Experiment 7

N-Grams Smoothing

Aim:

Write a program to create, append, and remove lists in python.

Theory:

One major problem with standard N-gram models is that they must be trained from some corpus, and because any particular training corpus is finite, some perfectly acceptable N-grams are bound to be missing from it. We can see that bigram matrix for any given training corpus is sparse. There are large number of cases with zero probability bigrams and that should really have some non-zero probability. This method tend to underestimate the probability of strings that happen not to have occurred nearby in their training corpus.

There are some techniques that can be used for assigning a non-zero probability to these 'zero probability bigrams'. This task of reevaluating some of the zero-probability and low-probability N-grams, and assigning them non-zero values, is called smoothing.

WHY IS SMOOTHING SO IMPORTANT?

● A key problem in N-gram modeling is the inherent data sparseness.

● For example, in several million words of English text, more than 50% of the trigrams occur only once; 80% of the trigrams occur less than five times (see SWB data also).

● Higher order N-gram models tend to be domain or application specific. Smoothing provides a way of generating generalized language models.

● If an N-gram is never observed in the training data, can it occur in the evaluation data set?

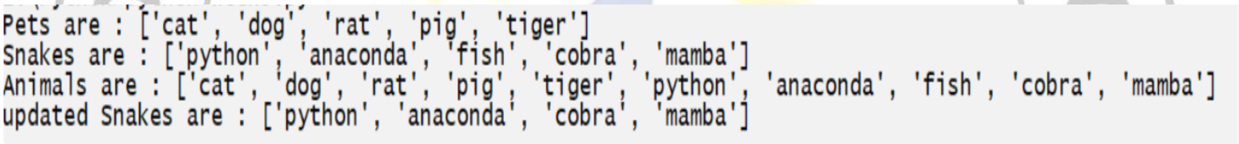
● Solution: Smoothing is the process of flattening a probability distribution implied by a language model so that all reasonable word sequences can occur with some probability. This often involves broadening the distribution by redistributing weight from high probability regions to zero probability regions

Zagdu Singh Charitable Trust's (Regd.) THAKUR COLLEGE OF ENGINEERING & TECHNOLOGY tcet Estd. in 2001 ISO 9001 : 2015 Certified NBA and NAAC Accredited

Code:

```
pets = ['cat', 'dog', 'rat', 'pig', 'tiger']
snakes=['python','anaconda','fish','cobra','mamba']
print("Pets are :",pets)
print("Snakes are :",snakes)
animals=pets+snakes
print("Animals are :",animals)
snakes.remove("fish")
print("updated Snakes are :",snakes)
```

Output:



```
Pets are : ['cat', 'dog', 'rat', 'pig', 'tiger']
Snakes are : ['python', 'anaconda', 'fish', 'cobra', 'mamba']
Animals are : ['cat', 'dog', 'rat', 'pig', 'tiger', 'python', 'anaconda', 'fish', 'cobra', 'mamba']
updated Snakes are : ['python', 'anaconda', 'cobra', 'mamba']
```

Conclusion:

Thus, in this experiment we have studied about N-gram models. We have also studied what is smoothing and why is it so important. We have implemented the code for the same and got relevant output.

Lagadu Singh Charitable Trust's (Regd.)
THAKUR COLLEGE OF ENGINEERING & TECHNOLOGY
Estd. 2001
ISO 9001:2015 Certified
NBA and NAAC Accredited