**构建智能BI系统**

1. **项目理解**

开发一个web平台使得非专业技术人员能够通过输入自然语言获得想要的信息，比如用户输入（2017年上半年财务报表），系统自动显示可视化报表和图形。项目对研究的数据内容不限。

**相关概念：**

NLP被广泛用在语音识别，语音合成，自动分词，句法分析，语法纠错，关键词提取，文本分类/聚类，文本自动摘要，信息检索（ES,Solr），信息抽取，网络爬虫，知识图谱，机器翻译，人机对话，机器写作，情感分析，文字识别，问答系统，推荐系统，高考机器人等项目中。

**二、项目计划**

1. **数据来源**

**学习阶段：**

选取小的数据集学习训练模型，数据集的类型可以多样，比如：文本分类、机器翻译、图像字幕、自动文摘等。

参考：<http://blog.csdn.net/u010708470/article/details/52749535>

<http://baijiahao.baidu.com/s?id=1580265790553760692&wfr=spider&for=pc>

**项目实践阶段：**

确定系统构建的数据集类型，由小组某个成员使用python或者R等编程语言爬取所需数据，或者到开源平台下载

1. **数据处理**

**学习阶段：**

将数据集特征化和数值化后转化为可训练的数据模型

**项目实践阶段：**

对于爬取的数据进行清洗：编写python或R脚本，过滤清洗噪音

对于爬取的数据进行转换：将数据转换成便于存储的数据结构，存入数据库中

1. **NLP技术**

**学习阶段：**

掌握NLP基本概念，理解NLP技术实现一些项目的原理，如人机对话，机器翻译，机器写作，情感分类，观点挖掘等。

掌握词性标注和实体识别的算法原理和应用场景，用选择的模型把词语、句子、文本等映射为向量，用于进一步训练。

掌握NLP核心算法，机器学习算法和NLP技术结合使用，实现文本处理。

掌握jieba介绍，安装及在Wing Ide使用，Gensim介绍，下载，安装及在Wing Ide中使用

**项目实践阶段：**

模拟用户输入语句，进行句法分析和数据格式化输出，比如json格式，再转化成sql查询语言到数据库中查询。可能涉及到的技术点如下：

* 词性标注与命名实体识别

以新闻类数据举例说明：实现新闻中人名、地名、机构名自动识别与词性标记

实体识别和词性标注开发包：Stanford NLP（斯坦福NLP）  
1）安装stanford nlp自然语言处理包：pip install stanfordcorenlp

2）下载Stanford CoreNLP文件

<https://stanfordnlp.github.io/CoreNLP/>

3)下载中文模型jar包

<http://nlp.stanford.edu/software/stanford-chinese-corenlp-2017-06-09-models.jar>

* 分词技术

常见分词工具：ICTCLAS，Tokenize，jieba分词

加载自定义字典

* NLP常用开发包之Gensim

Gensim模型训练

1. **平台开发**

**学习阶段：**

了解java框架技术，熟悉SSM框架搭建的全过程，掌握JEasyUI的API及使用技巧

**项目实践阶段：**

前台：用户界面，注册，登录，信息修改，提供供用户搜索的信息框，实现文章搜索引擎；信息展示界面

后台：管理员用户管理，信息类别添加，修改，删除，查询（时间不够的话这一步不做成动态的）

1. **开发环境**

开发工具：wing ide6;

开发语言：python、R

其他工具：Stanford NLP，ICTCLAS，Tokenize，Jieba，Gensim，

JEasyui+SSH2/SSM+Maven+MySQL技术