# 基于embedding方法的yelp餐厅推荐

## Yelp数据集

### 1.1 概况

Yelp是美国最大的点评网站，囊括了各地餐馆、购物中心、酒店、旅游等领域的商户。用户可以在Yelp网站中给商户打分，提交评论，交流购物体验等。在Yelp搜索一个餐厅或者旅馆等，就能看到它的简要介绍与网友的点评，包含给的星级评分以及评论等。Yelp与国内的大众点评是同一类型的产品。

Yelp在2017年公开了其内部数据集。据官网介绍，这是一个通用的数据集，开放的主要目的是帮助学习。该数据集是Yelp涵盖的商户、点评和用户数据的一个子集，主要可以用于学习NLP、推荐等。

### 1.2 数据集组成

Yelp数据集有两种格式：JSON和SQL，包含470万条用户评价，15万条商户信息，20万张图片，12座城市。此外，数据集还涵盖110万用户的100万条tips，超过120万条商家属性，随着时间推移在每家商户签到的总用户数等。

从Yelp网站下载的数据集包含6个JSON文件：business.json（商户信息），checkin.json（用户签到信息），photo.json（图片信息），review.json（用户评论），tip.json（tips）以及user.json（用户信息）。对于餐厅推荐而言，我们需要的是餐厅信息、用户对餐厅的点评等。选择Yelp数据集的原因是它的数据量大、数据丰富，有很多东西可以挖掘。

### 1.3 数据清洗

鉴于简化实验的需要，我们只选择一个城市作为实验的地区。在众多城市中，拉斯维加斯拥有着最多数量的餐厅，也是一个旅游大城，因此我们最后选择了拉斯维加斯的餐厅用来做推荐。

总体来看，餐厅的推荐需要餐厅的一些基本信息，包括菜系、环境、平均评分等；还需要用户的信息，包括口味偏好、需求的餐厅类型、质量等，这些一部分可以从历史记录中得到。在本实验中，我们认为历史记录即用户评论过的餐厅。

数据清洗大致分为以下几个步骤：

1. 从原始数据中提取拉斯维加斯的餐厅数据；
2. 提取在拉斯维加斯用餐过的用户数据（即review和tips中的餐厅位于拉斯维加斯的）；
3. 去除冗余的数据列和无用的数据列；
4. 处理缺失值，规范化列的属性值（如餐厅WiFi的情况统一为No, Free和Paid三种）；
5. 将时间戳标准化为日期；
6. 进一步筛选餐厅和用户，为了提高推荐效果，剔除了review数目较少的用户，这样也能同时减少数据量。

数据清洗完成后，总计6450家餐厅以及7610个用户。

## Embedding方法

### 2.1 Embedding方法简介

Embedding简单来说就是用一个低维向量表示一个物体等，能够在使用低维向量对物体进行编码的同时还能保留其含义特点，这使得它非常适合深度学习。

在自然语言处理领域，最为著名的embedding方法之一是word embedding。如果使用one-hot编码词向量，整个向量维度会因为语料库的体积而十分庞大，造成维度灾难。使用word embedding方法可以将one-hot编码的向量映射到低维的实数向量。Google提出了一个著名的word embedding的框架Word2Vec，该框架可以在百万数量级的词典和上亿的数据集上进行高效的训练，而且训练得到的词向量可以很好地度量词与词之间的相似性。

### 2.2 Word2Vec

很多人误以为word2vec是一种深度学习算法，其实它的背后是一个浅层的神经网络，分为两种模型：CBOW模型和Skip-gram模型。

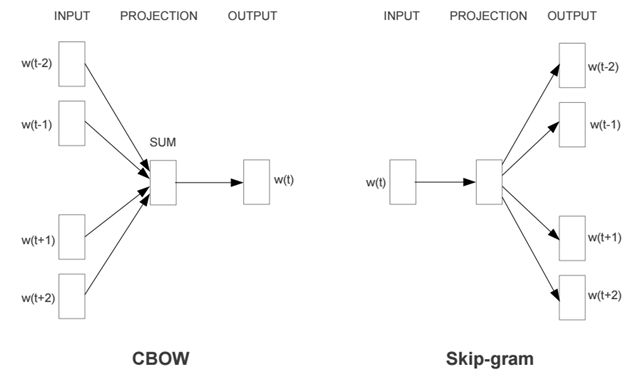


图2.1 CBOW模型（左）和Skip-gram模型（右）

图2.1给出了两种模型的基本框架，可以看出，两种模型的思路其实是刚好相反的。CBOW模型输入的是某一个中心词的上下文相关的词对应的词向量，输出是该中心词的词向量；而Skip-gram模型输入时一个中心词对应的词向量，输出是该中心词上下文相关的词对应的词向量。CBOW模型适合小型数据，而Skip-gram模型在大型语料库中表现更好。

与普通的神经网络相比，word2vec的隐藏层后不跟激活函数，而是在输出层加一个softmax函数，隐藏层的神经元个数即embedding后的向量维度。

使用神经网络实现两种模型的计算量是非常大的，每一次迭代都需要更新整个权重矩阵。为了加速计算，Google提出了两种加速方法：分层softmax和负采样。

1. 分层softmax：如图2.2所示，该方法实际上是采用哈夫曼树代替神经网络中隐藏层到输出层的映射，这样可以避免计算所有词的softmax概率，其本质是把N分类问题变成log(N)次二分类问题。具体每一次二分类问题采用二元逻辑回归的方法，判别正负类别使用sigmoid函数，目标是找到适合所有节点的词向量和内部节点使得训练样本达到最大似然。由于哈夫曼树本身的特点，高频词所在的层非常浅而生僻词所在的层非常深，这就导致它对生僻词的效率比较低，因此适合小样本。

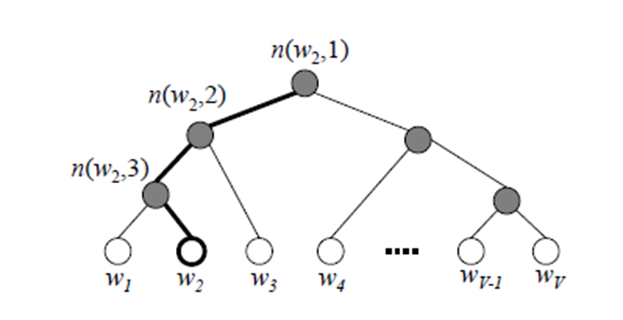


图2.2 分层softmax

2）负采样方法：对于一个中心词为的训练样本，上下文为 ，这是个真实的正例。通过负采样，得到 个和 不同的中心词 ，这样 和 组成了 个并不真实存在的负例，进行二元逻辑回归，得到负采样对应的每个词对应的模型参数 和每个词的词向量。负采样方法更适合大型语料库的训练。