From last notebook...

In [1]:

```
from nltk.corpus import treebank as tb
words = [(fileid,sentnum,wordnum,word)
         for fileid in tb.fileids()[:1]
         for sentnum,sent in enumerate(tb.tagged_sents(fileid))
         for wordnum,(word,pos) in enumerate(sent)
         if pos in ['NNP','NNPS']]

from sets import Set
sents = [
         {'fileid':fileid,
          'sentnum':sentnum,
          'nnp':Set([word for word,pos in sent if pos in ['NNP','NNPS']]),
          'nnp_loc':{word:wordnum for wordnum,(word,pos) in enumerate(sent) if pos i
          'word_loc':{word:wordnum for wordnum,(word,pos) in enumerate(sent)}
          }
         for fileid in tb.fileids()
         for sentnum,sent in enumerate(tb.tagged_sents(fileid))
         ]
```

Generate proper PB output now.

In [8]:

```
from nltk.tokenize import word_tokenize
from nltk.metrics import edit_distance
from FB_to_PB import *
f2p = fb_to_pb_map('FB_to_PB.csv')
fb_dir = 'tmp.business_Freebase'
for fbr in ['acquisition']:
    pbr = f2p[fbr]['PROPBANK_RELATION']
    pb = pb_instances(fb_dir,fbr,f2p)
    for pbi_num,pbi in enumerate(pb):
        if len([word for arg,word in pbi.iteritems() if word.strip() == '']) > 0:
            continue

        pbi_words = Set([token
                         for arg,word in pbi.iteritems()
                         for token in word_tokenize(word)
                         if token not in [',','.']])

        tb_sents = [sent for sent in sents if pbi_words.issubset(sent['nnp'])]
        if(len(tb_sents) > 0):
            for sent in tb_sents:
                rel_dist = [(wordnum,word,edit_distance(pbr,word)) for wordnum,word
                (rel_wordnum, rel_word, rel_edit_dist) = sorted(rel_dist, key=lambda

                roleset = '%s.01'%pbr
                inflection = '-----'
                print '%s %d %d FB2PB %s %s' % (sent['fileid'], sent['sentnum'], rel
                for arg,word in pbi.iteritems():
                    (long_tok_len, long_tok) = sorted([(len(token),token) for token
                    arg_word_num = sent['word_loc'][long_tok]
                    print '%d:0-%s' % (arg_word_num,arg),

                print ''
```

```
wsj_1794.mrg 11 14 FB2PB acquire.01 ----- 6:0-ARG0 9:0-ARG1
wsj_1856.mrg 72 6 FB2PB acquire.01 ----- 3:0-ARG0 6:0-ARG1
wsj_0111.mrg 19 5 FB2PB acquire.01 ----- 39:0-ARG0 42:0-ARG1
wsj_0111.mrg 20 6 FB2PB acquire.01 ----- 24:0-ARG0 35:0-ARG1
wsj_1809.mrg 55 20 FB2PB acquire.01 ----- 18:0-ARG0 1:0-ARG1
wsj_1324.mrg 6 4 FB2PB acquire.01 ----- 7:0-ARG0 0:0-ARG1
wsj_0162.mrg 30 24 FB2PB acquire.01 ----- 34:0-ARG0 34:0-ARG1
wsj_0666.mrg 13 11 FB2PB acquire.01 ----- 24:0-ARG0 24:0-ARG1
wsj_1157.mrg 31 26 FB2PB acquire.01 ----- 13:0-ARG0 14:0-ARG1
wsj_1157.mrg 36 5 FB2PB acquire.01 ----- 23:0-ARG0 24:0-ARG1
wsj_1317.mrg 42 6 FB2PB acquire.01 ----- 3:0-ARG0 4:0-ARG1
wsj_1317.mrg 43 4 FB2PB acquire.01 ----- 35:0-ARG0 36:0-ARG1
wsj_1856.mrg 70 2 FB2PB acquire.01 ----- 1:0-ARG0 2:0-ARG1
wsj_1856.mrg 72 6 FB2PB acquire.01 ----- 6:0-ARG0 36:0-ARG1
wsj_1875.mrg 112 0 FB2PB acquire.01 ----- 5:0-ARG0 6:0-ARG1
wsj_1885.mrg 5 25 FB2PB acquire.01 ----- 24:0-ARG0 25:0-ARG1
wsj_0502.mrg 5 0 FB2PB acquire.01 ----- 18:0-ARG0 24:0-ARG1
wsj_1192.mrg 13 0 FB2PB acquire.01 ----- 16:0-ARG0 9:0-ARG1
wsj_0477.mrg 9 18 FB2PB acquire.01 ----- 6:0-ARG0 3:0-ARG1
wsj_1015.mrg 48 12 FB2PB acquire.01 ----- 16:0-ARG0 8:0-ARG1
wsj_0111.mrg 1 27 FB2PB acquire.01 ----- 27:0-ARG0 32:0-ARG1
wsj_2394.mrg 10 3 FB2PB acquire.01 ----- 51:0-ARG0 34:0-ARG1
```

```
In [ ]:
```

```
In [ ]:
```