**Typical errors of stemming:**

Overstemming: Happens when too much is removed. For example, ‘wander’ becomes ‘wand’; ‘news’ becomes ‘new’; or ‘universal’, ‘universe’, ‘universities’, and ‘university’ are all reduced to ‘univers’. A better result would be ‘univers’ for the first two and ‘universi’ for the last two.

Understemming: Happens when words are from the same root but are not seen that way. For example, ‘knavish’ remains as ‘knavish’; ‘data’ and ‘datum’ become ‘dat’ and ‘datu’ although they’re from the same root.

Misstemming: Usually not a problem unless it leads for false conflations. For example, ‘relativity’ becomes ‘relative’.

**Performance Metrics for Stemming:**

**Index Compression Factor (ICF):**

The index compression factor represents a percent by which a collection of distinct words is reduced by stemming. The large number of words stemmed will give the greater strength of the stemmer. This is calculated as :

𝐼𝐶𝐹 = ( 𝑁 − 𝑆 /𝑆) ∗ 100

where

N- Number of distinct words before stemming S- Number of distinct stems after stemming.

**Word Stemmed Factor (WSF):**

It is the percentage of words that have been stemmed by the stemming process out of the total words in a sample. Improved the number of words stemmed, greater the strength of the stemmer. Minimum threshold for this factor should be 50%.

𝑊𝑆𝐹 = (𝑊𝑆/𝑇𝑊) ∗ 100

**Correctly Stemmed Words Factor (CSWF):**

It is the percentage of the number of words stemmed. Higher the percentage of this factor, higher will be the accuracy of the stemmer. Minimum threshold for his factors should be 50%.

𝐶𝑆𝑊𝐹 = 𝐶𝑆𝑊 ∗ 100

**Average Words Conflation Factor (AWCF):**

This indicates the average number if variant words of different conflation group that are stemmed correctly to the root words. To calculate AWCF, first to find out the number of unique words after conflation as:

𝑵𝑾𝑪 = 𝑺 − 𝑪𝑾

where,

CW- Number of correct words not stemmed. Thus Word conflation Factor is obtained as:

𝐴𝑊𝐶𝐹 = 𝐶𝑆𝑊 − 𝑁𝑊𝐶 ∗ 100 𝐶𝑆𝑊

Greater the percentage of AWCF, greater will be the accuracy of the stemmer.