Noise Distribution

When we do *negative sampling* for a given input word, we say that we sample, at random, k incorrect context words to use in our loss function. But what exactly do we mean by *random*, in this instance?

The most simple way would be to pick these words at random would be to pick them *uniformly* at random. I.e. for a given word w, the probability of picking that word would be

$$\text{Uniform}(w) = \frac{1}{V}$$

where V is the size of the vocabulary.

Perhaps a better way would be to pick these words *randomly* in proportion to how often they appear in the corpus. This is called the *Unigram distribution*, for which the probability of picking a word w would be

$$\text{Unigram}(w) = \frac{\text{the number of occurrences of } w \text{ in the corpus}}{\text{the length of the corpus}}.$$

The distribution which the authors of find to work well, empirically, is a power of the *Unigram distribution*, namely

$$\frac{\text{Unigram}(w)^{3/4}}{\sum_{i=1}^{V} \text{Unigram}(w_i)^{3/4}}.$$

Continue
Ask question