

Autumn 2024 Exercise class: in 302 on September 24th

EXERCISES FOR COMPUTATIONAL TOOLS FOR DATA SCIENCE (02807) WEEK 4: SIMILAR ITEMS

References and Reading

1. Chapter 3 of Mining of Massive Data Sets, Jure Leskovec, Anand Rajaraman, and Jeff Ullman.

Exercise 1: Setup

Download the test data and template file similarity.py. Also install the mmh3 library. See https://pypi.org/project/mmh3/

Exercise 2: q-shingles

Implement a function shingle that takes an integer q and a string and produces a list of shingles, where each shingle is a list of q words.

Exercise 3: Minhashing Solve the following exercises.

- 3.1 Implement a minhash algorithm minhash that takes a list of shingles and a seed for the hash function mapping the shingles, and outputs the minhash. Feel free to use the listhash function in the template.
- **3.2** Extend the minhash algorithm to output k different minhashes in an array. Use different seeds for each minhash, e.g., $1, \ldots, k$.

Exercise 4: Signatures

Construct a function **signatures** that takes the **docs** dictionary and outputs a new dictionary consisting of document id's as keys and signatures as values.

Exercise 5: Jaccard Similarity

Implement a function jaccard that takes two document names and outputs the estimated Jaccard similarity using signatures.

Exercise 6: Find Similar Items

Implement a function similar that finds all pairs of documents whose estimated Jaccard similarity is ≥ 0.6 . Test your program for different values of k and q. Compare your results for most similar documents with your own visual impression of the similarity of files.

Exercise 7: Locality-Sensitive Hashing

Use locality-sensitive hashing to speed up your solution to the find similar item exercise.