Mapping n-grams to feature indices

If your dataset is small you can store {n-gram → feature index} in hash map.

But if you have a huge dataset that can be a problem

- Let's say we have 1 TB of texts distributed on 10 computers
- You need to vectorize each text
- You will have to maintain {n-gram → feature index} mapping
  - May not fit in memory on one machine
  - Hard to synchronize
- An easier way is hashing: {n-gram → hash(n-gram) % 2^{20}}
  - Has collisions but works in practice
  - sklearn.feature_extraction.text.**HashingVectorizer**
  - Implemented in **vowpal wabbit** library