**PS：**第一次参加公开课学习计划，有点激动！在此，特别感谢DataWhale组织的各位帅哥靓妹，愿意无私付出这么多贡献！

那么先进入正题！

# DataWhale零基础入门NLP-Task01赛题理解

**赛题简介**：

以自然语言处理为背景，对新闻文本进行分类。在赛程引导中，会接触NLP的预处理、模型构建、模型训练等知识点，完整地了解一遍NLP比赛流程。

赛题地址：<https://tianchi.aliyun.com/competition/entrance/531810/learn>

**赛题数据**：



该赛题的数据，建立在各类新闻基础上，但内容是新闻文本的压缩编码形式，应该可以理解为已经过初步预处理。这种形式的数据，居然也可以应用NLP，真是厉害！

赛题数据为新闻文本，并按照字符级别进行匿名处理。整合划分出14个候选分类类别：财经、彩票、房产、股票、家居、教育、科技、社会、时尚、时政、体育、星座、游戏、娱乐的文本数据。

赛题数据由以下几个部分构成：训练集20w条样本，测试集A包括5w条样本，测试集B包括5w条样本。

**评测标准**：

评价标准采用mean(F1score)，结果越大越好。

**解题思路**：

文本数据是一种非结构化数据，因此需要特征提取和分类模型两个部分。

DataWhale的小伙伴们给出了4种方案，在机器学习和深度学习都有体现，很赞！

**1. TF-IDF+分类器**

TF-IDF就是特征提取方法，然后分类器可以用SVM、LR、XGBOOST。够经典！

**2. FastText**

FastText是入门级词向量，可以快速构建分类器！（待了解）

**3.WordVec+深度学习分类器**

WordVec是进阶款的词向量，所以词向量应该是特征！深度学习分类器可以是TextCNN、TextRNN、BiLSTM。

**4.Bert词向量**

Bert是更高级更强大的词向量，所以可以从这个开始学了。待学习了解！