Practical - 5 Virtual Lab

18BCE080 ISHAN TEWARI

Part - 1: Ngrams

What is a N Gram?

- *Definition:*
    - An n-gram is a contiguous sequence of n items from a given sample of text or speech.
- *Calculating Probability:*
    - The probability of any word n, can be calculated as
    - P(w(1), w(2),...,w(n)) = P(w(2)|w(1))*P(w(3)|w(1),w(2))*...*P(w(n)| w(1),w(2)...w(n-1))
    - Probability of a given sentence can be calculated by multiplying n gram probability of each word.

Bi-Grams

- *Definition:*
    - An n-gram is a contiguous sequence of n items from a given sample of text or speech.
    - So, a Bi-gram is a contiguous sequence of 2 items from a given sample of text or speech.
    - The assumption made here is “the probability of a word depends on the probability of the previous word only”. This assumption is called Markov assumption and the model is known as Markov model or bigram model.
- *Calculating Probability:*
    - $P(w(n)|w(n-1)) = P(w(n-1),w(n)) / P(w(n-1))$

Example:

- Let sentence = (eos) Can I sit near you (eos) You can sit (eos) Sit near him (eos) I can sit you (eos)
- Probability of

  “(eos) i sit you (eos)”

  = 0.2 * 0.5 * 0.25 * 0.67

  = 0.0167

	can	eos	him	i	near	sit	you
can	0.00	0.00	0.0	0.33	0.0	0.67	0.00
eos	0.20	0.00	0.0	0.20	0.0	0.20	0.20
him	0.00	1.00	0.0	0.00	0.0	0.00	0.00
i	0.50	0.00	0.0	0.00	0.0	0.50	0.00
near	0.00	0.00	0.5	0.00	0.0	0.00	0.50
sit	0.00	0.25	0.0	0.00	0.5	0.00	0.25
you	0.33	0.67	0.0	0.00	0.0	0.00	0.00

Part - 2: Bigram Smoothing

Need for Bigram Smoothing

- *Definition:*
    - 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.
- *Calculating Probability:*
    - $P(w(n) \mid w(n-1)) = \frac{C(w(n-1)w(n)) + 1}{C(w(n-1)) + V}$

Example:

- Let sentence = (eos) Can I sit near you (eos) You can sit (eos) Sit near him (eos) I can sit you (eos)

	eos	can	i	sit	near	you	him
eos	0.0	300.0	300.0	300.0	0.0	300.0	0.0
can	0.0	0.0	300.0	600.0	0.0	0.0	0.0
i	0.0	300.0	0.0	300.0	0.0	0.0	0.0
sit	300.0	0.0	0.0	0.0	600.0	300.0	0.0
near	0.0	0.0	0.0	0.0	0.0	300.0	300.0
you	600.0	300.0	0.0	0.0	0.0	0.0	0.0
him	300.0	0.0	0.0	0.0	0.0	0.0	0.0

Before Smoothing

	eos	can	i	sit	near	you	him
eos	0.0002	0.0527	0.0527	0.0527	0.0002	0.0527	0.0002
can	0.0002	0.0002	0.0527	0.1053	0.0002	0.0002	0.0002
i	0.0002	0.0527	0.0002	0.0527	0.0002	0.0002	0.0002
sit	0.0527	0.0002	0.0002	0.0002	0.1053	0.0527	0.0002
near	0.0002	0.0002	0.0002	0.0002	0.0002	0.0527	0.0527
you	0.1053	0.0527	0.0002	0.0002	0.0002	0.0002	0.0002
him	0.0527	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002

After Smoothing

Thank You :)