1. **引言**
   1. **概述**

在近十几年来，计算机领域发展迅猛，个人计算机，终端互联设配也越来越普及，几乎是人人随时随地都能上网，在当今互联网这个巨大的生态圈里面，每年互联网用户产生的数据正以惊人的速度增涨。这些逐年增长的数据中绝大多数的数据都是非结构化的数据，这些数据主要存储的形式都是文本、网页、图片、视频、音频等非结构化的形式。结构化的数据它是经过了一定的存储格式，处理规则加以封装，所以它能够被计算机根据其存储的规则来进行识别，而非结构化的数据比如说文本数据，这些数据没有规则，格式，通常情况下它是不能被计算机直接识别的。当人们想对这些数据进行分析，统计，挖掘的时候，由于其不是结构化的数据，很难通过数据挖掘的技术对其进行处理。人工筛选的方法在这个大数据时代显然是天方夜谭，虽然说人工筛选的方法在上世纪80年代之前都是的确存在的，有需求就有市场这句话放在科研的领域也是行的通的，正因为有对非结构化文本识别和分析的需求，所以研究者们也想出了各种相关的方法。文本挖掘（Text Mining），图像识别，模式识别等一些相关技术就对分析和识别这一些非结构化的数据起到很好的作用。在此篇论文中我们主要讨论一下文本挖掘技术。

文本挖掘(Text Mining)是一个从非结构化文本信息中获取用户感兴趣或者有用的模式的过程。首先它会把非结构化的文本数据通过一定的方法转换成为结构化的数据，然后再从这些结构化的数据中分析提取出我们用户所感兴趣的模式和东西。文本挖掘技术是一项非常复杂的技术，它是一个多学科混杂的一门学科，其中包含有信息技术、机器学习、统计学、数据可视化、数据库技术、数据挖掘。

集这么多门学科于一身，它在我们这个大数据时代，信息化时代发挥了重要的作用，也是在信息学科领域中比较热门的学科之一。

在生物医学领域，借助着计算机科学技术也在快速的发展，每年发表在互联网上的生物医学文献资料和研究者们在实验室产生的实验数据也正以惊人的速度增长，产生的这些文献资料和实验数据大部分都是以文本的形式存储在计算机设配上面。所以也是非结构化的数据，不能被计算机直接识别。那么我在此篇论文里主要讨论针对文本形式存储的非结构化数据的处理以及信息的提取。

* 1. **选题的背景和意义**

基因和蛋白质一直都是生物医学工程方面的两大热点词汇，人类从刚开始发现它们到现在一直都在研究它们，并从未停止过对它们的研究和实验。经过科学家们的众多次探索和实验让我们对它们有了一定的了解和认识，并且也享受到了研究成果给人类带来的益处.蛋白质是由氨基酸排列组合生成的，人体氨基酸种类有20种，这20种氨基酸排列组合可以生成成千上万种的蛋白质，随着研究者们不断深入研究蛋白质和不断实验，发表在科学文献期刊上的文献数量也随之增加，这个数量正在指数级地增长，如此多的学术文献给研究者们带来了无比多的信息和知识，但是蛋白质的种类如此繁多，有关蛋白质的文献也就越多，这么多的科学文献，研究者是不可能全部阅读完的，这样一来数量如此庞大的科学文献对于研究者们来说是一种浪费，然而研究者们发现他们在阅读每一篇文献时，仅仅只是要获取其中的蛋白质的关键信息，但是他们不得不逐一整片阅读，这大大阻碍了他们的研究效率，所以研究者们迫切需要借用计算机大数据的手段来支持他们完成更深入的探索和研究。他们希望通过计算机的文本挖掘和文本分析来帮他们迅速定位到他们想要得到的蛋白质信息，所以文本挖掘技术迫切需要应用与生物医学文献挖掘上，这也是我本论文的核心研究的方向。

在如此庞大的生物医学文献库当中加之以文本挖掘技术进行适当分析和处理，处理的结果给研究人员提供了一个客观可靠的数据基础，这大大地减少人工检阅和人工处理等繁杂的工作，提高了研究者们的效率，同时这些数据基础上研究人员们能够更深一步的论证、假设、探索、实验。从而加快了生物医学领域的发展进程。

此外再说一点个人观点，人类文明几千年，计算机技术从上个世纪问世到现在不到一百年的时间里，取得了飞快的发展，如此同时各个行业在享受着计算机带来的高效率之下，也取得了飞快的发展，所以人类社会在过去的一百年里取得了空前地发展。计算机，它作为一个20世纪最伟大的发明之一，造福于人类，那么它有义务去帮助人类解决各行各业的问题，因为它最初被设计出来就是为了解决辅助人类分担一些繁琐，重复的工作，解放了人类的双手，充分发挥了人类的大脑。所以解决像这种生物医学领域，或者其他领域一些棘手的里上遇到的问题，计算机有义不容辞的义务。试问如果计算机学科不承担这样的义务，那还有哪门学科能解决呢？

1. **入门准备**

**2.1蛋白质以及翻译后修饰概述**

蛋白质（protein）是生命的物质基础，是有机大分子，是构成细胞的基本有机物，是生命活动的主要承担者。中学课本学习过，蛋白质是由氨基酸构成的，而氨基酸是由基因表达产生的。所以可以认为蛋白质是基因表达的产物。蛋白质翻译是蛋白质合成过程中的第一步，翻译是根据遗传密码的中心法则，将成熟的信使RNA分子（由DNA通过转录而生成）中“碱基的排列顺序”（核苷酸序列）解码，并生成对应的特定氨基酸序列的过程。此过程产生的蛋白质是没有活性的，想要具有活性，常常要进行一个系列的翻译后修饰和加工，从而使其具有一定的功能和特性。据研究统计，人类有50%~90%的蛋白质都在翻译后进行了修饰。蛋白质的翻译后的化学修饰一般是对蛋白质的附加一些生物化学官能团或是改变其结构来让它改变其化学性质，从而让它具有某项特性或功能。翻译后的蛋白质正因为有了修饰加工这一个步骤，就可能使相同的蛋白质具有完全不一样的功能和特性，也大大地增加的蛋白质的复杂性和多样性。所以这种类远远地超过了有天然20中氨基酸经过排列组合可生成的蛋白质的种类。所以，一个基因表达产生一个蛋白的说法就不存在了。据统计，人体类的可以编码合成蛋白质的基因组有两万多个，但是却可以表达出高达20万种不同功能的蛋白质。

到目前已经发现了超过400种的蛋白质翻译后修饰的方式。但是目前发现的蛋白质翻译后修饰的方式还是比较少的，还有大量的蛋白质修饰过程还是未知的。然而我们比较常见的修饰方式有磷酸化、甲基化、泛素化、乙酰化、亚硝酸化、糖基化、氧化等。

**2.2文本挖掘技术概述**

文本挖掘（Text Mining）是以计算语言学、统计数理分析为理论基础 ,结合机器学习和信息检 索技术 ,从文本数据中发现和提取独立于用户信息需求的文档集中的隐含知识。它是一个从文本信息描述到选取提取模式 ,最终形成用户可理解的信息知识的过程。它其实也是数据挖掘技术的一个分支学科，它和数据挖掘不同的是，数据挖掘技术主要是对结构化的数据进行挖掘处理，比如说关系表（Table），事务表等，这些表里面的数据（特征）在积累的时候就已经通过键值对（key-map）的方式做好了结构化的存储，所以计算机在解析数据的时候根据不同key去取相应的数据，所以结构化的数据是能被计算机所识别的，但是文本挖掘技术主要针对的是文本形式存储的非结构化数据，比如说新闻，网页，书籍等，这些文本数据是通常是由人类所撰写，文本中充满了人类语言规则和定义，这些规则因地区而异，比如说不同国家的语言不一样，计算机必须为每种语言去适应其语法规则，或者时间而异，比如说随着科研不断发展，新的名词可能被创造，这个时候必须让计算机学习人类创造的新的词语，由此可见计算机是无法直接识别这些非结构化的数据。

**3.3 需求以及需求分析**

需求最初是由中山大学生物工程学院的几位师姐提出来的，她们研究蛋白质翻译后修饰的时候会经常去Pubmed生物医学文献检索系统里面去找有关蛋白质翻译后的文献，比如想要找找有关蛋白质磷酸化修饰的文献，学习这一类修饰后的蛋白质的知识，她们可以在Pubmed文献检索系统里输入phosphorylation关键字，接着就可以找到几千或者上万篇有关磷酸化的一些文献。这些根据关键字检索到的文献的数量有点庞大，人工每一篇地去阅读显得有些吃力，但是她们发现她们每次阅读文献的时候都有一个共同点，就是只要找出文献里面的关键信息，比如说只要找出某某蛋白质被某某酶在某个位点上修饰了。这样问题就来了，能不能借助计算机来识别这些关键信息并找出这些关键信息呢？如果可以的话就不用通篇阅读了，大大地节省了时间，提高了效率。于是她们就提出了这样一个需求：通过计算机来识别并找出文献中的一些关键信息，关键信息就是蛋白质的名称，酶的名称，酶修饰蛋白质的位点。

接到这个需求的时候，给我的第一感觉是