved by fetching the last 51 columns
investment_dummies = dfs[1].iloc[:, -51:]
investment_dummies.sum()
```

Out[5]:
```
Software                  24641.0
Mobile                    18065.0
Biotechnology             13954.0
E-Commerce                12354.0
Enterprise Software       11766.0
Curated Web                8999.0
Advertising                8528.0
Health Care                7639.0
Social Media               7587.0
Analytics                  7548.0
SaaS                       7288.0
Finance                    6010.0
Internet                   5410.0
Technology                 5200.0
Games                      5167.0
Education                  4871.0
Apps                       4772.0
Hardware + Software        4724.0
Clean Technology           4412.0
Health and Wellness        4345.0
Security                   4049.0
Big Data                   3779.0
Marketplaces               3726.0
Cloud Computing            3644.0
Semiconductors             3579.0
Video                      3477.0
FinTech                    3037.0
Services                   2920.0
Fashion                    2913.0
Search                     2574.0
Manufacturing              2536.0
Web Hosting                2523.0
Travel                     2439.0
Sales and Marketing        2267.0
Medical                    2258.0
Messaging                  2223.0
Networking                 2187.0
Retail                     2160.0
Startups                   2127.0
Media                      2122.0
Digital Media              2084.0
News                       2027.0
Music                      2024.0
Financial Services         2019.0
Android                    1998.0
Information Technology     1906.0
Social Network Media       1903.0
iPhone                     1833.0
Hospitality                1800.0
Entertainment              1761.0
Other                     92949.0
dtype: float64
```

In [6]:
```
for df, name in zip(dfs, names):
    #Save to file
    df.to_csv(name + '_with_dummies.csv')
```

In [ ]:
```

```