**自然语言理解实验指导书**

**实验一、语料库的收集与整理**

实验目的：

1.理解词频统计的概念

2.理解Ngram文法的概念

3.掌握从文本中统计生成词典的技术

4.了解文本不同的编码方式

实验内容：

1.熟悉开发环境（可采用Python，C/C++或者Java）

2.根据文本的编码方式，编写词频统计程序

实验步骤：

二，系统设计

1.打开开发环境，根据自己熟悉的语言，确定开发环境。

2.下载语料库（ci.txt和新闻语料库）到特定目录下，或从https://github.com/chinese-poetry/chinese-poetry中下载。

3.根据文本编码，加载语料库文本

4.分别统计n-gram（n=1，2）的词频，存储到相应的数据结构，该数据结构包括词（词本身）和词的频度（出现次数）

5.将内存中的数据结构存储到文本中，方便后面随时加载。

**实验二、词汇知识库使用技术**

实验目的：

1.理解词典的概念

2.理解文本生成的概念

3.掌握从词典中生成文本内容的技术

实验内容：

1.熟悉开发环境（可采用Python，C/C++或者Java）

2.根据词典中的词和词频，随机生成文本内容，或采取优化方案生成流畅的文本内容。

实验步骤：

二，系统设计

1.打开开发环境，根据自己熟悉的语言，确定开发环境。

2.将实验一中生成的词典到特定目录下

3.将词典加载到内存中，主要包括词和词频

4.采用随机生成，或者n-gram等算法，生成文本序列

5.构建前端，演示文本生成。其中包括，根据宋词词典，自动生成宋词，根据人民日报的词典，自动生成文本内容。

**实验三、中文分词技术应用**

实验目的：

1.理解中文分词的基本概念

2.掌握2-3种中文分词算法

3.掌握从文本中统计生成词典的技术

4.了解文本不同的编码方式

实验内容：

1.熟悉开发环境（可采用Python，C/C++或者Java）

2.根据文本的编码方式，编写词频统计程序

实验步骤：

二，系统设计

1.打开开发环境，根据自己熟悉的语言，确定开发环境。

2.导入相应的分词词典，如果基于模型，则导入相应的模型。

3.实现FMM和BMM，或者基于n-gram的分词方法

4.实现界面，输入一句话，输出分词结果

5.存储分词结果到文件中。

**实验四、文本分类技术应用**

实验目的：

1.理解文本分类的概念

2.理解文本分类算法的概念

3.掌握文本特征提取和分类算法包应用技术

4.了解多标签分类与单标签分类

实验内容：

1.熟悉开发环境（可采用Python，C/C++或者Java）

2.基于Libsvm等分类算法包，实现文本分类

实验步骤：

二，系统设计

1.打开开发环境，根据自己熟悉的语言，确定开发环境。

2.下载语料库（搜狗新闻分类语料）到特定目录下

3.根据不同的文本特征提取方法，提取文本特征，实现文本向量化，即将文本转换成向量。

4.将向量转换为libsvm等算法包的格式，使用libsvm进行训练，获得模型。

5.搭建演示系统，输入一段文本，输出分类结果。

**实验五、自动问答系统构建**

实验目的：

1.理解问答系统的概念

2.理解问答匹配算法

3.掌握从基于序列到序列模型，构建问答系统的技术

实验内容：

1.熟悉开发环境（可采用Python，C/C++或者Java）

2.基于序列到序列的神经网络模型，实现问答

实验步骤：

二，系统设计

1.打开开发环境，根据自己熟悉的语言，确定开发环境。

2.下载语料库（电影字幕语料库）到特定目录下

3.根据文本编码，加载语料库文本

4.基于TFIDF或者神经网络等模型，将文本转化为向量，作为模型训练的输入。

5.训练循环神经网络，实现问答模型。

6.搭建演示系统，输入问句，输出答案。

**实验六、文本摘要系统**

实验目的：

1.理解文本摘要系统的概念

2.理解序列到序列训练算法

3.掌握从基于序列到序列模型，构建文本摘要的技术

实验内容：

1.熟悉开发环境（可采用Python，C/C++或者Java）

2.基于序列到序列的神经网络模型，实现摘要系统

实验步骤：

二，系统设计

1.打开开发环境，根据自己熟悉的语言，确定开发环境。

2.下载语料库（文本摘要语料库）到特定目录下

3.根据文本编码，加载语料库文本

4.基于TFIDF或者神经网络等模型，将文本转化为向量，作为模型训练的输入。

5.训练循环神经网络，实现摘要模型。

6.搭建演示系统，输入原文本，输出摘要。