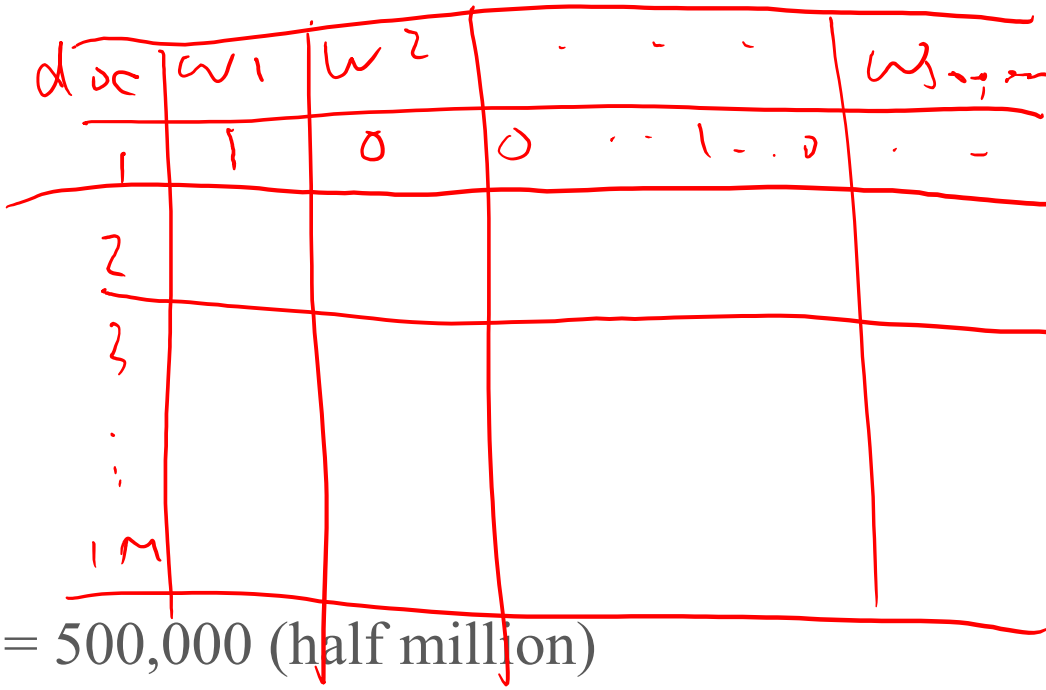
Sparse Matrix

School of Information Studies
Syracuse University

Sparse Matrix



N = 1 million docs

Each doc = 1,000 words

Each word = 6 bytes

Total corpus size = 6 GB

Total distinct words (types) = 500,000 (half million)

A matrix of 500,000 X 1 million:

- Half-a-trillion cells (0s or 1s): $5*10^{11}$
- 99.8% values are zeros
  - The number of 1s is up to 1 million*1,000 = $1*10^{9}$

Compressed Text Representation

Look-up dictionary of "index:word" pairs

- 1: "a," 2: "great," 3: "is," 4: "this," 5: "movie," 6: "terrible," …

Compressed storage of sparse vector:

- Each word that occurs in a document will be recorded as an "index:value" pair. For example:
  - Document: "this is a great movie"
  - Boolean vector: "1:1, 2:1, 3:1, 4:1, 5:1"