Inverted Index Compression

- In general, leverage skewed distribution of values and use variable-length encoding
- TF compression
  - Small numbers tend to occur far more frequently than large numbers (why?)
  - Fewer bits for small (high frequency) integers at the cost of more bits for large integers
- Doc ID compression
  - “d-gap” (store difference): d1, d2-d1, d3-d2,…
  - Feasible due to sequential access
- Methods: Binary code, unary code, γ-code, δ-code, …