Conversion Techniques in NLP (Natural Language Processing)

- Megh Desai

Natural Language Processing is a field of Artificial Intelligence used to analyse human input and generate useful patterns and insight from it. It highly relies on distributional semantics that is the way in which words appear in the same sentence. Each word is assigned a numeric representation and the entire sentences are stored in the form of vectors. These are then fed into the machine for training and useful outputs are gathered. There are various processes to perform the above action some of which are

1. Bag of Words (BOW) model

The bag of words model works on the principle of how many times a word occurs in a sentence. The first step is to turn the sentence to lower case so that the uppercase and lowercase versions of the same word are not treated differently. There are some exceptions like country names example US (United States) if turned to lowercase may be interpreted as us (pronoun) which we need to be careful about. Then we use stop keywords to remove non essential words. Let us understand what happens to the following sentences after using stop keywords.

i) The King is kind - - > King kind

ii) The Queen is kind - - > Queen kind

iii) The King and Queen are kind - - > King Queen kind

Next we turn these to vectors using a table where f1, f2 and f3 are independent features. If we are using Binary BOW, we put one if the word exists and zero if it doesn’t. If we use normal BOW, then it removes the total frequency of that word. The below example uses Binary BOW

	f1	f2	f3	Output Vector
sentence	King	Queen	kind	
i)	1	0	1	1 f1 + 1 f3
ii)	0	1	1	1 f2 + 1 f3
iii)	1	1	1	1 f1 + 1 f2 + 1 f3

These vectors are then fed to the ML models for analysis. The disadvantage of the above method is that it is difficult to enable sentiment analysis like in the first sentence both King and kind have the number 1 before their feature although kind should get a higher priority as it describes the kings nature.

2. Term Frequency and Inverse Document Frequency (TF-IDF)

This is another method used to determine the level of importance of a word in a given sentence. A high TF – IDF score means that the word is more informative compared to that with a lower TF – IDF score. There are three tables, one for TF, one for IDF and one for TF X IDF. Let us understand with the three sentences used above

TF = Number of representations of word in sentence / Number of words in sentence			
	i)	ii)	iii)
King	1/2	0	1/3
Queen	0	1/2	1/3
kind	1/2	1/2	1/3

IDF = log (Number of sentences / Number of sentences containing the word)	
Words	IDF
King	log(3/2)
Queen	log(3/2)
kind	log(3/3) = 0

TF X IDF				
	f1	f2	f3	Output vector
	King	Queen	kind	
i)	1/2 * log(3/2)	0	0	1/2log(3/2) * f1
ii)	0	1/2 * log(3/2)	0	1/2log(3/2) * f2
iii)	1/3 * log(3/2)	1/3 * log(3/2)	0	1/3log(3/2) * f1 + 1/3 * log(3/2)

Hence from the above vectors we are able to get a better understanding as compared to BOW.

3. N – grams
It is similar to the BOW model but instead of using single words it uses a sequence of words. Let us assume two sentences
a) The pasta was delicious - - > pasta delicious
b) The pasta was not delicious - - > pasta not delicious
if we use only uni-grams (1,1) then we get the following table

	pasta	not	delicious
a)	1	0	1
b)	1	1	1

The above vectors are almost similar and hence to further differentiate we use uni-grams and bi-grams (1,2)

	pasta	not	delicious	pasta delicious	pasta not	not delicious
a)	1	0	1	1	0	0
b)	1	1	1	0	1	1

Hence as we go on further that is tri-grams.. etc we get more different vectors and a better understanding. The ML model is also able to differentiate better between the vectors.

Despite the usefulness of the above methods recently more advanced methods like word embeddings , hash vectorizing etc are gaining importance.