NLP PROJECT EVALUATION METRIC

**MAANAV BHAVSAR | 19070122098**

**ISHA PENDHARKAR | 19070122069**

What is ROUGE?

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It is essentially of a set of metrics for evaluating automatic summarization of texts as well as machine translation. It works by comparing an **automatically produced summary** or **translation** against a set of **reference summaries** (typically human-produced).

Precision and Recall in the Context of ROUGE

Simplistically put, recall in the context of ROUGE simply means how much of the *reference summary*is the *system summary* recovering or capturing? If we are just considering the individual words, it can be computed as:



A machine generated summary (system summary) can be extremely long, capturing all words in the reference summary. But, much of the words in the system summary may be useless, making the summary unnecessarily verbose. This is where precision comes into play. In terms of precision, what you are essentially measuring is, how much of the system summary was in fact relevant or needed? Precision is measured as:



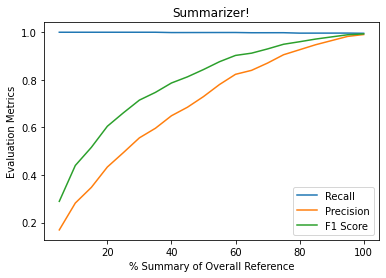
### So What is ROUGE-1, ROUGE-2 & ROUGE-L ?

ROUGE-1 refers to overlap of ***unigrams***between the system summary and reference summary.

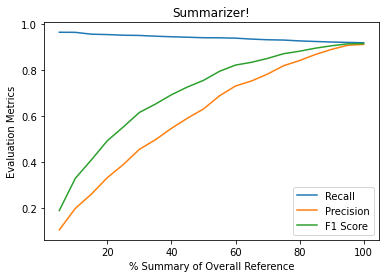
ROUGE-2 refers to the overlap of ***bigrams***between the system and reference summaries.

ROUGE-L measures ***longest matching sequence*** of words using LCS. An advantage of using LCS is that it does not require consecutive matches but in sequence matches that reflect sentence level word order. Since it automatically includes longest in-sequence common n-grams, you don’t need a predefined n-gram length.

**Rouge-1 Evaluation Metrics**



### ****Rouge-2 Evaluation Metrics****



### ****Rouge-L Evaluation Metrics****

