Homework 1
Due: Midnight, Feb 16st, 2022

Description:
We are going to use what we learned in the class to analyze real-world single-cell RNA-seq and single-cell ATAC-seq data. You will need to use the dimension reduction techniques introduced in the class. And you will need to use clustering algorithms to cluster the data. In the end, you will need to use manifold learn algorithm to visualize the clustering results.

Write everything in a jupyter notebook. Write your code in the "code" section in jupyter notebook and your answers to the questions in the "Markdown" section. You can use python package to do the following tasks. The only algorithm you need to program from starch is 2.2 and 2.3.

1. Data preparation (10points):
1.1 Download the single-cell RNA-seq data from the GEO website:
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126074
You need to download the following three files

Supplementary file	Size	Download	File type/resource
GSE126074_AdBrainCortex_SNAREseq_cDNA.barcodes.tsv.gz	48.3 Kb	(ftp)(http)	TSV
GSE126074_AdBrainCortex_SNAREseq_cDNA.counts.mtx.gz	28.3 Mb	(ftp)(http)	MTX
GSE126074_AdBrainCortex_SNAREseq_cDNA.genes.tsv.gz	93.7 Kb	(ftp)(http)	TSV

Hint: you can use scanpy.read_mtx to read in the data.
1.2 Download the single-cell ATAC-seq data from the GEO website:
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126074
You need to download the following three files

GSE126074_AdBrainCortex_SNAREseq_chromatin.barcodes.tsv.gz	54.1 Kb	(ftp)(http)	TSV
GSE126074_AdBrainCortex_SNAREseq_chromatin.counts.mtx.gz	77.7 Mb	(ftp)(http)	MTX
GSE126074_AdBrainCortex_SNAREseq_chromatin.peaks.tsv.gz	1.9 Mb	(ftp)(http)	TSV

Hint: you can use scanpy.read_mtx to read in the data.

2. Dimension reduction (40points):
2.1 Use the built-in pca function in scanpy to do dimension reduction on the scRNA-seq data.
2.2 Use the NMF to do dimension reduction on the scATAC-seq data.
2.3 Use the LDA to do dimension reduction on the scATAC-seq data.

3. Clustering (20points):

3.1 Clustering the raw scRNA-seq data and the one after dimension reduction (the result in 2.1).
3.2 Clustering the raw scATAC-seq data and the one after dimension reduction (the result in 2.2).
3.3 Clustering the scATAC-seq data after dimension reduction (the result in 2.3).

4. Visualization (15points):
4.1 Using Umap to plot the clustering results in 3.1.
4.2 Using Umap to plot the clustering results in 3.2 and 3.3.

5. Evaluation (15points):
The scRNA-seq data and scATAC-seq data are co-assayed. The cells in two datasets have correspondence. We need to check whether the clusters in scRNA-seq data (after dimension reduction) correspond to the clusters in scATAC-seq data (after dimension reduction). Compare the cluster overlap fraction similar to the plot below.