一、项目要求阐述

项目原始要求：给定自然语言文档，设计一套标签，手工嵌入到需求文档中，使其成为流程式的需求。并使用机器学习方法来自动插入标签。然后根据标签来提取信息，形成结构化的需求。

在此要求的基础上，明确给定的自然语言文档按照GWT形式进行描述，产生的结构化需求描述遵循RUCM格式并组成完整文档。

本项目立足软件需求描述方法，旨在建立一个将GWT形式描述的需求文档进行自动化处理转换为RUCM文档的实用系统。RUCM的需求描述方法相比GWT描述形式既保留了自然语言的表达能力与易懂性，又减少了表达的随意性与二义性。项目预期主要应用在软件工程的需求描述环节，在实现上要综合GWT需求文档与RUCM文档的特点设计文档标签，使用机器学习方法实现标签自动嵌入，由机器根据标签提取信息，转换成RUCM形式的需求文档。

项目要求的具体条目有

a)给定自然语言需求文档：自然语言描述的需求文档是系统的输入数据。给定的自然语言文档以GWT形式描述。在软件开发过程中，需求文档记录了需求分析和导出的结果，是对用户和系统需求的全面详细的描述，自然语言的描述方法是最直观易懂且应用广泛的需求描述方法，但是自然语言的二义性使得这种方式有着明显的弊端，GWT格式的需求描述相比一般的自然语言描述方式准确性更高，但仍然可以转换成结构化程度更高的描述方式。

b)设计一套标签：标签可以较高的层次上对需求文档的结构进行抽象，反映出软件需求的流程特性，在本项目中，需求文档标签是GWT描述与RUCM描述之间的桥梁，通过标签标注出GWT描述中与RUCM描述对应的实体与关系，在此基础上可以迅速地根据标签抽取信息产生RUCM文档。。

c)使用机器学习方法自动插入标签

a)使用机器学习方法：使机器能够对自然语言描述的需求文档进行处理需要自然语言处理技术的支撑，在本项目中主要体现为自然语言的信息抽取（Information Extraction）过程，作为信息抽取的子过程，分词、标注、分块等文本处理操作都可以结合机器学习技术，通过引入机器学习方法可以得到准确性更高、灵活性更好的文档处理工具。

b)自动插入标签：自动插入标签是项目中应用机器学习方法的目的，通过对自然语言描述的处理，在对描述内容达到一定理解程度的基础上可以由机器自行判断文档标签的位置，特别是在在处理规模较大的文档时可以拥有高效率。

d)根据标签提取信息：标签完成了对原始的自然语言描述的结构划分，显示出了需求的流程特点。在为文档插入标签后，GWT描述中的RUCM内容被标记出来，根据标签可以快速提取出文档中对应于RUCM描述的内容。

e)形成结构化的需求：在通过标签提取出信息后，提取出的信息遵循RUCM描述格式，在此基础上，可以将信息保存为RUCM描述的需求文档，获得结构化的需求。

二、领域定位

项目所属的具体领域在现实世界中有真实系统存在的，具体来说，项目所属的领域是软件工程中需求工程领域的需求描述环节，并且，根据项目的具体要求，需求文档的处理过程中，还会涉及到其他相关领域，例如，标签、AI、自然语言处理以及对软件提出不同具体要求的不同行业领域等。

对相关领域定位的具体分析如下。

软件工程-需求工程-需求描述环节：由于软件开发具有复杂性、一致性、可变性、不可见性的特征，软件开发往往面对着客户不满意、风险和成本问题、项目过程失控、无力管理团队等挑战。软件工程一直致力于探索软件开发问题的解决之道，将系统性的、规范化的、可定量的方法应用于软件的开发、运行和维护，即工程化应用到软件上。著名软件工程专家B.Boehm综合有关专家和学者的意见并总结了多年来开发软件的经验，于1983年在一篇论文中提出了软件工程的七条基本原理：（1）用分阶段的生存周期计划进行严格的管理。（2）坚持进行阶段评审。（3）实行严格的产品控制。（4）采用现代程序设计技术。（5）软件工程结果应能清楚地审查。（6）开发小组的人员应该少而精。（7）承认不断改进软件工程实践的必要性。其中，需求工程在软件工程中的重要性是毋庸置疑的。需求是人们要解决的某个问题或达到某种目的的需要，是系统或其组成部分为满足某种书面规定（合同， 标准， 规范等）所要具备的能力。需求将作为系统开发、 测试、验收、提交的正式文档依据。需求不只是软件过程的阶段之一，同时也涉及到软件工程的方方面面。本次项目的领域可以具体到需求工程的需求描述环节，即将原始的GWT需求描述转变为RUCM标准格式的需求描述。这种标准格式的描述可以便于人们对于需求的统一的理解，也可以用以后机器对于需求文档地进一步转化过程，如转化为模型或转化为设计等。

在需求描述环节，自然语言描述是最初被用来描述软件需求的方法，并且也是需求描述使用最广泛的方法，它易于表达，直观且通用，但是往往具有二义性，具有不同专业背景的读者对同一个需求描述可能会得到不同的理解。结构化描述要求按照一种标准方式来用自然语言表达需求，这种方式保留了自然语言描述的好处，包括表达能力和易懂性，同时对描述施加了一致性的约束。使用结构化的RUCM描述方法来描述系统需求，以表格形式突出了需求的主要内容，清晰地描述了各个需求间的关系。

标签：标签是无等级的关键字或分配某项物品的信息（如互连网书签，数字图像，或电脑档案）。这样的数据有助于说明一个项目，并允许它再次发现浏览或搜寻。根据系统，标签本身被非正式创作者或其他观察者选择。 在许多用户标记许多项目的网站上，标记的汇集成为分众分类法。本项目中的标签类似于html标签，即是对某一类文本进行定义的标签，标识出原始的需求描述中与RUCM表项相关的信息，是在分析理解文本的基础上对文本内容的划分。

AI：即人工智能（artificial intelligence），人工智能一般指通过普通计算机程序的手段实现人类智能技术。其核心问题包括建构能够跟人类似甚至超越的推理、知识、规划、学习、交流、感知、移动和操作物体的能力等。针对这些子领域中让机器学会“学习”的问题，派生而出了机器学习的方法，而机器学习已经成为了人工智能领域的一个重要的分支。机器学习理论主要是涉及和分析一些让计算机可以自动“学习”的算法，可分为监督性学习、非监督性学习和强化学习。根据项目要求，对标签的自动插入需要使用机器学习的方法。而机器学习的方法选取将直接影响到标签插入的效果，最终将决定对需求文档的处理结果。

自然语言处理：是人工智能和语言学领域的分支学科，研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言理解方向，主要目标是帮助机器更好理解人的语言，包括基础的词法、句法等语义理解，以及需求、篇章、情感层面的高层理解。在本项目中，给出的输入是自然语言写的文档，而对自然语言的处理有着种种困难，尤其是要消除歧义的问题，让程序正确理解需求文档。这就需要参考自然语言处理中对自然语言理解所采用的方法，来保证自然语言文档描述的需求能够被程序正确识别。

对软件提出不同具体要求的不同行业领域：需求会根据干系人的不同而不同，而干系人往往会由于所在领域的不同而不同。所以在具体的软件开发项目中，软件需求与约束会表现出较强的行业特性会与很多方面相关，如政府[机构](https://baike.baidu.com/item/%E8%A1%8C%E4%B8%9A%E5%88%86%E7%B1%BB/2847868?fr=aladdin#2_1)、医药卫生、交通运输、信息产业、生活服务、教育教学、游戏社交等，应用于不同行业的软件产品需要充分考虑与行业相适应的需求。

三、主要术语及解析

（1）软件需求相关术语

1.需求工程（Requirements engineering，RE）：指在工程设计过程中定义、记录和维护需求的过程。

2.（需求的）符合性（Conformity（of requirements））：需求规格说明书符合某些标准的程度。

3,。（需求的）一致性（Consistency（of requirements））：对一套需求的描述无前后矛盾程度。

4.约束，边界条件（Constraint）：一种需求，它限制解决方案在那些必要的满足给定的功能需要和质量需求之外的空间。

5.实体（Entity）：一般而言，实体是指一个元素或元素的集合，它可以表示任何可想象的产品，例如，一个系统、现实的一部分、一件东西、一个组织、-一个过程等。在实体关系建模中指单个对象，它具有一个标识，并且不依赖于另一个对象。

6.模型（Model）：一个存在的现实或一个将 要创建的现实的抽象表达。广义的模型是现有实体或将要创建的实体的表达，而实体表示现实的任何部分或任何其它可能的元素或现象，包括其它模型的组合。就模型而言，实体称为原件。在需求工程中，需求可用模型来说明。

7.功能需求：包括对系统应该提供的服务、如何对特殊输入做出反应，以及系统在特定条件下的行为的描述。在某些情况下，功能需求可能还需明确声明系统不应该做什么。

8.非功能需求：指不直接关系到系统向用户提供的具体服务的一类需求，或对系统提供的服务或功能给出的约束。包括时间约束、开发过程的约束和所受到的标准的约束。非功能需求经常适用于整个系统而不是个别的系统功能或服务。

①产品需求：这些需求定义或约束软件的行为。包括系统执行速度和内存消耗等性能需求，包括指定系统可以接受的出错率等系统可靠性需求，也包括信息安全性需求和可用性需求。

②机构需求：指广泛的系统需求，起源于客户所在的机构和开发者所在的机构中的政策和规定。包括过程标准，即机构中采用的过程标准；实现要求，如所采用的程序设计语言和设计方法；交付需求，即有关对产品及其文档交付的要求。

③外部需求：包括所有来自系统外部因素和开发过程的需求。如监管需求；立法需求和道德需求等。

9. 领域需求：系统的领域需求是从系统的应用城中得出的而不是从个别系统用户那里得到的。它们可能本身就是新的功能需求，也可能是约束现有的功能需求，或者是特别的计算必须如何进行的规定。

10.软件需求文档（软件需求描述或SRS）：是对系统开发者需要实现的功能的正式陈述。它应该包括系统的用户需求和一个详细的系统需求描述。

①GWT格式：即Given，When，Then。描述内容如下图所示。

②RUCM格式：即受限用例建模（Restricted use case modeling），如下图所示，包含必选的用例名、简述、前置条件、主次活动者、依赖、泛化、基本事件流信息和可选的特定分支流、全局分支流及有界分支流。

（2）机器学习相关术语解析

1．机器学习：机器学习是用数据或以往的经验，以此优化计算机程序的性能标准。机器学习目前已广泛应用于数据挖掘、计算机视觉、自然语言处理、生物特征识别、搜索引擎、证券市场分析、战略游戏和机器人等领域。

2. 深度学习：深度学习（deep learning）是机器学习的分支，是一种试图使用包含复杂结构或由多重非线性变换构成的多个处理层对数据进行高层抽象的算法。深度学习是机器学习中一种基于对数据进行表征学习的算法。深度学习的好处是用非监督式或半监督式的特征学习和分层特征提取高效算法来替代手工获取特征。

3. 监督学习：从给定的训练数据集中学习出一个函数，当新的数据到来时，可以根据这个函数预测结果。监督学习的训练集要求是包括输入和输出，也可以说是特征和目标。训练集中的目标是由人标注的。常见的监督学习算法包括回归分析和统计分类。

4.无监督学习：无监督学习与监督学习相比，训练集没有人为标注的结果。常见的无监督学习算法有生成对抗网络（GAN）、聚类。

5.半监督学习：介于监督学习与无监督学习之间。

6. 人工神经网络：人工神经网络（rtificial Neural Network，ANN），简称神经网络（Neural Network，NN）或类神经网络，在机器学习和认知科学领域，是一种模仿生物神经网络（动物的中枢神经系统，特别是大脑）的结构和功能的数学模型或计算模型，用于对函数进行估计或近似。神经网络由大量的人工神经元联结进行计算。大多数情况下人工神经网络能在外界信息的基础上改变内部结构，是一种自适应系统，通俗的讲就是具备学习功能。现代神经网络是一种非线性统计性数据建模工具。

（3）自然语言处理相关术语解析

1.自然语言处理：（natural language processing，NLP）是人工智能和语言学领域的分支学科。此领域探讨如何处理及运用自然语言；自然语言处理包括多方面和步骤，基本有认知、理解、生成等部分。自然语言认知和理解，让电脑把输入的语言变成有意思的符号和关系，然后根据目的再处理。

2. 中文分词：中文分词(Chinese Word Segmentation) 指的是将一个汉字序列切分成一个个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。在英文的行文中，单词之间是以空格作为自然分界符的，而中文只是字、句和段能通过明显的分界符来简单划界，唯独词没有一个形式上的分界符，虽然英文也同样存在短语的划分问题，不过在词这一层上，中文比之英文要复杂得多、困难得多。

3. 词性标注（Label）：基于机器学习的方法里，往往需要对词的词性进行标注。标注的目的是，表征词的一种隐状态，隐藏状态构成的转移就构成了状态转移序列。例如：苏宁易购/n 投资/v 了/u 国际米兰/n。其中，n代表名词，v代表动词，n,v都是标注。以此类推。

4．信息抽取或信息提取：Information Extraction，通常处理对象为非结构化文本，在一定程度分析和理解文本的基础上，从大量文字数据中自动抽取名称、事件等特定消息（Particular Information）。

5. 自动摘要：Automatic summarization，从一个或多个文本中自动摘取包含了原文中最重要信息的部分，通常用于提供已知领域的文章摘要，例如产生报纸上某篇文章之摘要。