A Needle in a Data Haystack

- Ex. 3 -

Motti Gold – 301260105

1)

a. This rule prevent words with SS suffix to remain the same without stemming one S. e.g. we want the word "Less" to remain the same and not changed to "Les"

b. PONI, TI, CIRCU

c. If we add the rule "$ES \rightarrow$ " we get CARESS. If we add it at the beginning we can even save the "$IES \rightarrow I$" rule.

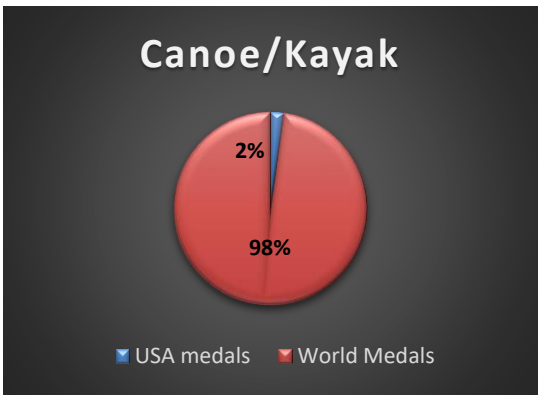
d. False. When we apply stemming we may actually reduce the size of the vocabulary since several word may combine to one single word. (e.g. plural and singular combine to one).

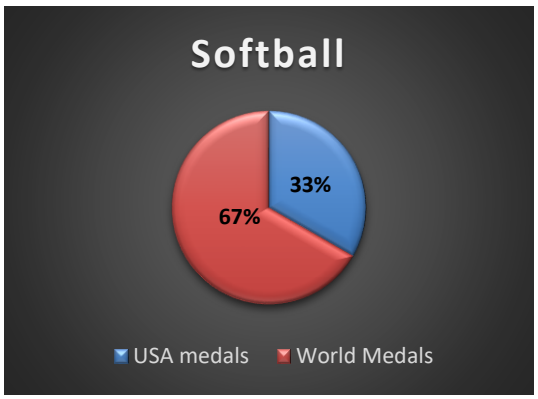
e. False. When we apply stemming we change the search to a shorter word, which makes it more general. So instead of the results that matches the original query, we get more results (adding more False-Positive), which lowers the precision.

f. True. Since relevance doesn't effect by stemming, and the retrieval may only increase by stemming (as explained in the previous article), we get that recall $\left(=\frac{relevance \cap retrieved}{relevance}\right)$ can only get higher by stemming and never lower.

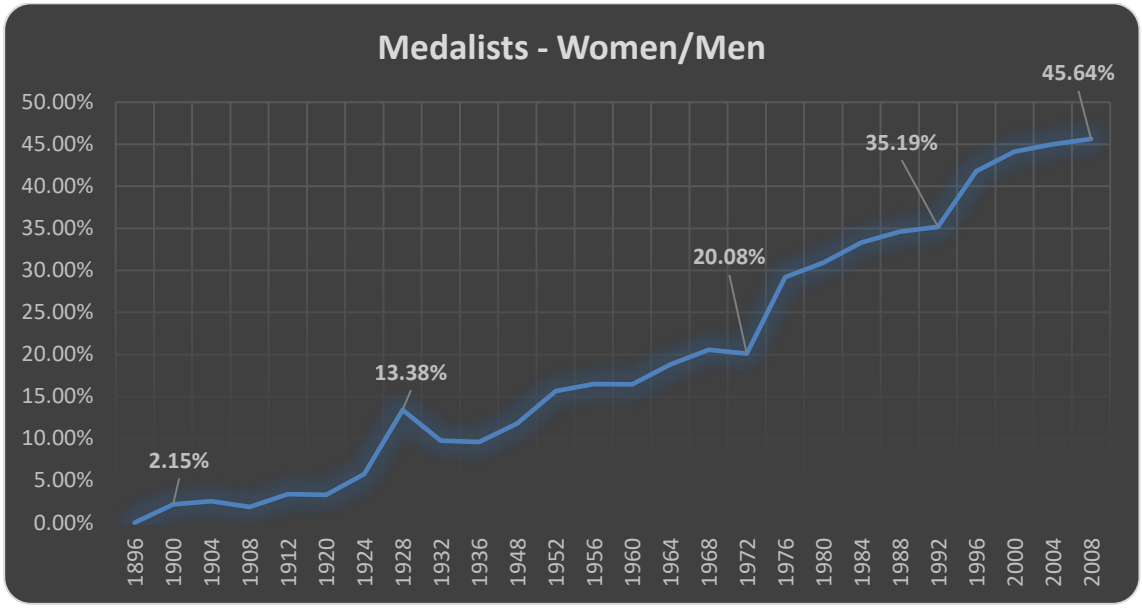
2)

a. **Highlight contrasts** – In which sport field USA is the best and in which the worst. Only fields with more than 100 medals (all over the world) are counted and we measure the ratio of medals the USA players won comparing to any other country. We can see that the **best field is Softball** and the **worst is Kayaking**:

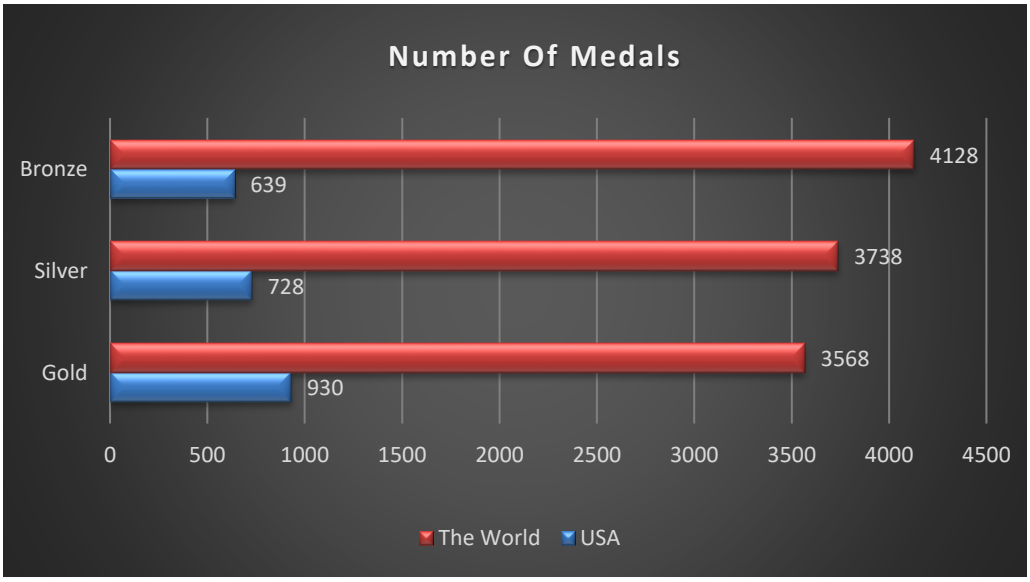




b. **Narrate change over time** – How the ratio $\frac{women}{men}$ winning medals change over time (1896-2008). We can see clearly how the ratio of women winning medals gets bigger and bigger over time**:**



c. **Start big and drill down** - How many Olympics medals (gold, silver and bronze) won around the world since 1896 till 2008? How many of them won by the USA players?

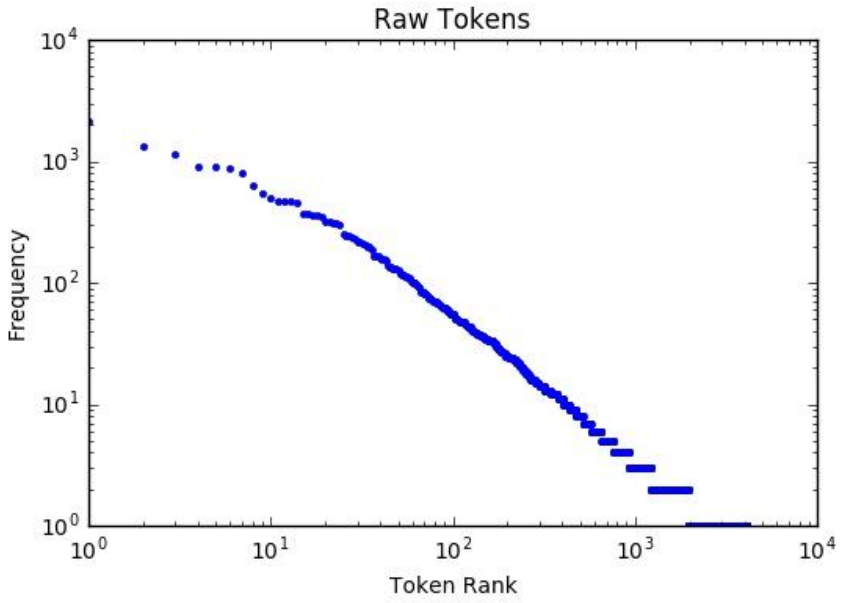


We can see that the sum of the each type of medal is not equal, probably caused by several players sharing the same medal, or by a problem with the data

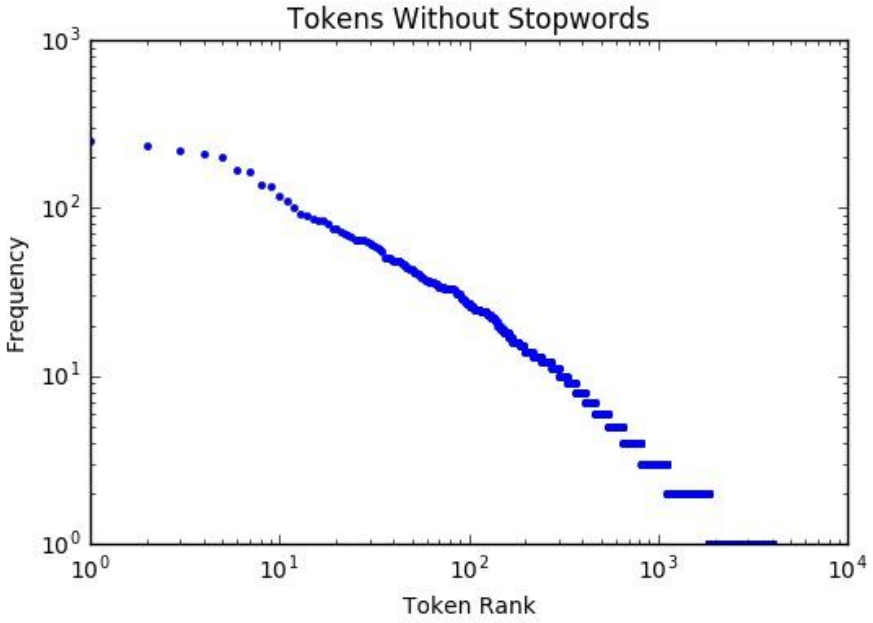
3)

a. "**Peter Pan**", By J. M. Barrie.

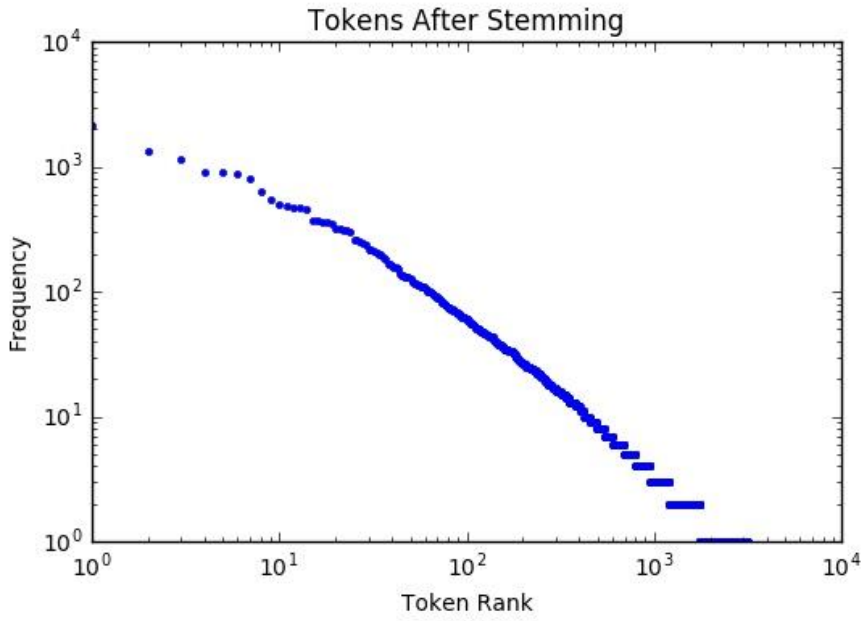
b. Top-20: the, and, to, a, was, he, of, in, that, had, she, they, it, his, but, not, her, for, you, is.



c. Top-20: I, Peter, said, would, Wendy, one, He, could, The, It, She, They, Darling, little, like, see, Hook, time, John, children.



d. Top-20: the, and, to, a, wa, he, of, in, that, had, it, she, they, hi, but, not, her, for, you, is.



e. Top-10: IN, DT, PRP, VBD, NN, RB, CC, JJ, VB, NNP.