机器学习和 NLP 工程师的能力结构和学习路线

学习目标：

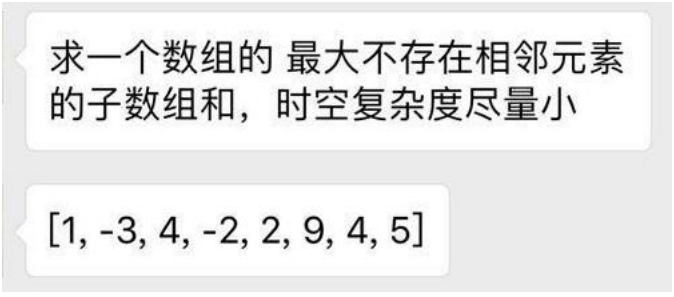
所谓纲举目张，介绍一下机器学习和 nlp 工程师的能力模型，为学习者明确学习路线图，为后面的学习明确方向。

1.说说面试

linux uniq sort sed awk 等命令灵活使用
二分查找 c 语言版本 python 版本
快排
泰勒公式
输入法怎么组织词表
用过 hadoop 没有
推荐 app

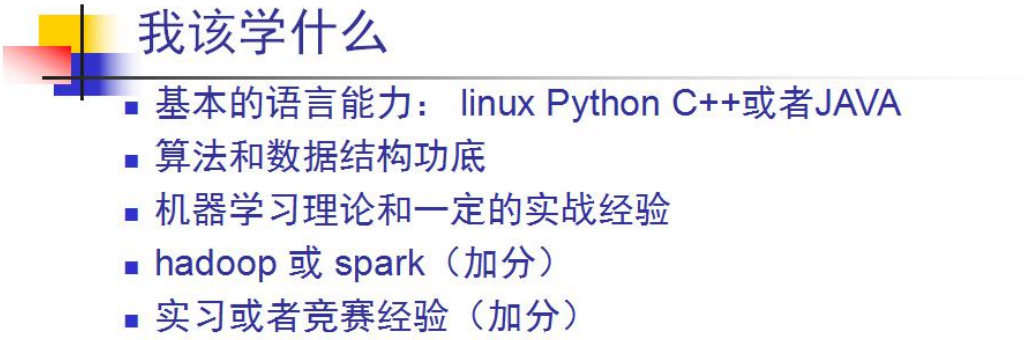
lda 原理
word2vec 原理
kmeans 的 map reduce 实现
lr 模型参数估计
boosting 的过程
cnn 文本分类的过程
字典树分词函数
hadoop 大小表

其他人的面试题
lstm 编辑距离 最长公共子串
手推最大熵



面试到技能目标

2.nlp 工程师的能力模型



2.1 编程基本功

python 和 c 或者 java 手写代码的能力
leetcode 类题目
shell 脚本熟练作用来处理文本
大数据方面 map reduce 思想解决具体任务

2.2.机器学习原理和工具

终极目标
各种模型的原理 要能够手推
各种模型造轮子 python 手写

数学基础:
深刻理解原理 需要数学底子

机器学习传统理论
深度学习
实践工具
（ numpy sklean pandas xgboost lightgbm libsvm liblinear weka 等）tensorflow(keras)

重要项目
Fasttext：facebook 短文本分类
gensim
word2vec
glovec

2.3.自然语言处理

知识点也很多
NLP 任务分类：
*序列标注问题
（命名实体 品牌词识别 中文分词（词性标注） 句法分析 新词发现）
*分类问题
（情感分析 行业分类 意图识别 ）
*改写问题
（query 扩展 改写 纠错 翻译）
*生成问题
（自动写稿 自动写诗 文本摘要 聊天机器人 自动问答 ）

工具：
序列标注经典工具：crf++
stanford corenlp
nltk spacy textblob
syntaxnet

思考：
下面两个任务 分别属于哪一类 NLP 任务？
判断一句话是不是 黄反言论

判断一句话中哪几个词是黄反词

nlp 高级任务
语言模型
kenlm 语言模型 rnn 语言模型
kenlm 语言模型：http://kheafield.com/code/kenlm/
rnn 语言模型 ：https://github.com/wpm/tfrnnlm

怎么训练一个语言模型
1、 语料获取及预处理
1 billion words
https://github.com/ciprian-chelba/1-billion-word-language-modeling-benchmark
下载其 training-monolingual 语料 并使用 get_data 脚本进行 tokenize 和 shuffle（ 不清楚 shuffle 对于语言模型有多大影响）
2、训练
传统方法 kenlm
https://kheafield.com/code/kenlm/
nohup /kenlm/build/bin/lmplz -o 5
<./training-monolingual.tokenized.shuffled/all > lm.arpa &
rnn 语言模型 ：https://github.com/wpm/tfrnnlm

语言模型可以对 句子通顺度建模；可以用于纠错

句法分析
Stanford parser 最经典
syntaxNet 最快、准确率最高

机器翻译
引入 seq2seq 模型的神经机器翻译，不仅仅是语言翻译

2.4 项目经验

文本分类问题的思路
特征选择
粒度:字粒度 词粒度
Ngram：1~5
长短文本：
长文本 tf-idf ； 摘要
短文本 word embedding

模型类型
传统机器学习模型：LR 朴素贝叶斯 SVM 最大熵 GBDT 随机森林 KNN 等等

深度学习模型：CNN ， RNN（包括普通 RNN LSTM 双向 LSTM GRU 等），CNN 与 RNN 结合，当然也有用普通的前馈神经网络做的（一般的 DNN）

其他任务：

聚类 Rank 模型 CTR 预估 其他（打标签 构建知识图谱 等）

3.总结

要沉下心去，不能浮在表面：

知识点很多

速成是骗人的：通过一个培训班学习十次课 每次课一两个小时就能成为机器学习工程师？学习机器学习很简单？都是培训班为了忽悠说的假话！

扎扎实实学下来 得 3 个月 按照 996 强度 目标： 进 BAT

学习很容易 学会难 要动脑思考 动手推导 动手实践

怎么成为技术高人：

读 paper

读源码

多看书

全面提高自己（数学 英语 编程 数据结构....）

共勉