title

wbg231

January 2023

1 Introduction

- question 1:Locality-sensitive hashing makes MinHash faster by distributing the computation to be located on the same machine as the data being processed.
- that is false, local sensitive hashing has nothing to do with distributed computation
- MinHash signatures are generated by applying multiple independent hash functions to the elements of a set, and choosing the minimum value produced by all hash functions.
- false, you pick the min value for each element
- Adding a new permutation to an existing MinHash signature table can result in a smaller candidate set.
- false:
- How does the estimation of Jaccard similarity change when you use imperfect hashes instead of permutations? Do you expect the estimated similarity to be higher, lower, or the same? Why?
- if we are using imperfect hashes we will have more collisions so we will estimate the Jaccard similarity to be higher than we would using permutations.
- version control is part of reproducibility
- true
- nyu library recommends we use git for version control
- true
- nyu recommends sensitive data is stored on
- box

- hat do NYU research librarians recommend where to put contextual information about the analysis and data files?
- in a readme at the front of the directory
- To maximize reproducibility, a project folder should contain the following subfolders (select all that apply)
- data, src, results, docs, readme