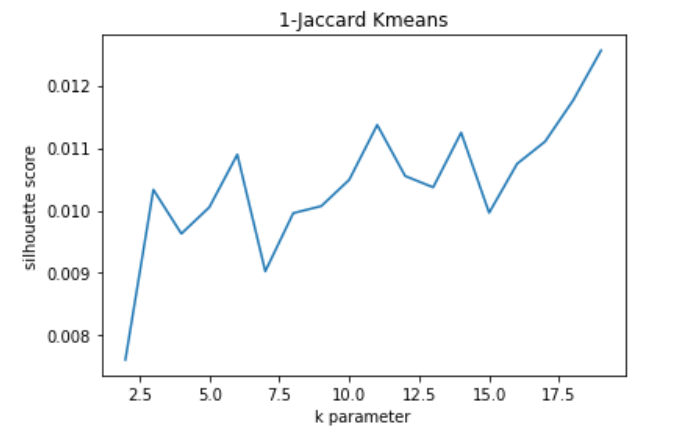
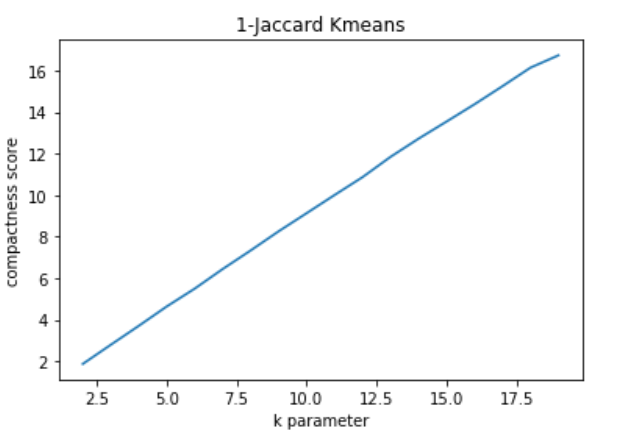
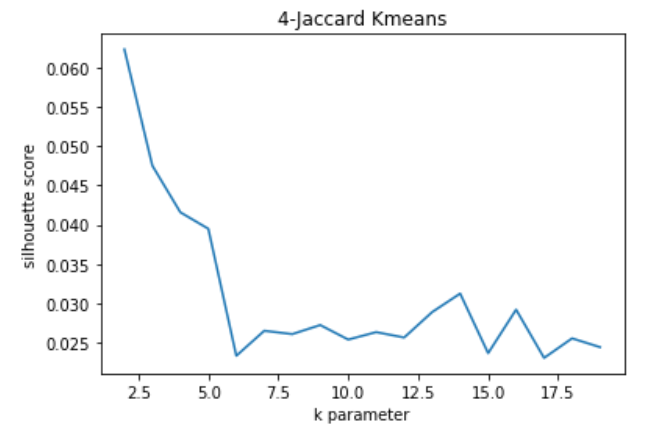
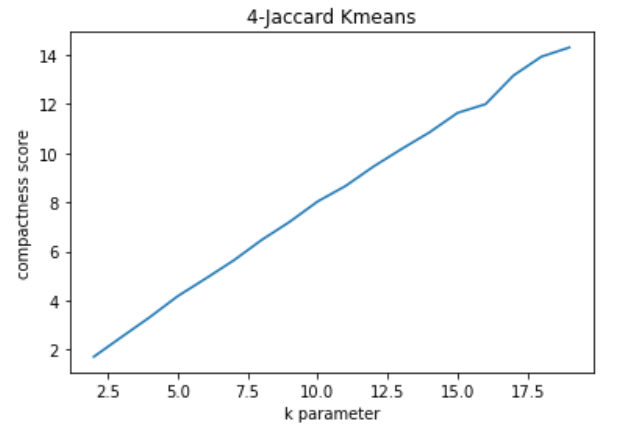
### 对于两种不同的Kmeans算法

选取不同的超参数K，得到的不同的评分如下



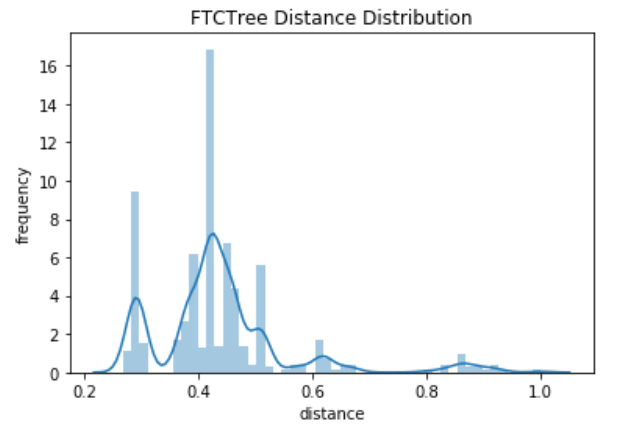


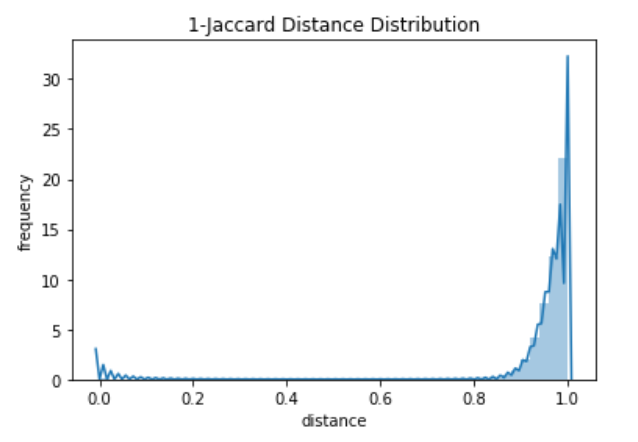


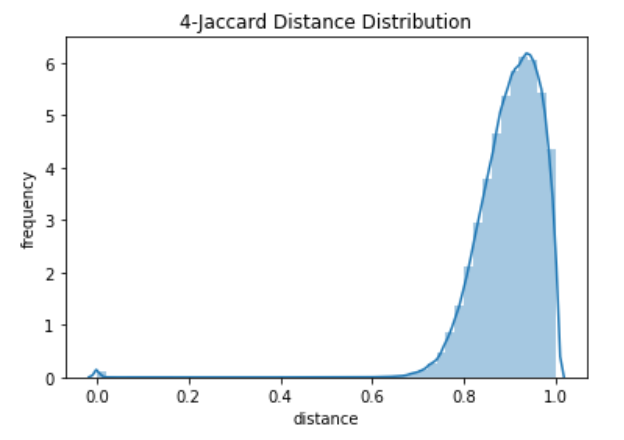


### 不同的距离度量和算法得出的距离分布比较

三种不同的算法得出的最终聚类结果的距离分布如下







### 讨论

1. 对KMeans算法
   1. Kmeans算法实质上对数据分布进行了隐式的假设：高斯分布假设，然而也会成为许多实际应用场景中的限制
   2. Kmeans算法同时也隐式假设了数据的先验概率分布：数据的平衡性，然而这在实际应用中也很难满足
   3. 超参K，比较难调，尤其是在对K的实际意义和限制条件比较模糊的时候
   4. 必须要在样本空间内求出均值
   5. 结果不一定是全局最优，只能保证局部最优
   6. 对噪声和离群点敏感
   7. 算法复杂度高
2. FTCTree算法
   1. 利用BIC准则进行簇的划分，不需要额外的参数来拟合
   2. 算法时间复杂度较高，但是相对于KMeans更低
   3. 距离分布比较均匀合理
   4. 非常适合稀疏矩阵
   5. 能很好地利用时间信息，对历史数据进行退化处理
3. Jaccard距离度量的KMeans算法
   1. 在1-Jaccard算法中，当K越大，compactness越低，而silhouette score却越高，无法很好地获得聚类效果
   2. 在4-Jaccard算法中，当K越大，compactness越低，silhouette score也越低，仍然无法很好地获得聚类效果
   3. 总体Compactness 和 Silhouette Score 都比较低，这是因为矩阵过于稀疏导致的
   4. 不能很好地利用时间信息