fastText

笔记本:	自然语言处理		
创建时间:	2018/7/30 9:18	**更新时间:**	2018/7/30 9:30
作者:	beyourselfwb@163.com		
URL:	http://www.52nlp.cn/category/text-classification		

不同的是，CBOW的输入是目标单词的上下文，fastText的输入是多个单词及其n-gram特征，这些特征用来表示单个文档；CBOW的输入单词被onehot编码过，fastText的输入特征是被embedding过；CBOW的输出是目标词汇，fastText的输出是文档对应的类标。

值得注意的是，fastText在输入时，将单词的字符级别的n-gram向量作为额外的特征；在输出时，fastText采用了分层Softmax，大大降低了模型训练时间。

于是fastText的核心思想就是：将整篇文档的词及n-gram向量叠加平均得到文档向量，然后使用文档向量做softmax多分类。这中间涉及到两个技巧：字符级n-gram特征的引入以及分层Softmax分类。

但是fastText就不一样了，它是用单词的embedding叠加获得的文档向量，词向量的重要特点就是向量的距离可以用来衡量单词间的语义相似程度，于是，在fastText模型中，这两段文本的向量应

该是非常相似的，于是，它们很大概率会被分到同一个类中。

使用词embedding而非词本身作为特征，这是fastText效果好的一个原因；另一个原因就是字符级n-gram特征的引入对分类效果会有一些提升 。