博学　审问　慎思　明辨　笃行

基于聚类的弱监督文本分类

SUN YAT-SEN UNIVERSITY

中山大學
SUN YAT-SEN UNIVERSITY

陈和港

--2023/5/23

中山大學 SUN YAT-SEN UNIVERSITY

Weakly-Supervised Text Classification
弱监督文本分类

➢无标记文本

ID	Documents
D_1	I cheered for Lakers winning NBA.
D_2	I am sad that Heat lost.
D_3	Great news! Scientists discovered …
D_4	The new film is not satisfactory.
	……

➢弱监督信息

- 少量有标记文本

ID	Documents
D_1	I cheered for Lakers winning NBA.

Sports

- 类的名称

sports
arts
science

- 类的关键词

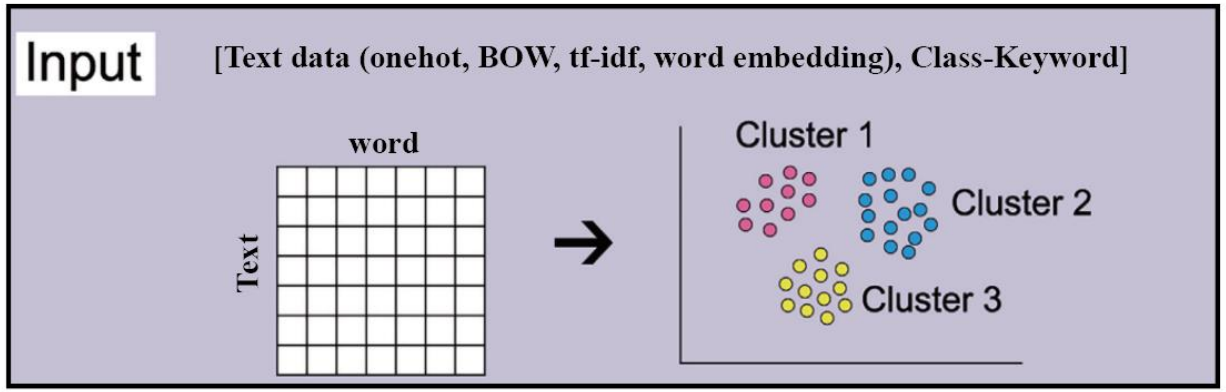
NBA, basketball, …

film, music, …

technology, computer, …

基于聚类的弱监督策略

第一步：聚类

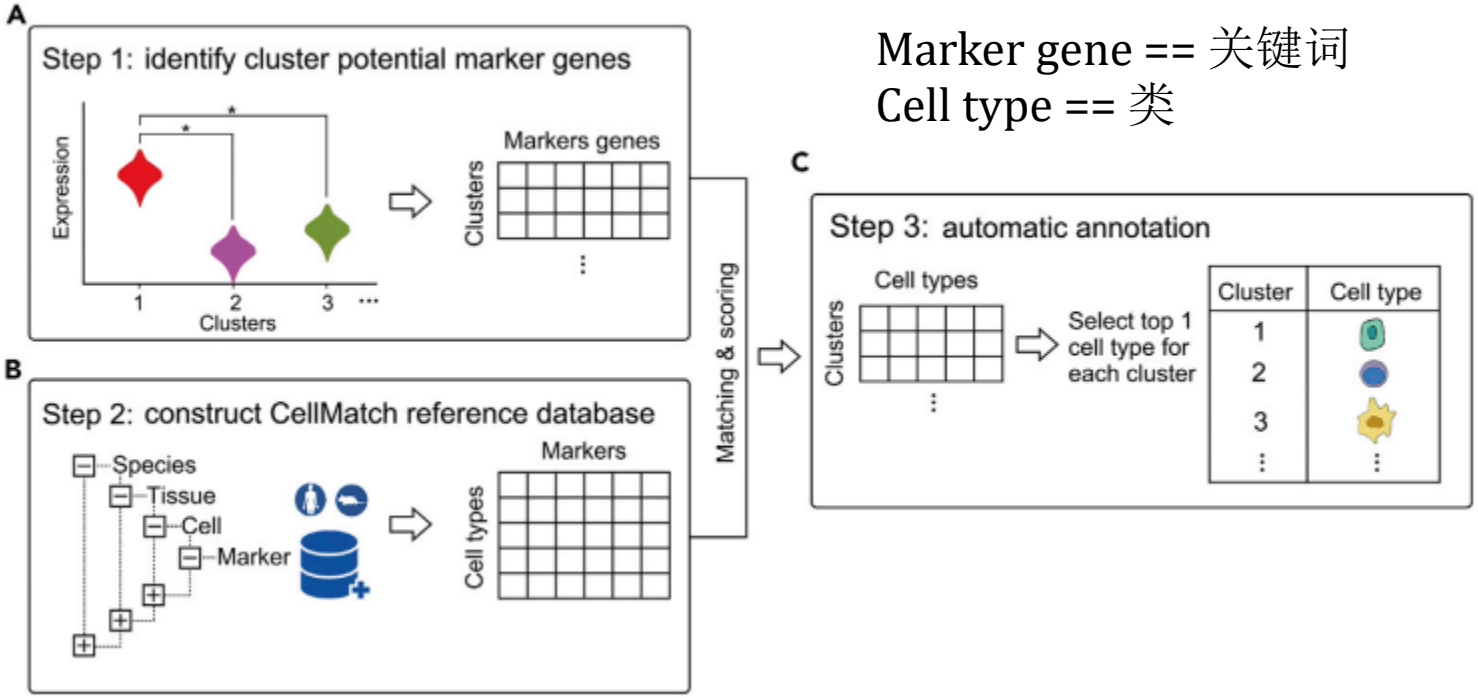


第二步：对于聚类出来的每个簇，找到其簇的关键词

第三步：根据簇的关键词和类的关键词（Class-Keyword），找到簇和类之间的关系

研究背景

ScCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data



Marker gene == 关键词
Cell type == 类

簇的关键词：简而言之就是在某个词在这个簇中的值和在其他簇中不同：

例如：词A，在簇i中的均值比簇j大一倍，我们可以认为相比于簇j来说，词A是簇i关键词。

$G_{i,1}$ ：簇i相对于簇1的所有关键词

$$M_i=G_{i,1}\cap G_{i,2}\cap G_{i,\dots}\cap G_{i,j}$$

M_i ：簇i相对于所有簇的关键词

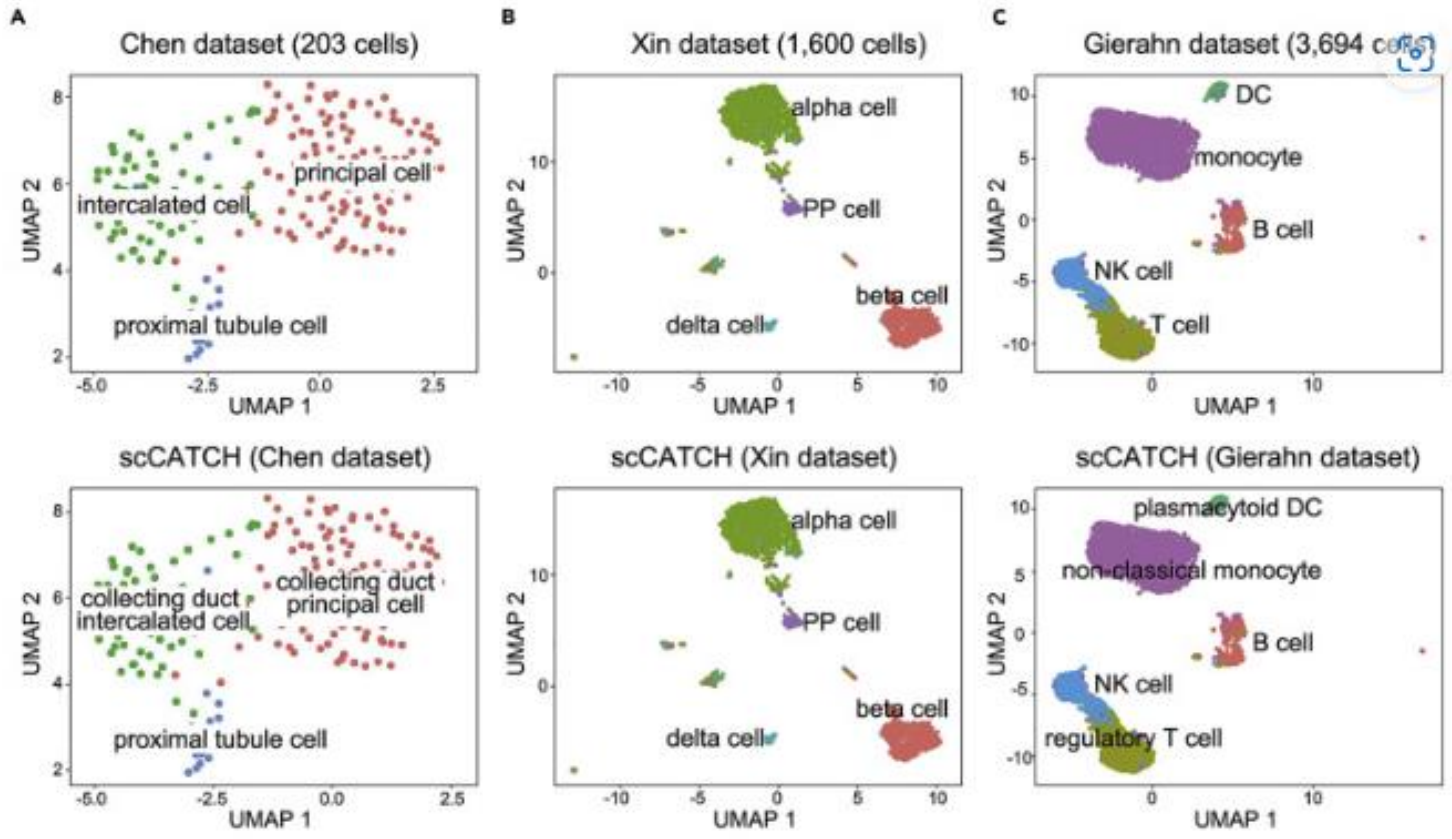
簇和类的关联：

$$ES_k=\sqrt{\frac{l_k}{l_k+1}\times\frac{g_k}{g_k+1}}$$

l_k ：类和簇总关键词数　　g_k ：类和簇相同关键词数

研究背景

聚类结果的低维可视化：通过降维算法将文本特征降到2维，然后将每个文本映射到图上，根据聚类结果标注每个簇的颜色（选择基于聚类策略的同学需实现）



未知类型识别（对于自训练的策略也同样适用）

由于存在多种未知类型时较难分析和预测，该任务由如下方式设置：

我们删除了一些文本类型（Class）对应的关键词（Keyword）（同学们在实现这部分是只需要将所有模型认为未知的文本类型定为Unknow即可）

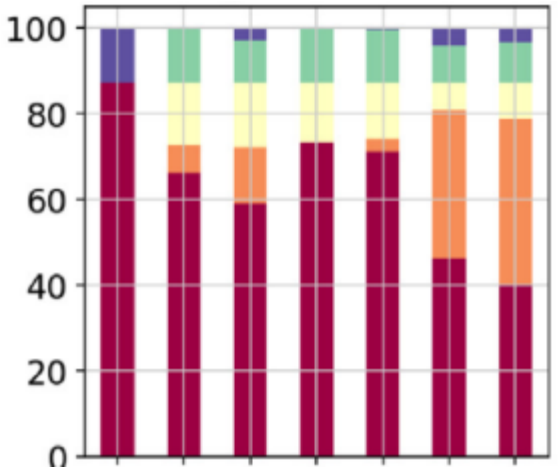
基于启发式拒绝策略：

1，当没有被以高置信度预测到某个类时

2，预测到每个类的概率类似

3，基于熵的拒绝 $$-\frac{1}{\log m}\sum_{i=1}^{m} q_i \log q_i > \gamma,$$

Correct (C)
Error (E)
Erroneously assigned (EA),
Correctly unassigned (CUA),
Erroneously unassigned (EUA)



期末项目：弱监督文本分类

- 基本要求：设计并实现一个弱监督文本分类模型
  - 报告模型在两个数据集上的性能(须使用同一套参数)
  - 使用至少两个评价方法
  - 推荐采用自训练算法（只是推荐，并不以此为评分标准）
  - 实现基于聚类策略的同学需要展示聚类图
- 进阶要求：未知类型识别
  - 数据集中有些文本的类不存在于Class-Keyword中，不实现进阶要求的同学在跑实验和计算指标时需删除这些文本
- 验收：第17、18周(暂定)自愿课堂展示，有加分
- 作业截止日期：第18周周一晚；6月19日，23:59
- 提交文件：按照先前作业的要求

研究背景

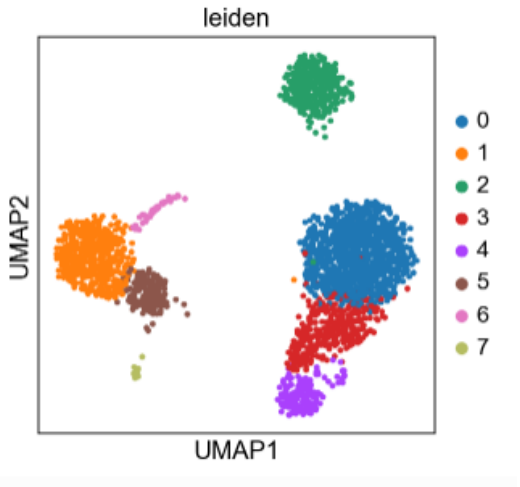
聚类和降维的策略可以使用sklearn集成的算法

https://scikit-learn.org/stable/modules/clustering.html#clustering
https://scikit-learn.org/stable/modules/decomposition.html#decompositions

Scanpy: 寻找簇的关键词和可视化

https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html

```
sc.pl.umap(adata, color=['CST3', 'NKG7', 'PPBP'])
```




```
sc.tl.rank_genes_groups(adata, 'leiden', method='t-test')
sc.pl.rank_genes_groups(adata, n=25
```