Python 进阶训练营

尹会生

11 语义情感分析

目　录

CONTENTS

01 使用 snowNLP 进行语意分析

02 使用 tesseract 自动识别验证码

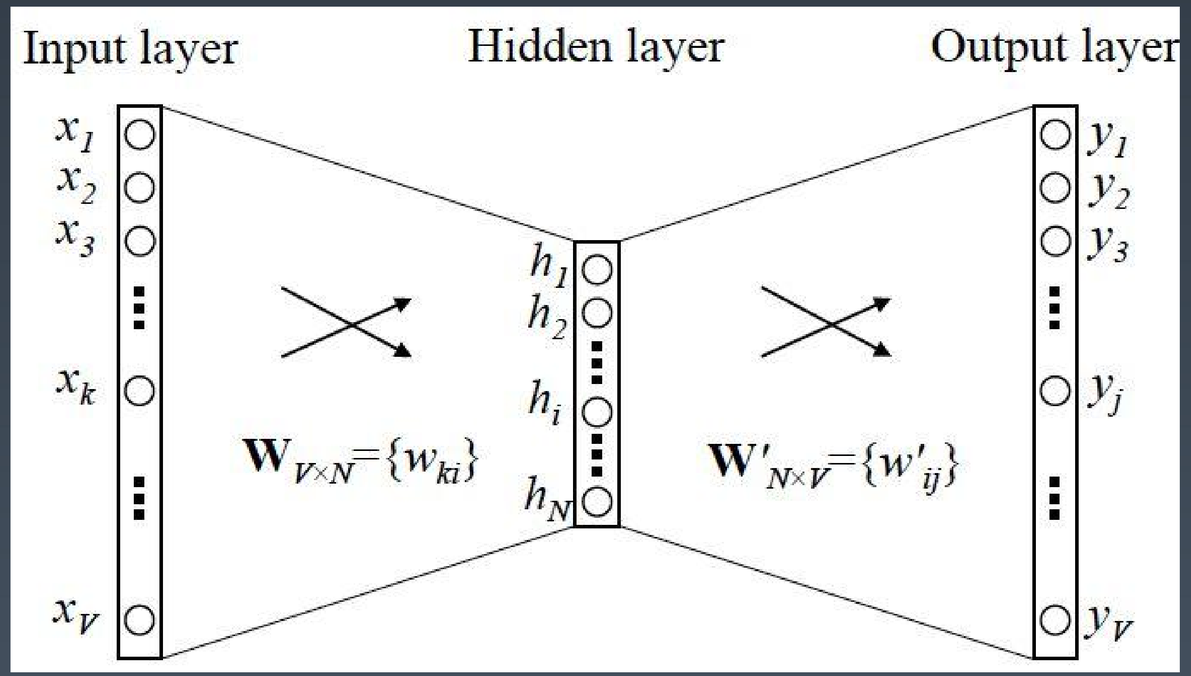
学习目标

1.学会使用自然语言库进行情感分析

1.学会训练自己的语义情感分析库

词向量

word2vec 模型



snowNLP

中文分词（Character-Based Generative Model）
词性标注（TnT 3-gram 隐马）
情感分析（现在训练数据主要是买卖东西时的评价，所以对其他的一些可能效果不是很好，待解决）
文本分类（Naive Bayes）
转换成拼音（Trie 树实现的最大匹配）
繁体转简体（Trie 树实现的最大匹配）
提取文本关键词（TextRank 算法）
提取文本摘要（TextRank 算法）
tf，idf（信息衡量）
Tokenization（分割成句子）
文本相似（BM25）

模型训练与排序

使用 CCF 数据进行 word2vec 训练
使用二部图算法进行排序

使用 tesseract 识别传统验证码

pytesseract 库的安装
使用 pytesseract 库识别字符验证码

THANKS!