HW3

实现思路

1. 数据清洗：我去除了原文中的英文停用词，并用正则表达式匹配删除了 {{URL}}，{{@YouTube@}} 这样的无用信息，然后通过 `nltk` token 化以后得到清洗后的数据。
2. TF-IDF向量化：通过 `sklearn` 的 `TfidfVectorizer` 库提取 Tfidf 向量矩阵
3. KMeans聚类算法：通过调用 `sklearn` 的 `Kmeans` 库对 Tfidf 向量进行聚类算法，聚类后我选取分到同一类的样本中 Tfidf 均值最高的五个词作为这一组的关键词输出

结果分析

```
===== Clustering with k=2 =====
Cluster 1 Keywords: ['game', 'time', 'news', 'day', 'love']
Cluster 2 Keywords: ['official', 'album', 'video', 'music', 'new']
===== Clustering with k=3 =====
Cluster 1 Keywords: ['apple', 'official', 'new', 'video', 'music']
Cluster 2 Keywords: ['love', 'year', 'new', 'happy', 'day']
Cluster 3 Keywords: ['time', 'game', 'news', 'love', 'new']
===== Clustering with k=4 =====
Cluster 1 Keywords: ['like', 'day', 'happy', 'thank', 'love']
Cluster 2 Keywords: ['music', 'happy', 'time', 'day', 'news']
Cluster 3 Keywords: ['supporters', 'album', 'music', 'trump', 'new']
Cluster 4 Keywords: ['super', 'bowl', 'football', 'win', 'game']
```

分析如下：

1. k=2 时，**Cluster 1** 包含的关键词更偏向于日常生活和感情相关的内容（例如：game, time, day, love）。**Cluster 2** 包含的关键词更集中在媒体和音乐相关的内容（例如：official, album, video, music）。
2. k=3时，**Cluster 1** 继续集中在媒体和音乐方面，同时加入了科技品牌（apple）。**Cluster 2** 更加偏向于感情和时间相关的内容（例如：love, year, happy, day）。**Cluster 3** 结合了时间和新闻以及游戏和感情方面的内容（例如：time, game, news, love）。
3. k=4时，**Cluster 1** 关键词更集中于日常生活和感恩相关的情感（例如：like, day, happy, thank, love）。**Cluster 2** 集中在音乐和时间相关内容（例如：music, happy, time, day）。**Cluster 3** 涉及到政治支持者和音乐专辑（例如：supporters, album, music, trump）。**Cluster 4** 明显集中在体育赛事（例如：super, bowl, football, win, game）。
4. 总体来看，随着k值增加，聚类的细化程度也在增加，例如k=2时只提供了一个简单的划分，将关键词分为日常生活和感情 vs. 媒体和音乐，而到k=4时，分类更加细化，还增加了日常感情、音乐时间、政治支持以及体育赛事的独立聚类。

可见，聚类数量k和划分的粒度强相关，因此在解决实际问题时，应该根据要解决的问题和数据集本身的特征，选择合适的k。