注意：考虑到助教批改以及反馈等原因，这次及以后的作业要求用纸质版的方式提交，请相互转告！

Web Content Mining: Classification

1. kNN with inverted index

1) 在 kNN 算法中，对一个 document 分类可以搜索所有的 documents，并找出其中距离最近的 k 个。试分析该过程的时间复杂度。

2) 如果我们使用 document 的 tf.idf 向量的余弦作为相似度，请思考是否可以使用倒排索引来查询一个 document 的 k 近邻。如果可以，请讨论该方法适用于何种情况，描述详细过程，并分析该方法和 1 中方法的区别。

2. Derivative of the logistic

为了使用梯度下降方法求解 $\min_{\mathbf{w}} \mathrm{J}(\mathbf{w})$，需要计算 $\mathrm{J}(\mathbf{w})$ 的导数。已知

$$\mathrm{J}(\mathbf{w}) = -\left[\sum_{i=1}^{N} y_i \log\, \sigma(\mathbf{w}^T\mathbf{x_i}) + (1-y_i)\log\,(1-\sigma(\mathbf{w}^T\mathbf{x_i}))\right]$$

求 $\frac{\partial}{\partial w_j}\mathrm{J}(\mathbf{w})$。

提示：

$$\frac{\partial}{\partial z}\sigma(z) = \frac{\partial}{\partial z}\frac{1}{1+e^{-z}} = \underbrace{-\left(\frac{1}{1+e^{-z}}\right)^2}_{\partial\sigma/\partial(1+e^{-z})} \times \underbrace{-e^{-z}}_{\partial(1+e^{-z})/\partial z}$$

$$= \sigma^2(z)\left(\frac{1-\sigma(z)}{\sigma(z)}\right) = \sigma(z)(1-\sigma(z)).$$