Week1

此笔记简略总结词向量表示方法。

1. 用相近的词语去解释：如 WordNet 方法，但此方法是主观的，会错过词其他的意思，也会错失新词的意思等；

2. 用离散的信号表示单词：如 One-Hot，但此方法会生成大量的词库规模，不能表示相近含义，没有通用性等；

3. 用上下文之间关系表示：如 N-gram 方法，会出现 OOV 的问题（数据稀疏），会出现维度灾难问题（N 增大），只能建模到前 n-1 个词，无法表示一词多义等；

4. 用 Word Vector 表示：把所有词都投到同一个语义空间，相似词性的词相近，能表示一定特征。如用 NNLM(2003)模型去构建，但存在词向量是副产品，且前文信息有限，计算量过大等问题；

5. Word2Vec：主要用到 CBOW 和 Skip-gram 两种方法，CBOW 是通过周边单词预测中间词，而 Skip-gram 则是用中间词预测周边单词。CBOW 比较简单，时间快，学习更多同词义单词，Skip-gram 能能学习到更多语义。但此方法存在计算量大，没有学习全局语义等问题；

5.1 H-Softmax(Hierarchical Softmax)：Word2vec 针对 word2vec 计算量大的问题，采用分层 softmax，把一棵树转化成二叉树，让其更有效率；

5.2 Negative Sampling：在 Skip-gram 方法中引入负样本，把预测周边单词的模型转化为逻辑回归任务，预测其是否是相邻单词。