Text Summarization-1

1. 如何表示文字的意义

1.1. WordNet

WordNet 是一种基于认知语言学的词典。它不是光把单词以字母顺序排列，而且按照单词的意义组成一个 “单词的网络”。WordNet 是一个覆盖范围宽广的英语词汇语义网。名词，动词，形容词和副词各自被组织成一个同义词的网络，每个同义词集合都代表一个基本的语义概念，并且这些集合之间也由各种关系连接。名词网络的主干是蕴涵关系的层次(上位/下位关系)，它占据了关系中的将近80%。

缺点：它是利用单词的相似性构造的词典，无法区别相似词的细微区别；更新词表困难；词表的构建有较大的主观性等等。

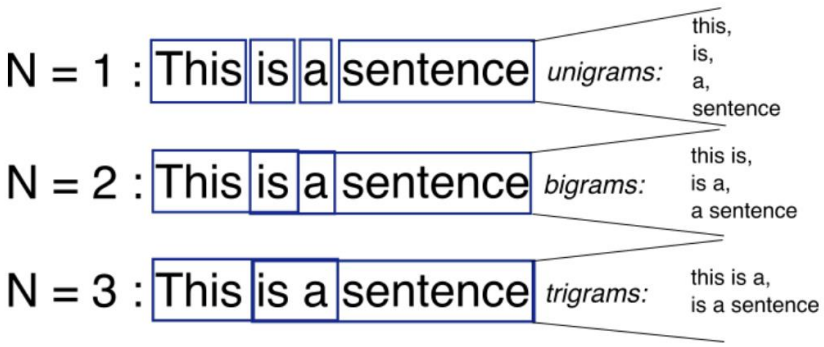
1.2. One-hot

根据预料中所有出现的词构建一个词表，词表大小为 V，词向量大小为 V，其中只有对应词的位置为 1，其余位置为 0。

$$motel = [0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0]$$
$$hotel = [0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0]$$

缺点：无法解决一词多义问题；无法计算单词的相似度；稀疏向量；词表过大，更新困难等。

1.3. N-gram



利用文章的前 N 个词去预测当前词，根据贝叶斯公式，

$$p(w_k|w_1^{k-1}) = \frac{p(w_1^k)}{p(w_1^{k-1})},$$

给定第 1 到 k-1 个词，第 k 个词的概率计算公式如上，假设一个词的出现概率只与它前面的 N 个词有关，则

$$p(w_k|w_1^{k-1}) \approx p(w_k|w_{k-n+1}^{k-1}),$$

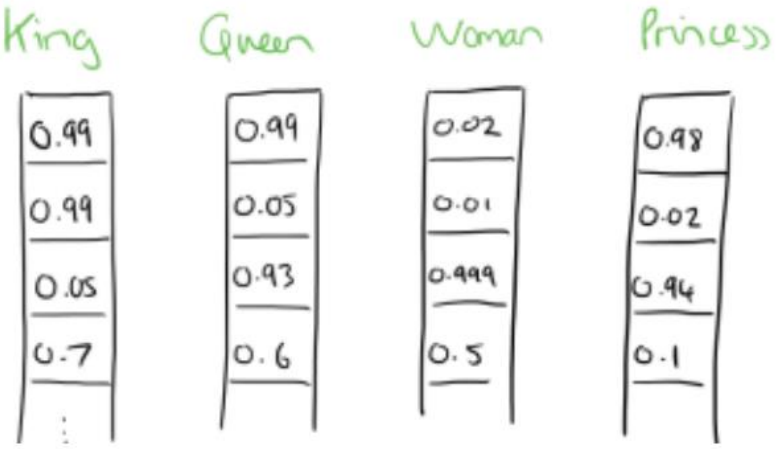
在预料足够大时，

$$p(w_k|w_1^{k-1}) \approx \frac{count(w_{k-n+1}^{k})}{count(w_{k-n+1}^{k-1})}.$$

缺点：数据稀疏，难免会出现 OOV 的问题；随着 n 的增大，参数空间呈指数增长（维度灾难）；缺少长期依赖，只能建模到前 n-1 个词；无法表示一词多义（语义鸿沟）

1.4. Word vector

利用一个固定长度的向量来表示一个词，不同的词之间，在词向量的不同维度上会有区别。例如：



在 2003 年提出的神经概率语言模型中，将预测词的前 n 个词拼接送入神经网络来进行训练，输出层接入 softmax 进行预测，网络中的参数就是词向量矩阵。

缺点：只有上文没有下文；无法解决一词多义；softmax 层计算量太大

2. Word2Vec 介绍

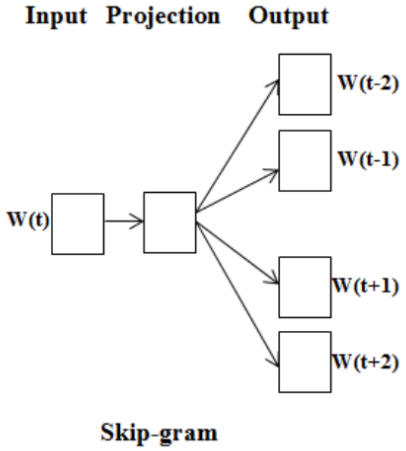
Word2Vec 于 2013 年提出，它包括两个训练模型：Skip-gram 和 CBOW，以及为了解决由于在输出层计算 softmax 计算量过大的问题，采用
两种近似训练方法：层级 softmax 和负采样。

2.1. Skip-gram

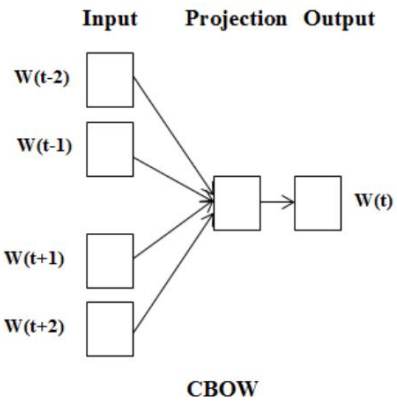
利用中心词预测背景词，根据窗口大小选择中心词两边的词作为背景词。模型通过中心词来预测背景词是哪些。容易学得语义关系(semantic relationshap)

2.2. CBOW

与 Skip-gram 相反，通过背景词来预测中心词。难度与 Skip-gram 相比较小，训练速度较快，容易学得句法关系(syntactic relationship)



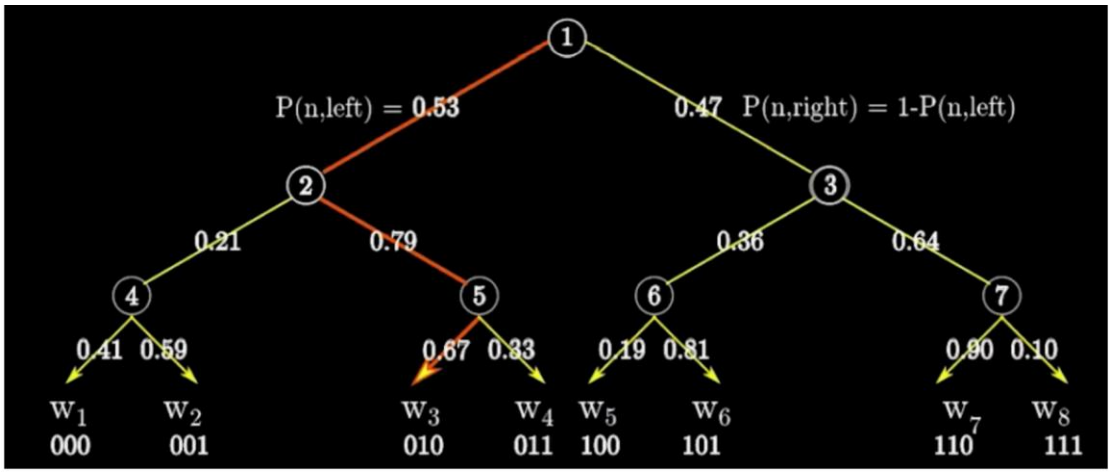
Skip-gram



CBOW

2.3. 层级 softmax

普通 softmax 需要将所有词的可能性都计算一边，因此时间复杂度是 O(n)，将其构建成树，则时间复杂度为 O(log(n))。



首先构造出一个树，树的叶子节点表示所有词表中的词。输出层为 树的中间节点，对中间节点进行 sigmoid 计算，选择左子树或者右子树，直到走到叶子节点，即为预测的单词。

2.4. 负采样

用逻辑回归的思路，将问题转换为二分类问题。将中心词对于的背景词划分为正样本，随机选取正样本之外的单词作为负样本，在训练的时候只需要判断输出词是输入词的正样本还是负样本即可。

dataset

input word	output word	target
not	thou	1
not	aaron	0
not	taco	0
not	shalt	1
not	mango	0
not	finglonger	0
not	make	1
not	plumbus	0
...	...	...

1 为正样本，0 为负样本