kdsec 技术栈上手白皮书

概述

目的

- 帮助新入团队的学生（员工）快速熟悉组内技术栈
- 帮助新入团队的学生（员工）进入工作角色

岗位划分

按照工作职责划分为岗位：

1. 前端研发岗（前端）
2. 后端研发（业务后端研发、大数据研发）
3. 算法研发（DL&ML类）

工具

- 服务器登录终端软件：secureCRT（推荐）、Finalshell、xShell
    - secureCRT注册：https://www.cnblogs.com/lccsdncnblogs/p/17517383.html
    - secureCT配色：https://blog.csdn.net/u014530704/article/details/78698060
- 代码编辑器：Vscode
    - Vscode远程连接服务器：https://blog.csdn.net/qq812457115/article/details/135533373
- 终端管理工具：tmux。http://mingxinglai.com/cn/2012/09/tmux/

基础知识

shell 基础知识

- 基本的shell脚本编程编写
    - 菜鸟学习教程
    - 变量定义（重要）
    - 传参（重要）
    - 数组（重要）
    - 循环（重要）
- set 命令教程
- 让进程在后台运行更可靠的几种方法
- linux定时执行脚本
- bash快捷键

常见linux命令

- linux 命令大全网址
    - https://www.runoob.com/linux/linux-command-manual.html

    - https://man.linuxde.net/

- 务必了解文件和目录管理的命令

    - ls：查看文件
    - cd：进入目录
    - pwd：查看当前路径
    - mkdir：创建文件夹
    - touch：创建文件
    - cp：拷贝文件
    - scp：远程拷贝文件
    - mv：修改文件名称
    - rm：删除文件
    - which：查找并显示给定命令的绝对路径
    - tar：解压缩文件
    - find：查找文件

- 务必了解的文本处理命令

    - cat：输出文件
    - more：全屏幕的方式按页显示文本文件的内容
    - less：与more类似，不同的是less命令允许用户向前或向后浏览文件
    - head：输出文本头部数据
    - tail：输出文本尾部数据
    - grep：文本搜索工具
    - tr：对来自标准输入的字符进行替换、压缩和删除
    - wc：计算数字
    - cut：显示行中的指定部分
    - paste：将多个文件按照列队列进行合并
    - diff：查看文件不同
    - sort：对结果排序
    - uniq：用于报告或忽略文件中的重复行
    - iconv：文本格式转换
    - sed：文本处理必备
    - awk：文本处理必备

- 务必了解的系统工作命令

    - top：实时动态地查看系统的整体运行情况
    - ps：报告当前系统的进程状态
    - kill：杀死进程
    - history：查看历史命令
    - curl or wget：下载资源
    - date：查看日期
    - echo：输出

python 基础知识

必须学会的知识点

- 基本语法
- python常见内置数据结构
    - list
    - dict
    - map
    - set
    - collection
- python 文本和字节序列
- 文件操作
- 面向对象知识
- 基础网络编程知识

拔高知识点

- 多线程
- 迭代器与生成器
- 装饰器
- 进程间通信

学习资源

- Python教程- 廖雪峰的官方网站
- 草根学python
- Python 3.7.6 中文文档

Git 基础知识

必须学会的知识点

- 安装配置git
- 使用github、gitlab进行代码管理(git clone、git status)
- git创建分支(git brach)
- git提交程序(git add、git rm、git commit、git push、git pull)
- 版本管理（git reset、git checkout）

拔高知识点

- 代码合并（git diff、git merge、git rebase）
- 日志管理（git log、git reflog）
- 打标签（git tag）

学习资源

- 廖雪峰：Git 基础教程
- 官方git 说明
- GUI客户端可使用sourcetree(推荐)

常见git 命令

- git init

- git clone
- git remote add origin ***.git
- git push -u origin master
- 推送到远程仓库的dev分支：git push origin dev
- git log
- git log --graph --pretty=oneline --abbrev-commit
- git status
- git diff
- git add *
- git commit -m "message"
- commit之后又改了一个小bug，但是又不想增加一个commit，可以用：git commit --amend --no-edit，直接将改动添加到上一次的commit中
- git push
- git pull
- touch .gitignore

vim 基础使用知识

- vim 的基本配置
- vim 进行文本的添删改查
- 从头到尾练习简明 VIM 练级攻略

组内 API 编写规范

组内API编写规范

前端

- 学生背景：只有薄弱后端研发基础甚至零基础的研一学生，有过简单html/css编程基础，熟悉简单的linux操作命令，有git使用经验
- 目标：帮助该学生尽可能上手组内的前端项目技术栈，2个月内能开发前端小型功能模块（eg：对接后端接口，开发支持不同搜索条件的检索单页面，可对检索结果进行分页、排序）

必须学会的知识点

- 掌握html、css，会进行简单的排版
- 掌握Javascript(ES6)语法
- 学会使用webstorm 或vs code进行开发环境搭建
- js操作dom 和jQuery
- 使用Vue构建基本应用
- 学会ajax 掌握网络传输协议
- 学会debug Chrome浏览器调试
- 异步/同步编程方式，及ES7的async await的使用
- 前端的webpack、npm、yarn等工具链
- 熟练使用Python，用flask构建简单的后端

拔高知识点

- 掌握1-2个框架，先学会bootstrap,并学会开发个简单应用
- 深入学习js 如this指向，es6特性
- 深入学习vue 特性或react特性
- 熟悉ant design Vue版本的api并能构建应用
- 掌握vuex、vue-router插件的使用
- 常用设计模式
- 了解MVC及MVVM
- 编译原理及Vue等框架的工作原理
- less、sass等css拓展与语法
- 掌握单元测试的编写
- 熟悉cytoscape的api并能构建图应用
- 熟悉echart、d3.js等库并能构建可视化组件
- webpack原理及配置
- RESTful接口规则

学习资源

- 菜鸟教程：先入门 html js jQuery vue等
- 书籍：javasrcipt权威指南
- HTML教程
- HTML-CSS基础教程
- JavaScript基础教程
- ES6标准入门
- Less、Sass入门
- Vue源码
- Vue源码解读
- Vuex教程
- Vue-router教程
- Ant design Vue版文档
- Cytoscape图可视化库文档

额外的建议

- 要动手练习，比如自己定一个目标（比如带数据库的TODO-list），使用上述技术栈构建一个前后端分离、MVVM、ES6、使用webpack/npm/yarn等工具链，完成一个页面布局精美、功能齐全的基础应用
- 以实际项目上手，效果更好一些，可能干巴巴的看些书籍，资料可能效果不佳。
- 最好以实际的例子，应用入手，如开发一个小型管理系统等。逐步丰富。

学习能力评价

任务1：实现kdsec小组博客前端页面

核心要求如下：

- [ ] 网页界面至少包括：主页、概况、成员、论著、项目、资源、动态等部分
- [ ] 支持文章的编辑、修改、发布等功能
- [ ] 支持文件的上传
- [ ] 支持pdf、word、ppt的在线预览
- [ ] 支持视频的在线播放

后端

- 学生背景：只有薄弱后端研发基础甚至零基础的研一学生，会利用python进行简单脚本的编写，熟悉简单的linux操作命令，有git使用经验
- 目标：帮助该学生尽可能上手组内的后端项目技术栈，2个月内能开发后端接口小型功能模块（eg：针对mysql中的公文数据，实现某条公文的添删改查接口）

必须学会的知识点

- Linux 知识
  - Linux基本命令，尤其是ssh命令要会用，以及kill，top，ps，iptables等等服务器管理命令
  - Linux用户，权限等
- Git的使用，尤其是分支功能的使用
- 熟悉SQL语言基本编写
- 书写mysql数据库的使用
- FastApi后端框架，是整个后台的基石，必须掌握
- Sqlalchemy的使用，增删查改
- Python：virtualenv的使用，为项目搭建虚拟环境；requests模块发送http请求
- Postman的使用
- 单元测试，python单元测试框架pytest
- Web项目架构，基于HTTP协议的前后端交互流程
- RESTful接口设计与规范

拔高知识点

- Web服务器gunicorn
- python-kafka的基本用法
- 任意一种爬虫框架，scrapy,xpaw等等，都大同小异，重点是要了解爬虫框架的整体工作流程
- Docker的使用，有些工具提供了docker版
- 熟悉ElasticSearch的基本使用
- 熟悉neo4j Cypher语言对图数据库进行CUID操作
- 数据Nebula Graph数据库的使用
- Pdb的使用，python的debug工具，定位bug十分方便
- 单机并发解决方案，多进程，多线程，协程，python GIL
- 前端模块化开发，VUE框架，webpack
- 软件开发设计模式
- 软件架构方法论，C4模型

学习资源

- 官方文档是最好的学习资源！建议要这个工具的官网去看Tutorial和Docs，要习惯看英文文档
- 博客是次好的资源，尤其是他人踩过的坑，CSDN，知乎，简书等等，都可以
- 谷歌大法好，遇过问题先去搜索一下，90%的问题都可以从谷歌上找到解决方法
- 百度大法好，如果谷歌用着不方便的话，百度也可以用，搜中文的话，百度结果会更好

额外的建议

- 多动手写例子，光看是学不会的

- 配环境既重要，又麻烦，要多动手去配环境
- 系统性学习： 推荐书籍，系列博客，付费专栏等，但不推荐通过购买书籍等方式学习开源系统和框架，这些东西会升级更新换代，书籍适合学习相对不过时的知识，比如设计模式等。
- 解决问题：最快的办法是搜索，包括中文和英文，github issue等，深入解决问题的办法是文档和源码。

学习能力评价

任务1：实现任一领域的图谱检索与存储服务

核心要求如下：

- [ ] 图谱实体数量不少于50万
- [ ] 能够实现实体的添加、删除、修改功能
- [ ] 能够实现实体的添加、删除、修改功能
- [ ] 能够实现基于关键词对实体名称进行检索
- [ ] 任一接口请求返回时间不高于100ms
- [ ] 包含所有接口的测试用例
- [ ] 能够以简单可视化形式对接口进行功能展示

算法

- 学生背景：本科毕业有一定基础的研一学生或者低年级直博研究生，会利用python进行中等难度脚本的编写，熟悉常见的linux操作命令，有git使用经验，了解过ML、DL的相关知识，但缺乏某一种深度学习框架（TF、pytorch等）进行模型训练的实战经验
- 目标：帮助该学生尽可能上手组内的算法技术栈，2个月内可以针对组内某个需求，从0到1开发出简单算法接口（eg：给定足量标注后的不同类别新闻数据，经过训练之后，实现新闻类别判断的API接口）

必须学会的知识点

- 神经网络与深度学习基础知识（概念及原理）
  - 前馈神经网络
  - 卷积神经网络
  - 循环神经网络
  - 激活函数
  - 梯度回传
  - 感知机模型
  - Logistic模型
- 面向自然语言处理的深度学习（原理及应用）
  - 注意力机制
  - 词向量
  - 预训练表示模型
  - 序列标注模型
- 各类损失函数的用法
  - 交叉熵
  - 合页损失
  - 均方损失
  - 三元组损失
- 神经网络参数调整方法

    - 初始化
    - 学习率
    - dropout
    - 各种Normalization方法（BN、LN等）
- pytorch
    - pytorch的基本用法（能够熟练掌握基本API的使用）
    - pytorch模型的训练、测试、封装与调用
    - pytorch模型的服务部署
- 常见自然语言处理工具的使用，包括但不限于LTP，CoreNLP，jieba，fastNLP等
- 学会灵活使用Google、Stack Overflow进行Debug

拔高知识点

- 常见ML&DL基本算法的原理
    - 支持向量机
    - 贝叶斯理论
    - 集成学习方法
    - 注意力机制
    - fasttext
    - 预训练语言模型
- pytorch_pretrained_bert 包的使用（使用0.6.2版本完成Mrpc数据分类）
- numpy 的使用方法和基本API
- 尝试阅读ACL、EMNLP、AAAI、IJCAI等顶会的NLP论文，逐渐形成NLP的知识体系

学习资源

- 自然语言处理入门练习，https://github.com/FudanNLP/nlp-beginner
- 吴恩达 机器学习系列课程CS229 https://www.coursera.org/learn/machine-learning
- 《神经网络与深度学习》，邱锡鹏， 重点学习1、3.1、3.2、3.3、4-8章，https://nndl.github.io/
- 《机器学习与深度学习》李宏毅，https://www.bilibili.com/video/BV1JE411g7XF
- 李航《统计学习方法》（小蓝书）https://item.jd.com/12522197.html
- 周志华《机器学习》（西瓜书） https://item.jd.com/11867803.html
- GoodFellow《深度学习》（花书）https://item.jd.com/12128543.html
- 邱锡鹏《神经网络与深度学习》https://item.jd.com/12851292.html
- 王斌《机器学习实战》https://item.jd.com/11242112.html
- IBM出品：pytorch seq2seq示例代码
- pytorch实现文本分类
- 莫烦pytorch教程

额外的建议

- 使用LaTeX或Markdown整理学习笔记，同步更新到个人主页中（可使用github.io构建），笔记格式参见https://github.com/Sakura-gh/ML-notes
- 阅读论文首先今三年定会论文开始看起，不要贪多求全，以打牢基础作为第一目标
- 阅读相关基础论文，培养从经典与前沿论文中汲取的能力
- 努力写出整洁、优雅、可读、可复用的代码

学习能力评价

待定

结语

-
- 受时间和能力限制，难免有不足之处，欢迎大家批评指正。
- 意见反馈邮箱：sutaoyu@iie.ac.cn

修订记录

- v1 2020.05 初稿
  - 初稿各部分内容由下列kdsec小组成员提供了建设性意见，向各位表示诚挚的感谢（排名不分前后，以拼音顺序为准）：
    - 基础知识：苏涛宇
    - 前端部分：李彦增、穆红章
    - 后端部分：从鑫、王玉斌
    - 算法部分：王栋、郁博文、张振宇
- v1.1 2021.06 修改部分错误
- v1.2 2023.07 删除过时内容
- v1.3 2024.12 增加工具使用

在线地址

- https://github.com/sutaoyu/Let-us-warm-up