BLOOM FILTERS

SET DATA STRUCTURE

ONLY TWO OPERATIONS: IsIn(e)
Insert(e)

PROPERTIES:

- FAST (BOTH OPERATIONS CONSTANT TIME)
- SMALL (PRIMARY) MEMORY
- SMALL PROBABILITY THAT IsIn(e) WILL ANSWER YES EVEN IF e IS NOT IN THE SET

IMPLEMENTATION:

- BOOLEAN HASH TABLE WITH k INDEPENDENT HASH FUNCTIONS
- NO HANDLING OF COLLISIONS

APPLICATIONS:

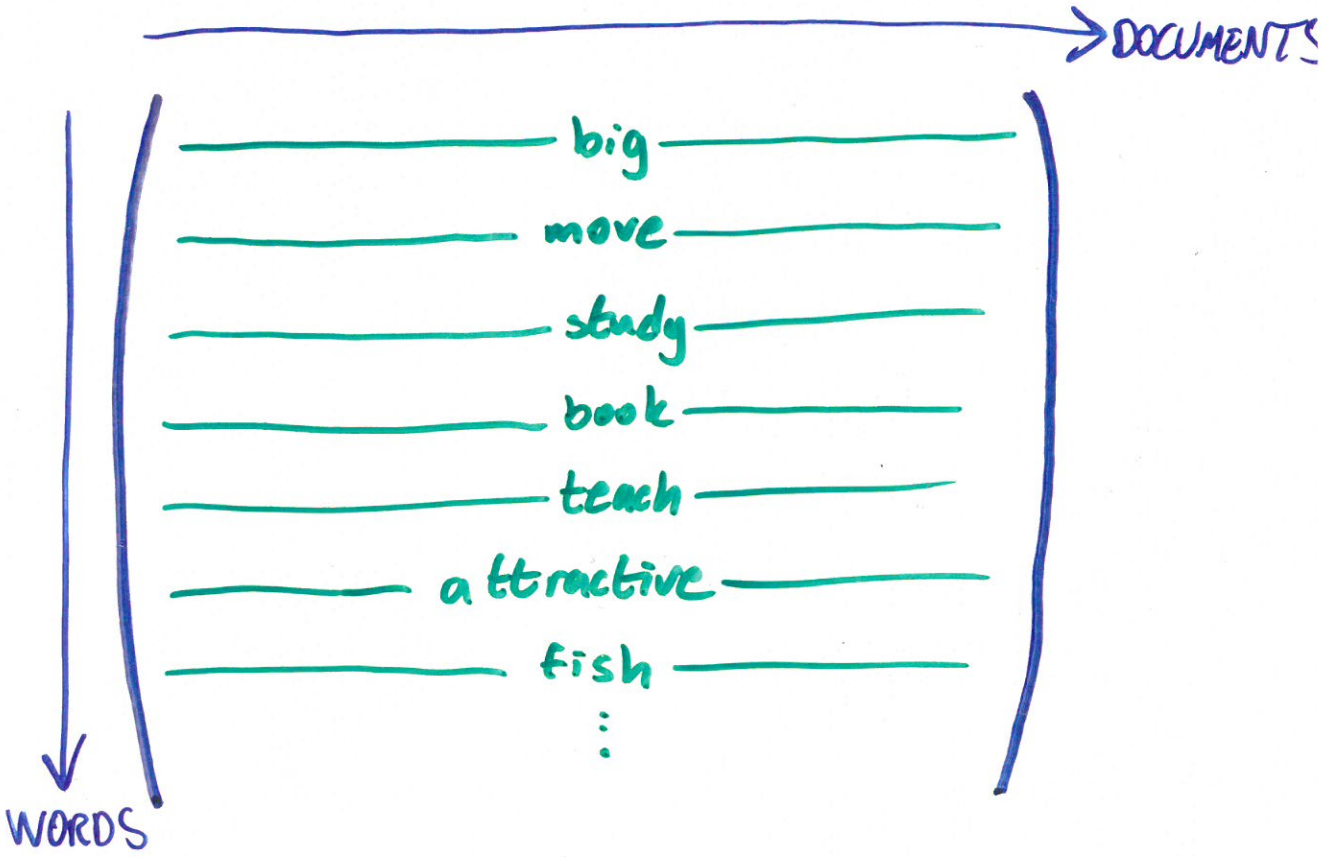
- WORD LIST FOR SPELL CHECKING
- IN SEARCH ENGINE: FAST CHECK WHETHER AN OBJECT MAY EXIST ON DISK (CASSANDRA DATABASE)
  SPOTIFY: SONGS

STATISTICAL LEXICAL SEMANTICS

THE VECTOR SPACE MODEL

> THE DISTRIBUTIONAL HYPOTHESIS:
>
> WORDS WITH SIMILAR MEANINGS TEND TO OCCUR IN SIMILAR CONTEXTS

STUDY THE WORD-DOCUMENT MATRIX:



THE ROWS ARE CONTEXT VECTORS REPRESENTING THE CONTEXTS IN WHICH A WORD HAS OCCURRED

+ VECTOR SPACES ARE MATHEMATICALLY WELL DEFINED AND WELL UNDERSTOOD

+ MAKES SEMANTICS COMPUTABLE (SIMILARITY IN TERMS OF VECTOR DISTANCES)

+ PURELY DESCRIPTIVE APPROACH TO SEMANTIC MODELLING

+ THE GEOMETRIC METAPHOR OF MEANING INTUITIVELY PLAUSIBLE, CONSISTENT WITH EMPIRICAL RESULTS FROM PSYCHOLOGICAL STUDIES

- MATRIX WILL SOON BECOME COMPUTATIONALLY INTRACTABLE

- SPARSE DATA PROBLEM

SOLUTION 1: DIMENSION REDUCTION TECHNIQUES

LSA - LATENT SEMANTIC ANALYSIS USING SVD - SINGULAR VALUE DECOMPOSITION

- COMPUTATIONALLY VERY COSTLY
- NEW DATA CAN'T BE ADDED
- REQUIRES THE WHOLE WORD-DOC-MATRIX

SEE CHAPTER 18 IN MANNING

SOLUTION 2: RANDOM INDEXING

INCREMENTAL WORD SPACE MODEL

[KANERVA 1988, 2000]

RANDOM INDEXING - RI

AN EFFICIENT INCREMENTAL WORD SPACE MODEL

1. ASSIGN A RANDOM INDEX VECTOR TO EACH WORD.

   THE INDEX VECTORS SHOULD BE
   - HIGH-DIMENSIONAL (E.G. 1800)
   - SPARSE
   - CONSIST OF 0s, +1s, −1s
   - NEARLY ORTHOGONAL

2. COMPUTE A CONTEXT VECTOR FOR EACH WORD X BY ADDING THE INDEX VECTORS FOR THE WORDS SURROUNDING X FOR EVERY OCCURRENCE OF X.

   IN ADVANCE CHOOSE THE SIZE OF THE CONTEXT WINDOW AND POSSIBLY WEIGHTS FOR DIFFERENT POSITIONS OF THE WINDOW.

3. THE CONTEXT VECTORS MAY NOW BE COMPARED USING A SIMILARITY MEASURE SUCH AS COSINE.

   WORDS WHOSE CONTEXT VECTORS ARE SIMILAR HAVE SOME SORT OF RELATION, SINCE THEY OCCUR IN SIMILAR CONTEXTS.