HW6

1. 18788
2. the 8651  
   to 4663  
   a 3673  
   in 3521  
   and 3446  
   of 2792  
   for 1771  
   is 1470  
   on 1432  
   was 1421  
   he 1244  
   with 1166  
   have 1152  
   at 1137  
   I 1126  
   his1 111  
   that 1060  
   has 965  
   be 950  
   but 931
3. graphs are on EXCEL, THERE ARE 2 seeds 1 for just the data and the other is the log data

the shape of the graph with raw data is fairly simple with it descending as the rank gets lowered, however, without a large graph it is hard to look at the trends and statistics of the graph since it follows a raw form.

With the graph with the log of data, the plot points have more range and is easier to see the trends of the graph rather than looking at the graph is the raw data. The log graph shows more variety and trends by graphing the log of data rather than the raw data itself.

1. contract = 0.5473102011324538  
     
   Ronaldo  
   contract  
   United  
   World.  
   tomorrow.  
   first-team.  
   Trafford.  
   five-year-deal,  
   knows,”  
   club.  
   knows  
   “Nobody  
   resolved  
   agreeing  
   star,  
   sides.”  
   renew  
   future.”  
   News  
   Portugal
2. bag of words = 0.9988489496852295  
   tf-idf = 0.9886336595935993

the values are different because for bag of words, it corresponds to the word count / total word count of that document, whereas the tf-idf value corresponds to how important a word in that document is for the entire corpus, therefore since the two values correspond to different aspects of the corpus and the documents, they are different.

1. The first major issue with my computer words is that I didn’t eliminate any periods, commas, or quotations that would make the words more unique when I. and I is actually the same word.

The second major issue with my computer words is that I did not eliminate any articles and useless words such as, the, a, and, it, and other irrelevant particles that will just add up the word count and other data forms in the corpus. Only accounting the specific words that matter should allow us to have more exact data without any irrelevant elements inside it.