《自然语言处理》课程作业

课程编号： 091M4044H　　**课程属性：** 核心专业课　　**学时/学分：** 60/3

预修课程： 概率论与数理统计、算法分析与程序设计

一、作业目的：

通过本课程作业加深对自然语言理解基础理论的认识和了解，锻炼和提高分析问题、解决问题的能力。通过对具体项目的任务分析、技术调研、数据准备、算法设计和编码实现以及系统调试等几个环节的练习，基本掌握实现一个自然语言处理系统的基本过程。

二、作业题目：

1. 实现一个汉语或英语的命名实体自动识别系统（Named entity identification）

命名实体一般指如下几类专用名词：人名、地名和组织机构名。本题目可实现汉语或英语中任意一种类型的命名实体识别，需进行实验分析。

2. 实现一个汉英人名自动互译系统（Chinese-English person name translation）

本题目要求实现一个汉语人名（包括中国人名和外国人译名）和英语人名的自动翻译系统，并进行实验分析。

3. 实现一个汉语自动分词系统（Chinese word segmentation）

本题目要求实现一个汉语自动分词系统，并在微博等非规范文本测试集上进行测试分析。如果在本题目中不考虑命名实体识别问题，歧义消解和集外词处理是汉语自动分词中的关键问题。

4. 实现一个汉语或英语的词类自动标注系统（Automatic part-of-speech tagging）

本题目要求实现一个汉语或英语的词类自动标注系统，并进行实验分析。

5. 实现一个汉语和英语两种语言中数字、日期或时间、货币数量表达的自动识别和翻译系统

数字、日期或时间、货币数量等在自然语言中有特殊的表达方式。如汉语：“2018 年 3 月 18 日”的英语表达是：“March 18, 2018”或“18 March 2018”等。选做本题目时可实现某一种表达的识别和单向翻译，也可实现双向互译。

6. 实现一个（汉语/英语）词义自动消歧系统（Word sense disambiguation）

很多词汇具有一词多义的特点，但一个词在特定的上下文语境中其含义却是确定的。本题目要求实现一个能够自动根据不同上下文判断某一词的特定含义的系统，并进行实验分析。针对汉语或英语均可。

7. 实现一个（汉语/英语）名词短语自动识别系统（Noun phrase recognition）

本题目要求实现一个汉语或英语文本中名词短语自动识别系统，并进行实验分析。

8. 实现一个汉语或英语句子谓语成份的自动识别系统（Predicate recognition）

本题目要求设计并实现一个汉语或英语句子中谓语成份的自动识别系统，并进行实验分析。

9. 实现一个句法分析器（Chinese parser）

针对汉语或英语实现一个句法分析器，并做实验分析。

10. 设计实现一种术语自动识别方法（Term identification）

可以利用维基百科、百度百科等网络内容，或者基于某种专业领域的文本数据，设计实现一种术语识别方法。

11. 设计实现一种汉语句子自动改写方法（Chinese sentence paraphrasing）

设计实现一种汉语句子的自动改写方法，并进行实验分析。

12. 设计实现一种汉英双语可比语料自动获取方法（Chinese-English comparable corpus acquisition）

设计实现一种方法，能够从互联网上自动获取汉英双语可比语料，并进行实验分析。

13. 设计实现一种汉语文字自动检查系统（Chinese text proofreading）

设计实现一种汉语文本的拼写错误自动检查方法，并进行实验分析。

14. 设计实现一种基于文本内容/情感的文本自动分类方法（Text classification）

依据某种文本分类标准实现文本自动分类，并进行实验分析。针对汉语文本或英语文本均可，但**不允许与《文本数据挖掘》课程的作业重复。鼓励自己从网上爬取数据，进行标注，并进行对比实验。**

15. 设计实现一种语种自动识别方法（Language identification）

任意给定一段文本，该方法应能自动识别该文本属于哪一种语言，如汉语、英文、德文、阿拉伯语、维吾尔语等。

三、基本要求：

(1) 每人可以选择其中的一个题目，也可以几个人合作完成一个题目，原则上合作人数不应超过 3 人，彼此之间必须有明确的分工和要求。

(2) 任何一个题目，都不限定采用的方法，可以采用基于规则的分析方法，也可以采用基于语料库的统计方法或基于深度学习的方法，还可以是几种方法的

结合，鼓励方法创新，但必须有理论根据或实验数据依据。

(3) 上述有的题目比较困难，如果不能找到合作的同学共同完成，可以选做上述某一题目中的部分工作，但请说明所做的部分与整个项目其它部分的关系。

(4) 完成一份技术报告，报告内容包括：项目目标、国内外相关工作、自己在本项目中承担工作的不同点、实现系统（或模块）的核心思想和算法描述、系统主要模块流程、实验结果及分析。

(5) 提交系统源代码和可执行程序，以保证实验系统可以正常编译和运行。如果是多人合作完成的，应提交最终集成的系统。

(6) ?月?日（北京时间 24:00）之前提交技术报告、系统代码和可执行程序。请留下作者的姓名、单位、联系电话和邮件地址。可以通过电子邮件提交或直接提交光盘。

四、严正声明：

(1) 鼓励充分使用网络资源和其它一切可以利用的资源（包括数据、语料、软件工具和论文资料等），但严禁侵害他人知识产权，技术报告中必须明确说明所用资源的真实来源。

(2) 鼓励相互交流、相互合作，但严禁抄袭他人工作，严禁伪造结果。

五、参考网站：

[1] 北京大学计算语言学研究所公开了 1998 年 1 月份《人民日报》的分词与词性标注语料，网址为：http://icl.pku.edu.cn/icl_groups/corpustagging.asp。若下载该语料时，请严格遵守该网站的有关规定。

[2] 关于最大熵的开源工具：

OpenNLP： http://incubator.apache.org/opennlp/
张乐： http://homepages.inf.ed.ac.uk/lzhang10/maxent.html
Malouf： http://tadm.sourceforge.net/
Tsujii： http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/maxent/
林德康： http://webdocs.cs.ualberta.ca/~lindek/downloads.htm
……