28.6 Uni-gram, Bi-gram and n-grams

Let us assume we have two reviews 'r1' and 'r2'.

r_1: this pasta is very tasty and affordable

r_2: this pasta is not tasty and is affordable

After removing the underlined stopwords, the reviews become

r_1: pasta tasty affordable

r_2: pasta tasty affordable

Now both 'r_1' an 'r_2' have become exactly the same and the distance between their corresponding vectors 'v_1' and 'v_2' is zero, as they both are the same. But here, now we are forced to conclude that both the reviews are similar. But these two reviews, in their original form(ie., before removing the stopwords) are quite opposite. So we shouldn't go with false conclusions. In order to solve this type of problem, we need bi-grams, tri-grams, n-grams, etc.

Let us consider the reviews in their original form again.

r_1: this pasta is very tasty and affordable

r_2: this pasta is not tasty and is affordable

Uni-grams

We have a d-dimensional vector for all the words in the corpus. The presence of each word is indicated by non zero values in each dimension.

	this	pasta	is	very	tasty	and	affordable	not
r1	1	1	1	1	1	1	1	0
r2	1	1	1	0	1	1	1	1

Bi-grams

Here we create a vector for each review, with a pair of words as each dimension and these words will be the consecutive words of both the reviews.

	this pasta	pasta is	is very	very tasty	tasty and	and affordable	is not	not tasty	and is	is affordable
r1	1	1	1	1	1	1	0	0	0	0
r2	1	1	0	0	1	0	1	1	1	1

Tri-grams

Each dimension is obtained by taking 3 consecutive words at a time.

	this pasta is	pasta is very	is very tasty	very tasty and	tasty and affordable	pasta is not	is not tasty	not tasty and	tasty and is	and is affordable
r1	1	1	1	1	1	0	0	0	0	0
r2	1	0	0	0	0	1	1	1	1	1