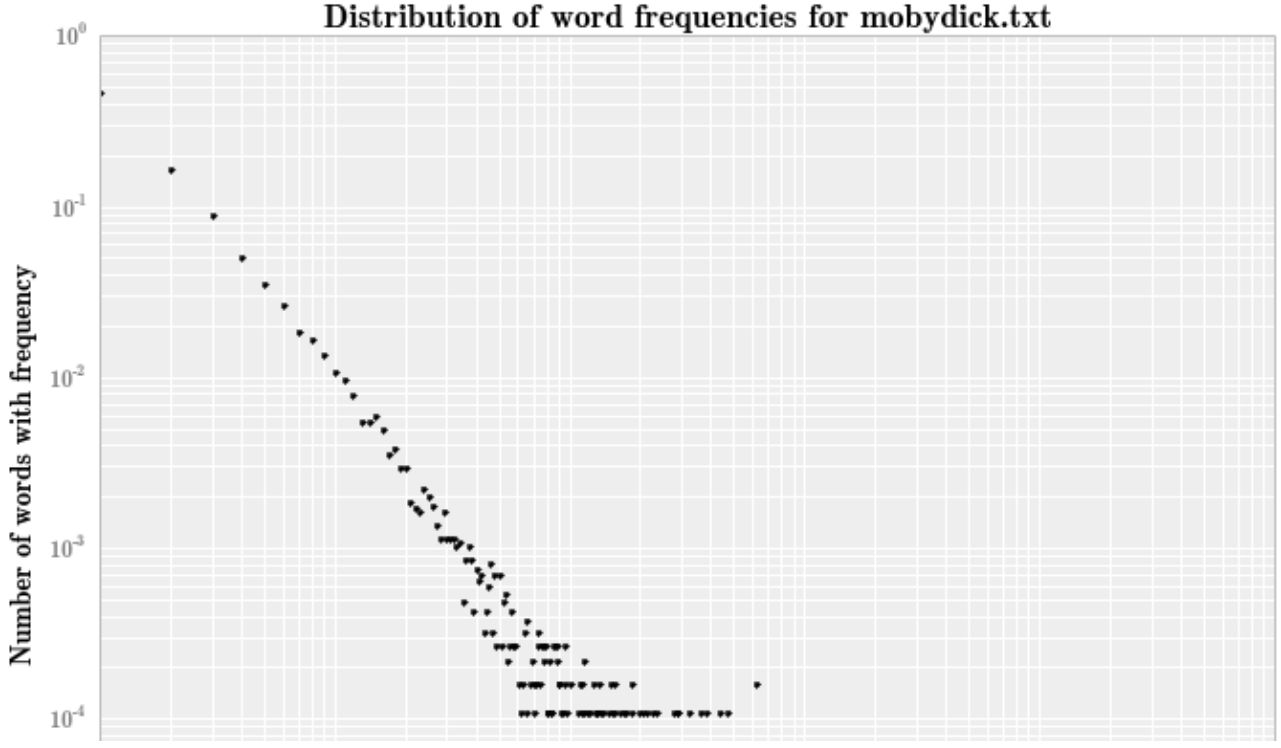
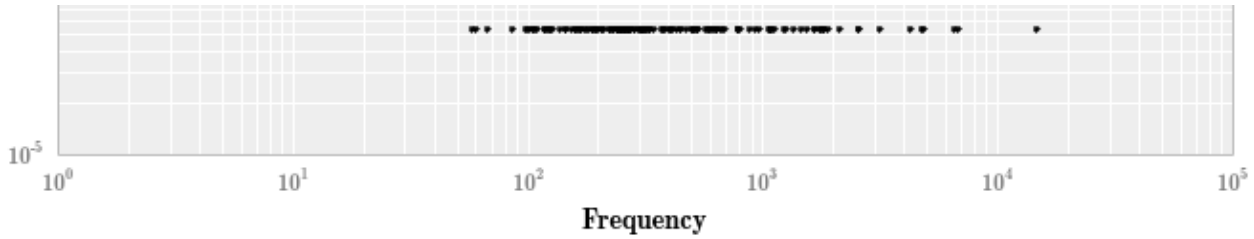
Homework 3: Question 3

In [1]:
```
from collections import Counter
from ggplot import *
import math
```

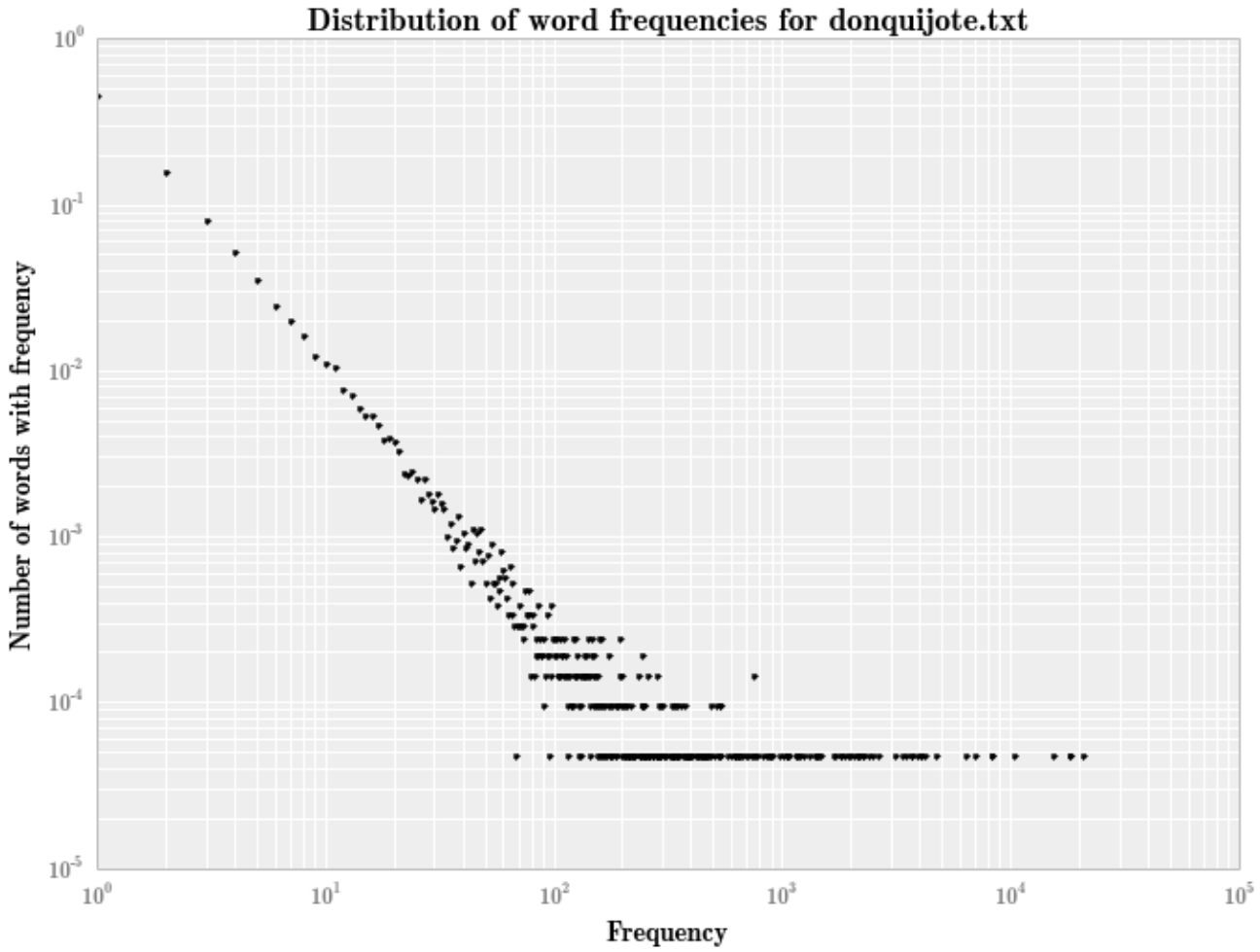
In [2]:
```
def plot_word_freq(filename):
    with open (filename, "r") as myfile:
        words=[s.strip() for s in myfile.readlines()]
        counter = Counter(Counter(words).values())

    total = float(sum(counter.values()))
    ctr = {k:(v/total) for k,v in counter.items()}
    x,y = zip(*(ctr.items()))
    loglog(x,y,'k.')
    title("Distribution of word frequencies for " + filename)
    xlabel("Frequency")
    ylabel("Number of words with frequency")
    return x,y
```

In [4]:
```
x,y = plot_word_freq("mobydick.txt")
```





In [5]:
```
x,y = plot_word_freq("donquijote.txt")
```



Q3.ii) Estimating α and x_{min}.

In [6]:
```
def mle_alpha(x,xmin):
    x_filt = filter(lambda i: i >= xmin, x)
    lxmin = log(xmin)
    logsum = sum([log(d)-lxmin for d in x_filt])
    return 1.0+len(x_filt)/logsum
```

In [9]:
```
def find_xmin(filename):
    with open (filename, "r") as myfile:
        words=[s.strip() for s in myfile.readlines()]
        freqs = Counter(words).values()
        for xmin in xrange(1,16):
            alpha = mle_alpha(freqs,xmin)
            print "xmin = %d, alpha = %g" % (xmin,alpha)
```

In [10]:
```
find_xmin("mobydick.txt")
```

```
xmin = 1, alpha = 2.15047
xmin = 2, alpha = 2.05763
xmin = 3, alpha = 2.03793
xmin = 4, alpha = 2.01739
xmin = 5, alpha = 2.02127
xmin = 6, alpha = 2.0227
xmin = 7, alpha = 2.0192
xmin = 8, alpha = 2.03006
xmin = 9, alpha = 2.02384
xmin = 10, alpha = 2.01437
xmin = 11, alpha = 2.00957
xmin = 12, alpha = 1.99772
xmin = 13, alpha = 1.98742
xmin = 14, alpha = 1.99372
xmin = 15, alpha = 1.99142
```

In [11]:
```
find_xmin("donquijote.txt")
```

```
xmin = 1, alpha = 2.01752
xmin = 2, alpha = 1.90257
xmin = 3, alpha = 1.87391
xmin = 4, alpha = 1.86769
xmin = 5, alpha = 1.86212
xmin = 6, alpha = 1.85952
xmin = 7, alpha = 1.86578
xmin = 8, alpha = 1.86857
xmin = 9, alpha = 1.86923
xmin = 10, alpha = 1.8771
xmin = 11, alpha = 1.87975
xmin = 12, alpha = 1.87398
xmin = 13, alpha = 1.87955
xmin = 14, alpha = 1.88191
```

```
xmin = 15, alpha = 1.88651
```

In [ ]: