gensim库中word2vec的重要参数

```
class gensim.models.word2vec.Word2Vec(
    sentences=None, corpus_file=None,
    vector_size=100, alpha=0.025, window=5, min_count=5, max_vocab_size=None,
    sample=0.001, seed=1, workers=3, min_alpha=0.0001,
    sg=0, hs=0, negative=5, ns_exponent=0.75, cbow_mean=1,
    hashfxn=<built-in function hash>, epochs=5,
    null_word=0, trim_rule=None, sorted_vocab=1,
    batch_words=10000, compute_loss=False,
    callbacks=(), comment=None, max_final_vocab=None, shrink_windows=True)
```

vector_size:最终期望提取的向量维度大小。

window：滑动窗口大小（为1+周边词）

min_count:单词频数小于该值的单词不参与训练。

sg：1（skip-gram),0(CBOW)'

hs:1表示层级softmax，0表示负采样。

negative：如果选择负采样，样本数是多少。

alpha:优化器学习率

sample:给定训练数据的单词重采样频率（在最终的构造数据集的时候，会删除部分单词），主要针对文档中出现的高频单词（比如是、的、了、过等），降低高频单词的参与模型训练的概率。

ns_exponent:在给定负采样的时候，计算各个类别被抽取的概率公式中的超参数，主要是降低高频词\类别被抽中的概率，增加低频词\类别被抽中的概率

重采样频率的计算方式

word2vec训练集的构造过程是通过滑动窗口，来获取中心词和周边词，一个中心词+2n个周边词构建出一个样本，我们可以看作每滑动一次窗口构造出一个样本。而重采样是通过减少高频词为中心词的样本个数，来达到降低高频词参与模型训练的目的。

在进行重采样的时候，需要给定一个重采样频率sample（gensim中默认为0.001），当构造训练集的时候，对一个滑动窗口而言（即一条样本），我们可以找到它的中心词$word_i$，通过初始化一个0-1分布的随机数来确定采不采用该词为中心词的样本，具体的概率计算公式如下：

$$p = 1 - (\frac{sample}{freq\_ratio_{word_i}})^{\frac{1}{2}}$$

当p为负数时（即中心词在全文档中出现的频率小于设定的频率阈值），表示不需要构造随机数，直接将该条样本加入训练集。当p为正数时，表示该条样本有p的概率加入训练集。由上述公式可以看出该中心词在全文中出现的频率$freq_ratio_{word_i}$越大，该条样本被删除的概率越大。