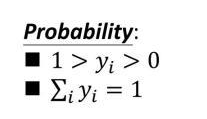
**语料库（corpus）：**指经科学取样和加工的大规模电子文本库。借助计算机分析工具，研究者可开展相关的语言理论及应用研究。

矩阵分解（Matrix Factorization）：就是将一个矩阵D分解为U和V的乘积，即对于一个特定的规模为m\*n的矩阵D，估计出规模分别为m\*k和n\*k的矩阵U和V，使得的值尽可能逼近矩阵D。

**Softmax：**把一些输入映射为0-1之间的实数，并且归一化保证和为1，因此多分类的概率之和也刚好为1。

**交叉熵（Cross Entropy）**：主要用于度量两个概率分布间的差异性信息。用该模型对文本识别的难度，或者从压缩的角度来看，每个词平均要用几个位来编码。

**TensorFlow：** 是一个基于[数据流编程](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%B5%81%E7%BC%96%E7%A8%8B/22735640)（dataflow programming）的符号数学系统，被广泛应用于各类[机器学习](https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0/217599)（machine learning）算法的编程实现，其前身是[谷歌](https://baike.baidu.com/item/%E8%B0%B7%E6%AD%8C/117920)的神经网络算法库DistBelief [1]  。

**ReLU 函数：线性整流函数**（Rectified Linear Unit, **ReLU**），又称**修正线性单元，**是一种[人工神经网络](https://baike.baidu.com/item/%E4%BA%BA%E5%B7%A5%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C)中常用的激活函数（activation function），通常指代以[斜坡函数](https://baike.baidu.com/item/%E6%96%9C%E5%9D%A1%E5%87%BD%E6%95%B0)及其变种为代表的非线性函数。

**逻辑回归（logistic regression）**进行分类的主要思想是：根据现有数据对分类边界线建立回归公式，以此进行分类。“回归”一词源于最佳拟合，表示要找到最佳拟合参数集。优点：计算代价不高，易于理解和实现；预测结果是界于0和1之间的值。缺点：容易欠拟合，分类精度可能不高；预测结果呈“S”型，概率变化很小，边际值太小，slope太小，而中间概率的变化很大，很敏感。 导致很多区间的变量变化对目标概率的影响没有区分度，不容易确定阀值。