Alberto Andrés Valdés González.
Degree: Mathematical Engineer.
Work position: Data Scientist.
Mail: anvaldes@uc.cl/alberto.valdes.gonzalez.96@gmail.com
Location: Santiago, Chile.

N-gram Score

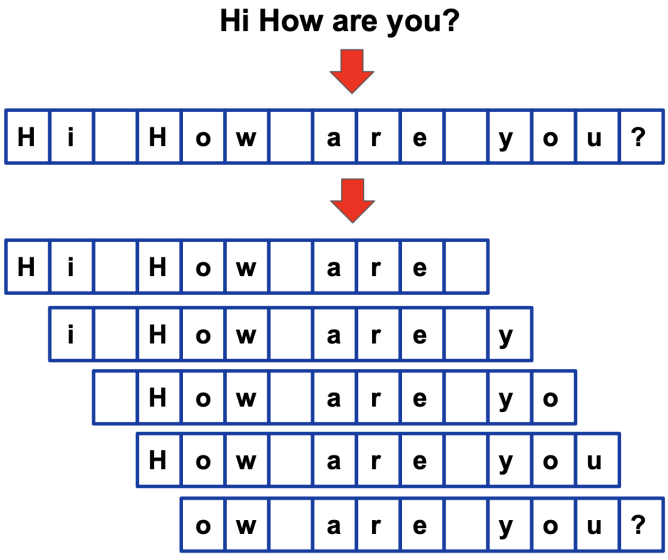
When we are working in NLP an important issue is measure the similarity between two sentences. There is many metrics to achieve that, however, we are going to aboard the $n-$gram score metric.

Example:

We have the next two sentences:

Sentence 1: Hi! How are you? I hope you find well.
Sentence 2: H How are you I hope you find well.

We going to divide these two sentences in fractions of n characters. An illustration of that is:



Here we used $n = 11$.

We do this split for the specific n and the results are $s1$ and $s2$.

$$\boxed{\text{n gram score } 1 = \frac{\text{Number of elements of s1 which are in s2}}{\text{Number of elements of s1}}}$$

You can see this as the percentage of the sentence 1 which are in the sentence 2.
($Sentence1 \subseteq Sentence2$).

If we want to measure of the two sentences are equal we can define:

$$\boxed{\text{n gram score} = \frac{\text{n gram score } 1 + \text{n gram score } 2}{2}}$$
