深度学习与自然语言处理

韩升华
Sy2303405
419130449@qq.com

摘要

词向量是对一个词在词典中的一组向量表示，用来抽象化描述该词在词典中的特征。基于词向量，可以将词与词之间的关系表述为数值，从而能够依靠数值的关系，来推测一个词的上下文，或者通过上下文来推测未知词。把 word 视为构成语言的基本单位，则 word embedding 则可以认为是将文本中某个 word 映射到某个数值向量空间的一个函数。神经网络是一种生成映射的方法。

对于词向量，输入是某个文本的词汇表 vocabulary，输出则是按照对应的编码方式，得到每个 word 对应的一组向量。

基于此种方法，可以使得自然语言在计算机中的可操作性大大提高，对文本的分析能力也剧增。

对于词向量的建立，可以选择词频或者预测来作为依据。

词向量在基于语义信息的计算发挥了较为明显的作用，能够判断语义的合理性，词之间的相似性。

引言

Word2Vec 模型是一类常用的词向量模型，属于浅层双层的神经网络，该类模型用以训练词向量，推测词相邻位置的输入词，不考虑词出现的顺序，训练结束后，能够建立每个词到向量的映射，用以文本处理。Word2vec 依赖 skip-grams 或 CBOW（连续词袋）来建立神经词嵌入。

最基本的词向量 One Hot Encode，即是建立一个 n 维零向量，对每个 word，将不同的位赋值为 1，此时，向量的维数等于次数。

经过训练之后，得到一组神经网络的权重，从而将 OneHotEncode 降维得到 Word2Vec 模型，并能将降维后的模型推测出其他词出现的概率。

对于复杂的 Skip-gram 模型，则是在一个词的基础上给出多个位置的频率。

而对于 CBOW 模型，则恰恰相反，是在给出上下文的情况下推测出该位置的词。因而是多个词向量下给出一个输出频率。

实验步骤

1. 文本预处理

在文本中，删去停词与多余的符号，采用了 python 中的 jieba 库，来进行停词操作。从而得到全文本基于词的 vocabulary。

2. 文本添加

添加人物，功法，门派相关文本，用以后续的训练以及测试。

3. 模型训练

在 Word2Vec 模型训练过程，采用了 gensim 库中的 w2v.Word2Vec 函数，其中 sg=1 表示使用 skip-gram 模型作为训练模型。

4. 结果

print(model.wv.similarity('人名 1', '人名 2'))#显示两个词向量之间的距离，进而体现两个人的关联程度。

print(model.wv.most_similar("人名", topn=n))#给出与某个人关系最为接近的十个人

d, _ = model.wv.most_similar(positive=[c, b], negative=[a])[0]#按照 a 与 b 的相对关系，寻找 c 的关系人

结果分析

find_relation("武当派","张三丰","天地会")

输出为天地会 陈近南，武当派领袖为张三丰，而天地会总舵主为陈近南。

print(model.wv.most_similar(u"张无忌", topn=10))

[('赵敏', 0.5588545799255371), ('周芷若', 0.5496096014976501), ('杨逍', 0.5072521567344666), ('小昭', 0.4658464789390564), ('范遥', 0.4538869559764862), ('殷天正', 0.45196396112442017), ('韦一笑', 0.4477475583553314), ('明教', 0.4476700723171234), ('谢逊', 0.44541293382644653), ('义父', 0.43570035696029663)]表示了与张无忌接近的关联词

print(model.wv.similarity('张无忌', '周芷若'))

print(model.wv.similarity('张无忌', '赵敏'))

给出了张无忌与赵敏和周芷若的相对距离，0.5496096 0.5588546

print(model.wv.most_similar("韦小宝", topn=10))

[('小桂子', 0.5290344953536987), ('索额图', 0.5247835516929626), ('多隆', 0.5214013457298279), ('康熙', 0.5192488431930542), ('建宁公主', 0.5164569020271301), ('苏菲亚', 0.5089921355247498), ('奴才', 0.508526086807251), ('施琅', 0.5018771886825562), ('费要多罗', 0.4917173981666565), ('吴之荣', 0.49091485142707825)]输出了与韦小宝接近的词

print(model.wv.similarity('韦小宝', '阿珂'))

print(model.wv.similarity('韦小宝', '双儿'))

print(model.wv.similarity('韦小宝', '建宁公主'))

print(model.wv.similarity('韦小宝', '苏荃'))

print(model.wv.similarity('韦小宝', '沐剑屏'))

print(model.wv.similarity('韦小宝', '曾柔'))

print(model.wv.similarity('韦小宝', '方怡'))

给出了七组词之间的距离，

0.35198486

0.47142348

0.51645684

0.43476936

0.39979503

0.360509

0.4352182

对于我比较了解的韦小宝一词，结果基本合理，周芷若，赵敏与张无忌也基本符合事实。