如果你是一个初学者，你每完成一个新项目后自身能力都会有极大的提高，如果你是一个有经验的数据科学专家，你已经知道这里所蕴含的价值。
本文将为您提供一个网站/资源列表，从中你可以使用数据来完成你自己的数据项目，甚至创造你自己的产品。

一.如何使用这些资源?

如何使用这些数据源是没有限制的，应用和使用只受到您的创造力和实际应用。使用它们最简单的方法是进行数据项目并在网站上发布它们。这不仅能提高你的数据和可视化技能，还能改善你的结构化思维。另一方面，如果你正在考虑/处理基于数据的产品，这些数据集可以通过提供额外的/新的输入数据来增加您的产品的功能。所以，继续在这些项目上工作吧，与更大的世界分享它们，以展示你的数据能力!我们已经在不同的部分中划分了这些数据源，以帮助你根据应用程序对数据源进行分类。我们从简单、通用和易于处理数据集开始，然后转向大型/行业相关数据集。然后，我们为特定的目的——文本挖掘、图像分类、推荐引擎等提供数据集的链接。这将为您提供一个完整的数据资源列表。如果你能想到这些数据集的任何应用，或者知道我们漏掉了什么流行的资源，请在下面的评论中与我们分享。（部分可能需要翻墙）

二.由简单和通用的数据集开始

1.data.gov(https://www.data.gov/)
这是美国政府公开数据的所在地，该站点包含了超过 19 万的数据点。这些数据集不同于气候、教育、能源、金融和更多领域的数据。

2.data.gov.in(https://data.gov.in/)
这是印度政府公开数据的所在地，通过各种行业、气候、医疗保健等来寻找数据，你可以在这里找到一些灵感。根据你居住的国家的不同，你也可以从其他一些网站上浏览类似的网站。

3.WorldBank(http://data.worldbank.org/)
世界银行的开放数据。该平台提供 Open Data Catalog，世界发展指数，教育指数等几个工具。

4.RBI(https://rbi.org.in/Scripts/Statistics.aspx)
印度储备银行提供的数据。这包括了货币市场操作、收支平衡、银行使用和一些产品的几个指标。

5.Five ThirtyEight Datasets (https://github.com/fivethirtyeight/data)
Five Thirty Eight，亦称作 538，专注与民意调查分析，政治，经济与体育的博客。该数据集为 Five ThirtyEight Datasets 使用的数据集。每个数据集包括数据，解释数据的字典和 Five ThirtyEight 文章的链接。如果你想学习如何创建数据故事，没有比这个更好。

三.大型数据集

1.Amazon WebServices（AWS）datasets
(https://aws.amazon.com/cn/datasets/)
Amazon 提供了一些大数据集，可以在他们的平台上使用，也可以在本地计算机上使用。您还可以通过 EMR 使用 EC2 和 Hadoop 来分析云中的数据。在亚马逊上流行的数据集包括完整的安然电子邮件数据集，Google Booksn-gram，NASA NEX 数据集，百万歌曲数据集等。

2.Googledatasets
(https://cloud.google.com/bigquery/public-data/)
Google 提供了一些数据集作为其 Big Query 工具的一部分。包括 GitHub 公共资料库的数据，Hacker News 的所有故事和评论。

3.Youtubelabeled Video Dataset
(https://research.google.com/youtube8m/)
几个月前，谷歌研究小组发布了 YouTube 上的“数据集”，它由 800 万个 YouTube 视频 id 和 4800 个视觉实体的相关标签组成。它来自数十亿帧的预先计算的，最先进的视觉特征。

四.预测建模与机器学习数据集

1.UCI MachineLearning Repository
(https://archive.ics.uci.edu/ml/datasets.html)
UCI 机器学习库显然是最著名的数据存储库。如果您正在寻找与机器学习存储库相关的数据集，通常是首选的地方。这些数据集包括了各种各样的数据集，从像 Iris 和泰坦尼克这样的流行数据集到最近的贡献，比如空气质量和 GPS 轨迹。存储库包含超过 350 个与域名类似的数据集(分类/回归)。您可以使用这些过滤器来确定您需要的数据集。

2.Kaggle
(https://www.kaggle.com/datasets)
Kaggle 提出了一个平台，人们可以贡献数据集，其他社区成员可以投票并运行内核/脚本。他们总共有超过 350 个数据集——有超过 200 个特征数据集。虽然一些最初的数据集通常出现在其他地方，但我在平台上看到了一些有趣的数据集，而不是在其他地方出现。与新的数据集一起，界面的另一个好处是，您可以在相同的界面上看到来自社区成员的脚本和问题。

3.AnalyticsVidhya
(https://datahack.analyticsvidhya.com/contest/all/)
您可以从我们的实践问题和黑客马拉松问题中参与和下载数据集。问题数据集基于真实的行业问题，并且相对较小，因为它们意味着 2 - 7 天的黑客马拉松。

4.Quandl
(https://www.quandl.com/)
Quandl 通过起网站、API 或一些工具的直接集成提供了不同来源的财务、经济和替代数据。他们的数据集分为开放和付费。所有开放数据集为免费，但高级数据集需要付费。通过搜索仍然可以在平台上找到优质数据集。例如，来自印度的证券交易所数据是免费的。

5.Past KDDCups
(http://www.kdd.org/kdd-cup)
KDD Cup 是 ACM Special Interest Group 组织的年度数据挖掘和知识发现竞赛。

6.DrivenData
(https://www.drivendata.org/)
Driven Data 发现运用数据科学带来积极社会影响的现实问题。然后，他们为数据科学家组织在线模拟竞赛，从而开发出最好的模型来解决这些问题。

五.图像分类数据集

1.The MNISTDatabase
(http://yann.lecun.com/exdb/mnist/)
最流行的图像识别数据集，使用手写数字。它包括 6 万个示例和 1 万个示例的测试集。这通常是第一个进行图像识别的数据集。

2.Chars74K
(http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/)
这里是下一阶段的进化，如果你已经通过了手写的数字。该数据集包括自然图像中的字符识别。数据集包含 74,000 个图像，因此数据集的名称。

3.Frontal FaceImages
(http://vasc.ri.cmu.edu//idb/html/face/frontal_images/index.html)
如果你已经完成了前两个项目，并且能够识别数字和字符，这是图像识别中的下一个挑战级别——正面人脸图像。这些图像是由 CMU & MIT 收集的，排列在四个文件夹中。

4.ImageNet
(http://image-net.org/)
现在是时候构建一些通用的东西了。根据 WordNet 层次结构组织的图像数据库(目前仅为名词)。层次结构的每个节点都由数百个图像描述。目前，该集合平均每个节点有超过 500 个图像(而且还在增加)。

六.文本分类数据集

1.Spam – NonSpam
(http://www.esp.uem.es/jmgomez/smsspamcorpus/)
区分短信是否为垃圾邮件是一个有趣的问题。你需要构建一个分类器将短信进行分类。

2.TwitterSentiment Analysis
(http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/)

该数据集包含 1578627 个分类推文，每行被标记为 1 的积极情绪，0 位负面情绪。数据依次基于 Kaggle 比赛和 Nick Sanders 的分析。

3.Movie ReviewData
(http://www.cs.cornell.edu/People/pabo/movie-review-data/)
这个网站提供了一系列的电影评论文件，这些文件标注了他们的总体情绪极性(正面或负面)或主观评价(例如，“两个半明星”)和对其主观性地位(主观或客观)或极性的标签。

七.推荐引擎数据集

1.MovieLens
(https://grouplens.org/)
MovieLens 是一个帮助人们查找电影的网站。它有成千上万的注册用户。他们进行自动内容推荐，推荐界面，基于标签的推荐页面等在线实验。这些数据集可供下载，可用于创建自己的推荐系统。

2.Jester
(http://www.ieor.berkeley.edu/~goldberg/jester-data/)
在线笑话推荐系统。

八.各种来源的数据集网站

1.KDNuggets
(http://www.kdnuggets.com/datasets/index.html)
KDNuggets 的数据集页面一直是人们搜索数据集的参考。列表全面，但是某些来源不再提供数据集。因此，需要谨慎选择数据集和来源。

2.Awesome PublicDatasets
(https://github.com/caesar0301/awesome-public-datasets)
一个 GitHub 存储库，它包含一个由域分类的完整的数据集列表。数据集被整齐地分类在不同的领域，这是非常有用的。但是，对于存储库本身的数据集没有描述，这可能使它非常有用。

3.RedditDatasets Subreddit
(https://www.reddit.com/r/datasets/)
由于这是一个社区驱动的论坛，它可能会遇到一些麻烦(与之前的两个来源相比)。但是，您可以通过流行/投票来对数据集进行排序，以查看最流行的数据集。另外，它还有一些有趣的数据集和讨论。

九.结尾的话

我们希望这一资源清单对于那些想项目的人来说是非常有用的。这绝对是一个金矿，好好加以利用吧！
