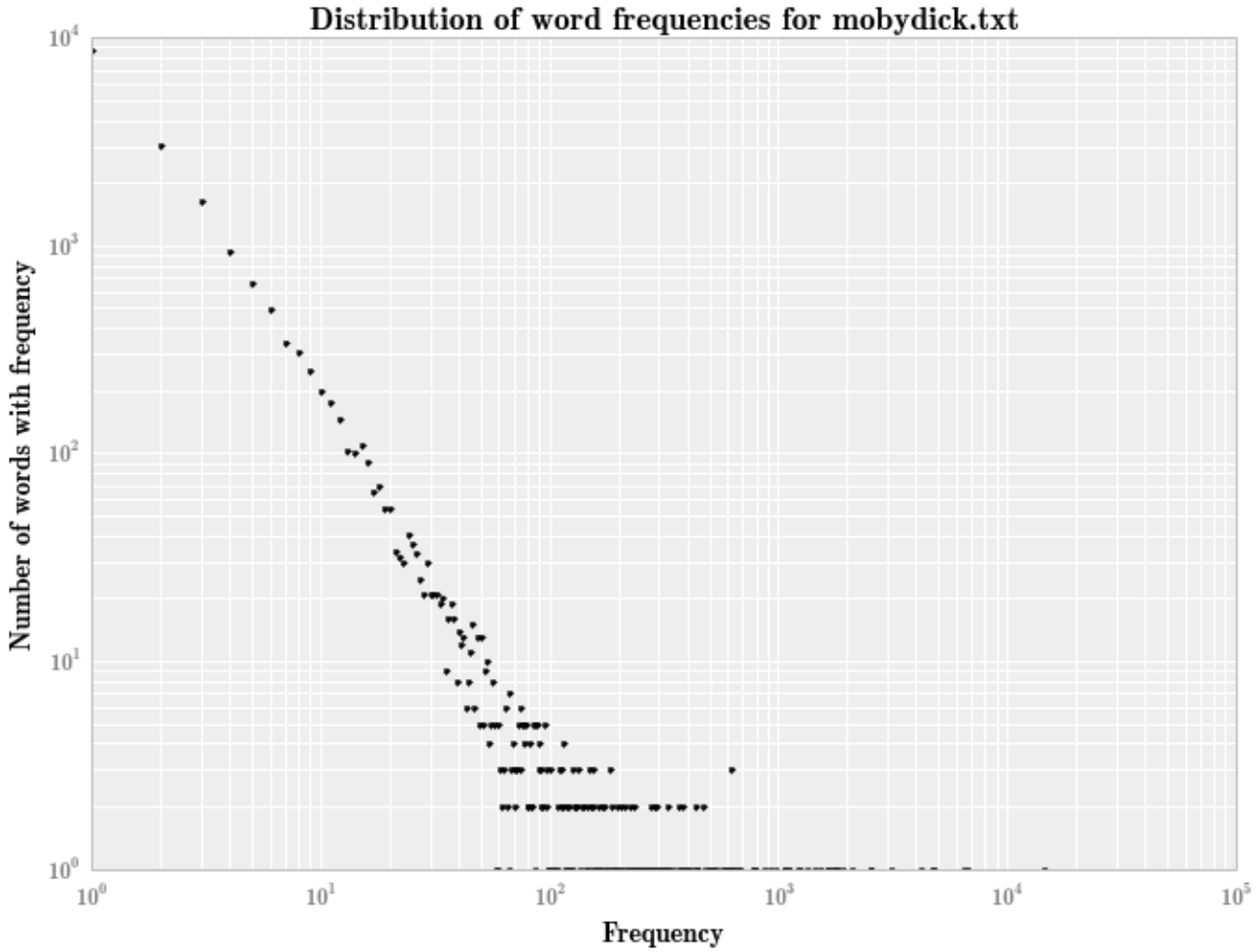
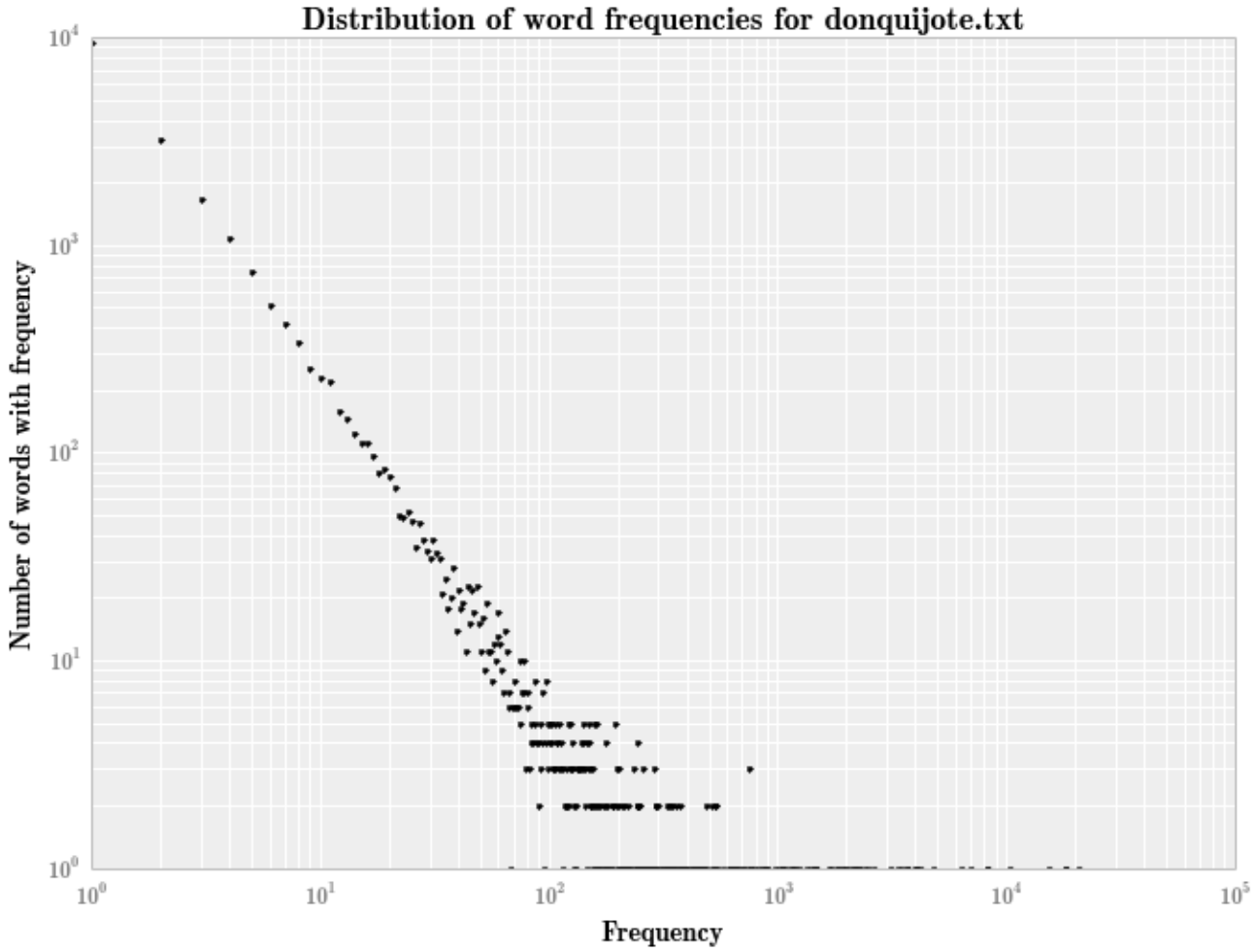
In [1]:
```
from collections import Counter
from ggplot import *
import math
```

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())
    x,y = zip(*(counter.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 pdf(x,a,xm):
    return exp(log(a-1)-log(xm)-a*(log(x)-log(xm)))

def KS_statistic(x,alpha,xmin):
    x_filt = filter(lambda i: i >= xmin, x)
    freqs = Counter(x_filt)
    total = float(len(x_filt))
    ctr = {}
    for f in freqs:
        ctr[f] = freqs[f]/total
    return max([abs(pdf(f,alpha,xmin) - ctr[f]) for f in ctr])
```

In [7]:
```
def mle_alpha(x,xmin):
```

In [7]:
```
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 [8]:
```
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(1,50):
            alpha = mle_alpha(freqs,xmin)
            ks = KS_statistic(freqs,alpha,xmin)
            print "xmin = %d, alpha = %g, KS=%g" % (xmin,alpha,ks)
```

In [9]:
```
find_xmin("mobydick.txt")
```

```
xmin = 1, alpha = 2.15047, KS=0.680908
xmin = 2, alpha = 2.05763, KS=0.219511
xmin = 3, alpha = 2.03793, KS=0.104255
xmin = 4, alpha = 2.01739, KS=0.0718568
xmin = 5, alpha = 2.02127, KS=0.0482799
xmin = 6, alpha = 2.0227, KS=0.0317109
xmin = 7, alpha = 2.0192, KS=0.0340513
xmin = 8, alpha = 2.03006, KS=0.0157586
xmin = 9, alpha = 2.02384, KS=0.0115836
xmin = 10, alpha = 2.01437, KS=0.0123542
xmin = 11, alpha = 2.00957, KS=0.0133526
xmin = 12, alpha = 1.99772, KS=0.0133924
xmin = 13, alpha = 1.98742, KS=0.0133019
xmin = 14, alpha = 1.99372, KS=0.00957057
xmin = 15, alpha = 1.99142, KS=0.0103967
xmin = 16, alpha = 1.97308, KS=0.00972808
xmin = 17, alpha = 1.95843, KS=0.009517
xmin = 18, alpha = 1.95636, KS=0.00996281
xmin = 19, alpha = 1.94448, KS=0.00969741
xmin = 20, alpha = 1.94097, KS=0.0099474
xmin = 21, alpha = 1.93246, KS=0.00974414
xmin = 22, alpha = 1.93948, KS=0.00891264
xmin = 23, alpha = 1.9459, KS=0.0083392
xmin = 24, alpha = 1.95195, KS=0.0088222
xmin = 25, alpha = 1.94279, KS=0.00895109
xmin = 26, alpha = 1.9345, KS=0.00907184
```

```
xmin = 27, alpha = 1.92767, KS=0.00920605
xmin = 28, alpha = 1.928, KS=0.00953894
xmin = 29, alpha = 1.93146, KS=0.00997438
xmin = 30, alpha = 1.92096, KS=0.00993668
xmin = 31, alpha = 1.92021, KS=0.0102195
xmin = 32, alpha = 1.91764, KS=0.0104291
xmin = 33, alpha = 1.91323, KS=0.0105539
xmin = 34, alpha = 1.90997, KS=0.0107119
xmin = 35, alpha = 1.90348, KS=0.0107131
xmin = 36, alpha = 1.91272, KS=0.00968958
xmin = 37, alpha = 1.90998, KS=0.0100431
xmin = 38, alpha = 1.90083, KS=0.010687
xmin = 39, alpha = 1.89506, KS=0.0112
xmin = 40, alpha = 1.90183, KS=0.0111258
xmin = 41, alpha = 1.89737, KS=0.0115887
xmin = 42, alpha = 1.89533, KS=0.0119393
xmin = 43, alpha = 1.89041, KS=0.0124462
xmin = 44, alpha = 1.89764, KS=0.0123053
xmin = 45, alpha = 1.90064, KS=0.0123836
xmin = 46, alpha = 1.89713, KS=0.0128248
xmin = 47, alpha = 1.88464, KS=0.0122028
xmin = 48, alpha = 1.88916, KS=0.0121961
xmin = 49, alpha = 1.8787, KS=0.0129677
```

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

```
xmin = 1, alpha = 2.01752, KS=0.563174
xmin = 2, alpha = 1.90257, KS=0.166209
xmin = 3, alpha = 1.87391, KS=0.0858485
xmin = 4, alpha = 1.86769, KS=0.0502038
xmin = 5, alpha = 1.86212, KS=0.03434
xmin = 6, alpha = 1.85952, KS=0.0319277
xmin = 7, alpha = 1.86578, KS=0.0226407
xmin = 8, alpha = 1.86857, KS=0.0178504
xmin = 9, alpha = 1.86923, KS=0.0202742
xmin = 10, alpha = 1.8771, KS=0.0131354
xmin = 11, alpha = 1.87975, KS=0.0126821
xmin = 12, alpha = 1.87398, KS=0.0130122
xmin = 13, alpha = 1.87955, KS=0.00924054
xmin = 14, alpha = 1.88191, KS=0.0102726
xmin = 15, alpha = 1.88651, KS=0.00883158
xmin = 16, alpha = 1.89033, KS=0.00684893
```

```
xmin = 17, alpha = 1.8891, KS=0.00719676
xmin = 18, alpha = 1.88908, KS=0.00760368
xmin = 19, alpha = 1.89292, KS=0.00825508
xmin = 20, alpha = 1.89135, KS=0.00856281
xmin = 21, alpha = 1.88897, KS=0.00879493
xmin = 22, alpha = 1.88796, KS=0.00911173
xmin = 23, alpha = 1.89441, KS=0.00824696
xmin = 24, alpha = 1.89928, KS=0.00886789
xmin = 25, alpha = 1.90003, KS=0.00926116
xmin = 26, alpha = 1.90163, KS=0.00971364
xmin = 27, alpha = 1.90929, KS=0.0064473
xmin = 28, alpha = 1.9077, KS=0.00663066
xmin = 29, alpha = 1.90973, KS=0.00690347
xmin = 30, alpha = 1.91303, KS=0.0072152
xmin = 31, alpha = 1.91715, KS=0.00755739
xmin = 32, alpha = 1.91433, KS=0.00770073
xmin = 33, alpha = 1.91375, KS=0.00790684
xmin = 34, alpha = 1.91329, KS=0.00811495
xmin = 35, alpha = 1.92, KS=0.00856901
xmin = 36, alpha = 1.92236, KS=0.00888291
xmin = 37, alpha = 1.92995, KS=0.00939757
xmin = 38, alpha = 1.93501, KS=0.00983357
xmin = 39, alpha = 1.93166, KS=0.00994448
xmin = 40, alpha = 1.94054, KS=0.00932365
xmin = 41, alpha = 1.94101, KS=0.00956121
xmin = 42, alpha = 1.94453, KS=0.00990414
xmin = 43, alpha = 1.94625, KS=0.0101906
xmin = 44, alpha = 1.95562, KS=0.00818367
xmin = 45, alpha = 1.95165, KS=0.00833321
xmin = 46, alpha = 1.95539, KS=0.00853768
xmin = 47, alpha = 1.9506, KS=0.00867737
xmin = 48, alpha = 1.95046, KS=0.00885182
xmin = 49, alpha = 1.94249, KS=0.00895795
```

In [85]:
```
max([1,2,3])
```

Out[85]: 3

In [93]:
```
log(10)
```

Out[93]: 2.3025850929940459

In [ ]: