数据挖掘竞赛题目

专利引用量预测

随着科技的发展，为了更好地保护产权人的智力劳动成果权益，知识产权制度应运而生并不断完善，知识产权保护意识也越来越深入人心。其中，我们所熟知的专利就是知识产权的一种重要形式。

专利，即专利权的简称。专利制度是让专利权人在法定期间（例如：20 年）内享有专利技术的排他权，使其享有商业上的特权利益，以鼓励其将知识公开分享。世界知识产权组织(WIPO)数据显示，2017 年中国申请的专利数量已超越日本排名全球第二，仅位居美国之后。WIPO 预计，中国将在三年内超越美国。WIPO 总干事高瑞(Francis Gurry)称：“中国利用国际专利系统的迅速增长表明，随着中国经济的快速转型，那里的创新者已瞄准国际市场，希望将他们的创意引入新市场”。可以说，专利作为一种知识产权保护的手段，在国际市场中扮演着极为重要的角色，受到了国内外越来越多的企业、机构或个人的广泛关注。

近年来，为了鼓励创新，专利数据越来越趋于开放。例如，USPTO 不仅公开了 1972 年以来美国所有授权专利的申请记录、授权记录、专利文本、专利权人等信息，而且提供了免费的查询和下载渠道（具体信息见 http://www.patentsview.org/download/），极大地方便了相关人员进行专利查询以及相关研究。

专利作为一个巨大的技术数据宝库，内容丰富、体量可观，不仅具有技术价值、经济价值，还具有法律价值。专利分析与挖掘可以帮助企业更加准确地抓住技术创新机遇，对技术创新成果进行全面的保护，培育和完善专利组合，因此一直备受企业关注。其中，专利的引用量预测就是一个极为重要的问题。同一时刻申请的专利，随着时间的推移，引用量常常是不同的，而高引用的专利往往被认为价值较高，更加值得企业或者个人继续进行保护。

请你利用 USPTO 公开数据集，针对专利引用量预测问题，结合自己所学的知识进行实践研究，从以下角度做出探讨。

问题 1：通常一个专利文档中存在大量的信息，包括申请时间、授权时间、授权时延（申请时间与授权时间之间的间隔）、权利项声明数量、参考文献数量、文档中词语数量、句子数量、图片数量、表格数量、专利所属类别和专利类别数量（常用的专利类别包括国际通用的联合分类体系（CPC 分类）、美国专利分类体系（UPC）等）。请讨论哪些因素可能会对专利未来（从申请时间起 10 年后）的引用量产生影响，如何从给定专利数据集中抽取这些因素，以及这些因素与专利引用量之间可能存在的关系。

问题 2：除了上述因素，讨论还有哪些因素会影响一个专利未来的引用量。

问题 3：结合以上讨论内容，建模模型，预测专利未来能够获得的总引用量。

问题 4：请思考对于该问题的建模依据。

- 为什么使用以上因素能够帮助预测专利的引用量？请给出几个你认为最重要的因素，并尝试解释这些因素是如何影响专利未来的引用量的；
- 你是如何思考该问题以及如何建立合适的模型的？请阐述一下建模思路，并尝试解释模型选择对于预测精度有什么影响？