Basic Observation and data summary

Load data

Data contains itemID, itemTitle, condition and price. The key to classify items from CSA to JWL is the itemTitle. We load itemTitle data from "CSA5k.txt" and "JWL35k.txt" file.

In [5]:

```
file_CSA = open("CSA5k.txt")
file_JWL = open("JWL35k.txt")
data_CSA = []
data_JWL = []
for line in file_CSA:
    title = line.strip("\n").split("\t")[1].decode('utf-8').lower()
    if title != -1:
        data_CSA.append((title,"CSA"))
for line in file_JWL:
    title = line.strip("\n").split("\t")[1].decode('utf-8').lower()
    if title != -1:
        data_JWL.append((title,"JWL"))
```

Classification model selection

Package used here

In [17]:

```
from textblob.classifiers import NaiveBayesClassifier
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.pipeline import Pipeline
from sklearn import linear_model
from sklearn.cross_validation import KFold
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.linear_model import SGDClassifier
import numpy as np
```

In order to observe the performance of the model in the future, we split data to two parts. One is train data, the other is test data. In this case, I choose the proportion is 7:3.

In [10]:

```
TRAIN_PROP = 0.7
train = data_CSA[:int(len(data_CSA)*TRAIN_PROP)] +
data_JWL[:int(len(data_JWL)*TRAIN_PROP)]
test = data_CSA[int(len(data_CSA)*TRAIN_PROP):] + data_JWL[int(len(data_JWL
)*TRAIN_PROP):]
```

The idea of classification is that we try to tokenize the title string into short words by space. We count the occurrences of tokens in these two categories. We build a classifier based on this information. In order to avoid the impact of tokens that occur very frequently in a given corpus are hence empirically less informative than features that occur in a small fraction of the training corpus, we transfer occurrences of raw data to tf-idf.We can choose SVM and Naive Bayes classifier. Here we use 10-fold cross validation to see which model is better. We split train data to 10 subsets. We use 1 piece as test data and the rest 9 pieces as train data to model. We will test the prediction error of two models and choose a better one.

In [19]:

```
np.random.shuffle(train)
kf = KFold(len(train), n_folds=10)
SVM_error = []
NB_error = []
for train_index,test_index in kf:
    train_val, test_val = np.array(train)[train_index], np.array(train)
[test_index]
    ##NaiveBayes classifier
    text_nb = Pipeline([('vect',CountVectorizer()),
('tfidf',TfidfTransformer()),('clf',MultinomialNB()),])
    text_nb = text_nb.fit([x[0] for x in train_val],[x[1] for x in train_val
])
    predicted_nb = text_nb.predict([x[0] for x in test_val])
    NB_error.append(np.mean(predicted_nb == [x[1] for x in test_val]))
    ##SVM classifier
    text_svm = Pipeline([('vect',CountVectorizer()),
('tfidf',TfidfTransformer()),('clf',SGDClassifier()),])
    text_svm = text_svm.fit([x[0] for x in train_val],[x[1] for x in train_v
al])
    predicted_svm = text_svm.predict([x[0] for x in test_val])
    SVM_error.append(np.mean(predicted_svm == [x[1] for x in test_val]))
```

In [31]:

```
print 1-np.mean(NB_error)
print 1-np.mean(SVM_error)
```

```
0.0303928571429
0.00649675324675
```

We calculate the average prediction cross validation error of two models. We found that SVM has less misclassification error so that SVM is a better model to choose.

Classify and test the performance

In [26]:

```
text_svm = Pipeline([('vect',CountVectorizer()),('tfidf',TfidfTransformer()
),('clf',SGDClassifier()),])
text_svm = text_svm.fit([x[0] for x in train],[x[1] for x in train])  ##fina
l model
predicted_svm = text_svm.predict([x[0] for x in test])
test_error = np.mean(predicted_svm == [x[1] for x in test])
```

In [29]:

In [29]:

```
print 1-test_error
```

```
0.00808333333333
```

We use train data to fit the model and use test data as new input to observe the performance of SVM model. Finally the test prediction error is 0.008. We cannot claim that this rate is tolerent or not. We have 12000 test data, which means around 100 of them are misclassified. If everyday listing contains large scale of data, 0.8% misclassification rate might cause problem when doing classification. Therefore, we might need to discuss about the tolerance of this classification model and try to observe those misclassified data to check the reason and the pattern of why this method lead these item misclassified.

In [37]:

```
index = np.where(predicted_svm != [x[1] for x in test])
index
```

Out[37]:

```
(array([   56,    71,    97,   113,   126,   128,   140,   147,   150,
          162,   163,   192,   218,   226,   239,   246,   252,   270,
          317,   320,   347,   453,   502,   505,   530,   533,   572,
          578,   599,   624,   629,   630,   632,   635,   640,   663,
          694,   768,   810,   825,   855,   866,   869,   880,   898,
          908,   959,   963,   987,  1019,  1046,  1053,  1065,  1067,
         1073,  1082,  1126,  1132,  1173,  1198,  1222,  1226,  1245,
         1272,  1273,  1280,  1325,  1345,  1351,  1359,  1371,  1452,
         1470,  1484,  3007,  3339,  3909,  4584,  4850,  4906,  5452,
         5833,  6486,  6916,  6922,  6993,  7975,  8157,  8546,  8711,
         9720,  9742, 10863, 10991, 11290, 11323, 11399]),)
```

In [38]:

```
np.array(test)[index]
```

Out[38]:

```
array([[u'vintage englert trucking co silvertone & goldtone #1 key chain',
        u'CSA'],
       [ u'shoe sneaker shoelace charm decoration i love heart names female
k kena',
        u'CSA'],
       [ u'peace hippie boho fair trade ethnic hill tribe nepal handbag pom
poms bells (29)',
        u'CSA'],
       [u'thanksgiving hair bow on an alligator clip', u'CSA'],
       [ u'skidlid original motorcycle half helmet in bomber pin up silver
/ blue xs-2xl',
        u'CSA'],
       [ u'corduroy breton cap with brass buttons ; john lennon fisherman s
tyle beatles',
        u'CSA'],
       [u'barely breezies s/2 seamless modern teardrop bras a211857',
        u'CSA'],
       [u'steve madden beasst bronze snake us 7', u'CSA'],
       [ u'big cheetah cheer bow glitter white purple ribbon girls uniform
accessories ties',
        u'CSA'],
       [u'authentic vans grey / black', u'CSA'],
```

```
       [ u'ganz women\u2019s scarf with silver beads & jewel accents
various colors er24452',
        u'CSA'],
       [u'vintage avenue 38" 23mm metal belt w/ dangling leaf charms silver
tone',
        u'CSA'],
       [u'native miller glow child in fire truck red sizes 4-13', u'CSA'],
       [ u'this soul belongs to jesus (christ gold silver cross icon holy o
rthodox) t-shirt',
        u'CSA'],
       [ u'corkys regan jeweled flip flops silver/gold super bling stones 6
-11 new in box',
        u'CSA'],
       [ u'vtg. brighton alligator stamped leather silver heart angel buckl
es sz.s 42409 #9',
        u'CSA'],
       [ u"1set men's groom crystal cufflinks & necktie tie clasp clip bar
pin fancy gift",
        u'CSA'],
       [u'taos pizazz pewter', u'CSA'],
       [u'vintage round metal eyewear  ', u'CSA'],
       [ u'thunder road motor oil highway to hell old school devil skull sw
eatshirt ws543',
        u'CSA'],
       [u'5" boutique stacked hair bow with gator clip', u'CSA'],
       [u'sassy candy monogrammed retractable id badge reel', u'CSA'],
       [u'mint$60 *express* brown gold sparkle glitter stretch wrap top s m
',
        u'CSA'],
       [ u'a crystal  hair pin glass pearl 269p, rose color. a good gift fo
r a girl',
        u'CSA'],
       [ u'hair chains for wedding prom bridal silver/gold crystal tone on
grips or combs',
        u'CSA'],
       [ u'clear lens eyeglasses metal & plastic horn rim frame glasses bla
ck silver',
        u'CSA'],
       [ u'new wedding garter french pink white prom homecoming
getthegoodstuff love charm',
        u'CSA'],
       [ u'durable aluminum credit card case box id credit card wallet hold
er case',
        u'CSA'],
       [u'the legend of zelda - 11pc. necklace gift set', u'CSA'],
       [ u'women bling big pearl silver rhinestone metal chain belt waist h
ip wedding ring ',
        u'CSA'],
       [ u'vtg feather wallet lee sands set peacock pheasant earrings nativ
e american w box',
        u'CSA'],
       [u'hippie beaded rope bracelet', u'CSA'],
       [ u'clearance closeout lot 5 reading glasses 4 cases 1 hard case men
+1.25 mrs5-4748',
        u'CSA'],
       [ u'wholesale 7pcs 60g clip in 100% human hair extensions many color
s 16" 20" 24" ',
        u'CSA'],
       [ u'christion audigier don ed hardy red 3 xl.new with tag  king dog
forever ',
```

```
        u'CSA'],
       [u'michael kors mens tie, gold with brown and blue squares',
u'CSA'],
       [u"i love you but i've chosen trance turquoise v-neck", u'CSA'],
       [u'gi watch caps - stylish and warm - white', u'CSA'],
       [u'25th birthday gift birthday tshirt gift for him birthday outfit b
d-101',
        u'CSA'],
       [u'scarf with jewelry - silver plated pendant', u'CSA'],
       [u'hairware color 2" bobby pins - 60 ct cp-72x', u'CSA'],
       [ u'5x fashion twinkle crystal faux pearls bridal bridemaid headband
tiara crown',
        u'CSA'],
       [ u'charismatico red and gold fire inspired mohawk styled red feathe
r headdress',
        u'CSA'],
       [ u'hair styling clip women hair comb band hairpin fashion hair acce
ssories dz88',
        u'CSA'],
       [ u'clown dots butterfly fairy angel  wings, elastic arm straps glit
ter & feathers ',
        u'CSA'],
       [u'pakistani indian shalwar kameez kurta summer sale leline', u'CSA'
],
       [ u'i love my cocker spaniel dog key chain key ring - zipper pull -
nwt - nice!',
        u'CSA'],
       [ u'rock and republic keidis crystal skull wide leg jean antrax wash
org. $267  ',
        u'CSA'],
       [u'bridal flower rhinestone crystal wedding necklace earrings set n2
93',
        u'CSA'],
       [u'kangol  gold   shine  links  adjustable  baseball  cap style k404
st',
        u'CSA'],
       [ u'indian party wear saree wedding designer new bollywood sari .eth
nic pakistani',
        u'CSA'],
       [ u'free shipping new sexy orange black bridal wedding garters\xa0-
getthegoodstuff',
        u'CSA'],
       [u'marvel tshirts size large bidding on 2 different designs',
u'CSA'],
       [u"i'm not fat coffee & tea mug", u'CSA'],
       [ u'hot girls hair ties pearl hairband elastic hair tie rope strap p
onytail holder',
        u'CSA'],
       [u'l.o.g.g. h&m', u'CSA'],
       [ u'genuine original prada calzature donna nappa sport ladies shoe b
lack nero in box',
        u'CSA'],
       [ u"big lot of women's handmade scarves 10 total beautiful styles &
colors new look!",
        u'CSA'],
       [ u'wedding desiner party wear saree pakistani ethnic festival tradi
tional sari',
        u'CSA'],
       [u'renaissance sailor/pirate satin sash with fringe', u'CSA'],
       [u'columbia tenacity ii', u'CSA'],
       [u'fash chic mod single houlon boho handbag', u'CSA'],
```

```
       [u'rasn chic mod circle bowler hobo handbag', u'CSA'],
       [ u'usa red mix df1090 lace front wig pin part heat ok iron safe han
d tied sas*',
        u'CSA'],
       [u'chinese laundry pacific size 7.5 org 89$', u'CSA'],
       [ u'7a unprocessed virgin hair loose wave lace closure brazilian fre
e part 4x4 inch',
        u'CSA'],
       [u'[uniq] 1st mini album [eoeo] cd+booklet+photo card+poster sealed
k-pop',
        u'CSA'],
       [u'jordan 13', u'CSA'],
       [u'new jp kids legging 4t purple free shipping', u'CSA'],
       [u'solid black totes "titan" max strength automatic folding umbrella
nwt',
        u'CSA'],
       [u'etienne aigner silver jeweled dianne sandals 11m retail $59',
        u'CSA'],
       [ u'embellished white rhinestone torrid forever rue 21 wet seal alfa
ni blouse m/l',
        u'CSA'],
       [ u'tan & gold sparkle womans size 7 dress & free choker necklace in
gold & teal',
        u'CSA'],
       [ u'bejeweled beaded trade show badge id license name tag holder lim
e lanyard',
        u'CSA'],
       [ u'iheartraves electric styles white light up el wire tie rave fest
ival wear',
        u'CSA'],
       [ u'vintage plastic brown cream white lucite art deco 2 piece small
belt buckle',
        u'JWL'],
       [ u'sapphire & diamond wedding engagement ring sz 5 sz 6 sz 7 sz 8 s
z 9 sz 10 prince',
        u'JWL'],
       [ u'buy 2 get 1 free mlb major league baseball silver suit shirt tie
bar clip clasp',
        u'JWL'],
       [ u"new cute sexy women's bikini set bandage swimwear swimsuit beach
wear oe",
        u'JWL'],
       [u'az womens cute fruit hair clips girls child lovely kid gift ',
        u'JWL'],
       [ u'yf007 new bridal waist sash satin belt bridesmaid wedding
evening party dress ',
        u'JWL'],
       [ u'chuunibyou unisex kids mens womens boys girls durable color wris
t watches',
        u'JWL'],
       [ u'bright snowflake bolo tie gemstone bola ties suit shirt necktie
necklace',
        u'JWL'],
       [ u'purse - unique brown leather fringe 4" coin purse w baby indian
broach pin',
        u'JWL'],
       [ u'emerald green post convertiblez jacket set  gold flower jacket a
nd convertiblez ',
        u'JWL'],
       [u'silk bracelet with 180.00ctw  pink rhondonites retail $99 nwt',
        u'JWL'],
```

```
        u'JWL'],
       [u'fashion jewelry black lace hollow mesh party half accessory mask
women',
        u'JWL'],
       [u'winter flora miniature sheet first day cover (isle of man
stamps)',
        u'JWL'],
       [u'"jj" jonette jewelry bronze pewter \'girl zipper pull\' ~ jacket/
purse',
        u'JWL'],
       [ u'lot ralph lauren white house black market my flat london treska
jessica simpson',
        u'JWL'],
       [ u'monet white daisy pin,gold tone,2.5" ,retro,summer dress
party,scarf,purse,coat',
        u'JWL'],
       [u'vintage rhinestone silk strap', u'JWL'],
       [ u'sexy womens harness tassel layered bikini beach necklace belly w
aist body chains',
        u'JWL'],
       [ u'men women clothing bag collocation alloy bird belt buckle high l
eather bracelet',
        u'JWL'],
       [u'girls headband with yellow/red flower', u'JWL'],
       [ u'set 2 brand new american eagle men leather bracelets one size o/
s  nwot',
        u'JWL'],
       [u'scarf 77x18 in dark blue cross-hatch 100 % indian cotton hand-bat
iked',
        u'JWL'],
       [ u'new womens orchids rhinestone-studded coconut tree classic cloth
es accessory ',
        u'JWL']],
      dtype='<U80')
```

Just by observation, we can see some titles are really confusing, such as "new bridal waist sash satin belt bridesmaid wedding evening party dress", which contains "dress" but belong to JWL category. These kinds of cases, we might need to dig into it in the future.