# C题：网络媒体的流行度

最近大火的TVB热搜剧《新闻女王》再一次把新闻业带到了大众眼前。社交媒体已经成为了人们交流和获取信息的重要渠道。每天，数以亿计的文本内容被发布到各种社交媒体平台上，包括微博、Twitter、Reddit等。这些文本内容涵盖了各种主题，从娱乐、体育、科技到政治等。在如此海量的文本信息中，如何预测哪些内容会受到用户的欢迎，哪些内容会被忽略，成为了社交媒体公司和研究者关注的焦点。



在社交媒体时代，每个人都可以成为信息的发布者和传播者。大量的用户生成内容（UGC）使得新闻传播的流行度预测变得更加复杂。流行度不仅受到文本内容本身的影响，还受到发布者、受众以及时间因素的影响。因此，建立一个准确的数学模型对于社交媒体公司的内容推荐、广告投放和战略决策等方面都具有重要的意义。

作为一家专注于人工智能和大数据分析的公司，你被委托设计一个数学模型，以预测社交媒体上文本内容的流行度。你们公司有关项目的甲方帮忙弄到了一份数据，包括时长、标题长度、文本内容长度等很多属性的数据。每条数据对应一个新闻（以url的形式给出）。最终的share就是最后的传播效果。你们需要挖掘这份数据：

1. 对这份数据进行数据处理与统计的工作，指出这份数据的统计特征。其中，LDA是一种语义模型，它抽取了有关文本的五个维度的特征并存储在数据中。
2. 构建数学模型，对传播量进行准确预测，阐述你所使用的模型以及它的表现效果。
3. 根据对前面问题的分析，你认为什么样的新闻更容易受到大众的传播？这对于新闻行业有什么启示？
4. 请将你们的发现写成一封建议信送往当地的某个网络媒体公司（你们公司的甲方）。