Word2vec

1. 负采样思想是什么？解决了什么问题？

softmax 是一个巨大的词表，要计算分母，而分母是词表级别的运算。为了优化分母，将多分类转化成二分类。

二分类就是正样本和负样本，以 skip-gram 为例，skip-gram 就是中间词预测两边词，如果两边的词出现则认为是正样本，那么没出现的词就是负样本，负样本太多要进行采样，采样要根据词频进行随机采样。

2. fasttext 思想

利用了子词概念，子词的作用是丰富词的信息，对出现稀少的词有更好的表示，可以表示未出现在词表中的词。

将词分为：

- for example, $n = 3$, i.e., 3-grams:
  - **word**: *"where"*,
  - **sub-words**: *"wh"*, *"whe"*, *"her"*, *"ere"*, *"re"*

对于 fasttext 的词可以通过加和子词来表示。

fasttext 的词向量如何训练得到？

to compute an un-normalised score with center word $\mathbf{v}_c$, given a word w, g_w is the set of n-grams appearing in w, z_g is the representation to each individual n-gram

$$u(w, c) = \exp\left[\sum_{g \in g_w} z_g^\top \mathbf{v}_c\right]$$

不过是两个词向量之间的内积，使用子词与词做内积然后加和而已。

fasttext 有啥用？

1. 可以用来计算词与词之间的相似度，因为加入来子词概念，效果可能相对于 word2vec 更好一点。

2. 使用 fasttext 计算词向量时，对于没有出现的词 word2vec 是无法知道其词向量的，对于出现稀少的词，word2vec 无法很好的学习到他的词向量表示，而 fasttext 可以!!!!

为什么呢？因为一个词首先分解为子词，然后查询子词的词向量，然后做加和得到词的向量表示，然后就可以计算相似度啦。

3. 为啥 fasttext 可以做文本分类效果那么好？

fasttext 所有词的词向量的加和来表示一个句子的向量，然后直接在这个向量上做分类。

很难理解的是，很多小伙伴在中文上做分本分类上用了 fasttext，但是 fasttext 没办法给中文做子词啊。。也就是说使用 fasttext 做分类，跟在文本上先训练 word2vec 然后加和做文本分类是一样的。。