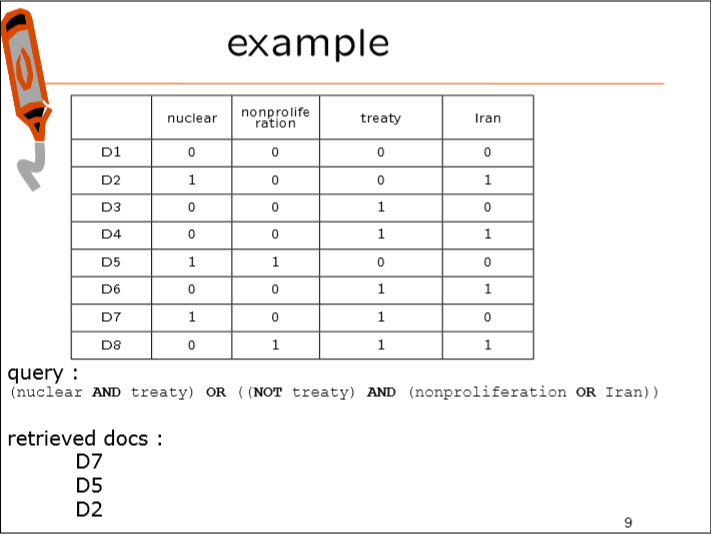
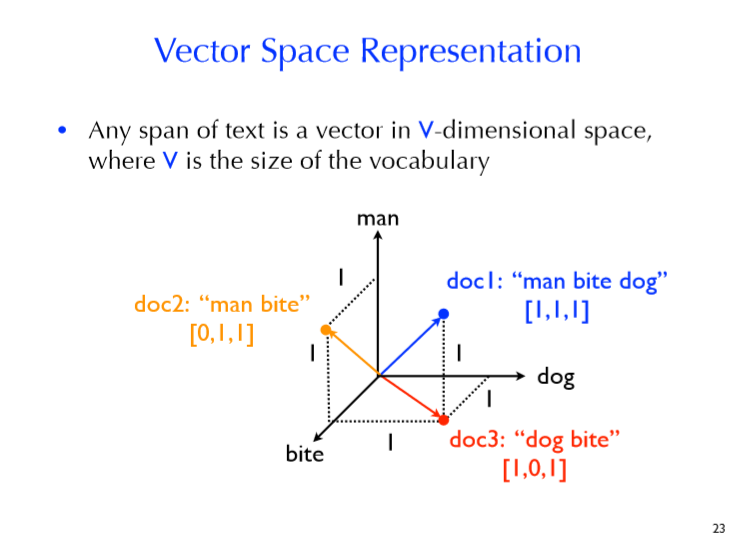
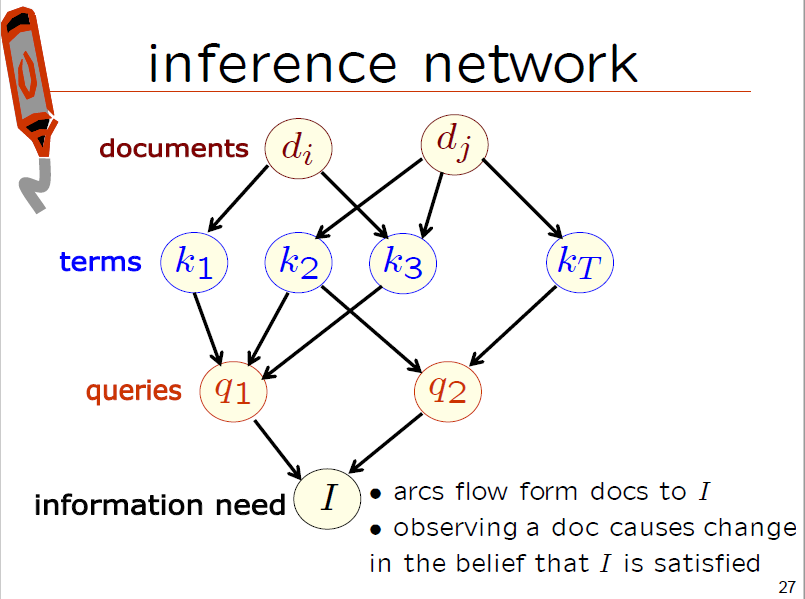
1. Boolean IR Model:
   1. Boolean model is most common exact-match model
      1. queries are logic expressions with document features as operands
      2. In Boolean model, retrieved documents could be ranked or not ranked
         1. Exact-match:
            1. query specifies precise retrieval criteria
            2. every document either matches or fails to match query
            3. result is a set of documents
   2. Boolean queries
      1. Used by Boolean model
   3. Use “Pure” Boolean operators: AND, OR, AND-NOT
   4. Most systems have proximity operators
   5. Most systems support simple regular expressions as search terms to match spelling variants



1. Vector Space Model:
   1. Using vector space to represent the appearing words in different document
   2. The inner product:
      1. Multiply corresponding components and then sum those products
      2. Using a binary representation, the inner product corresponds to the number of terms appearing (at least once) in both spans of text
      3. the inner-product favors long documents
   3. Cosine Similarity
      1. The numerator is the inner product
      2. The denominator is the product of the two vector-lengths
      3. Ranges from 0 to 1 (equals 1 if the vectors are identical)
   4. Term-Frequency
      1. To figure how important a term is in the document
   5. Inverse Document Frequency
      1. Log(number of document in collection/ number of documents in which term t appears)
      2. It is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.
   6. Summery
      1. Any text can be seen as a vector in V-dimensional space
      2. Rank documents based on their cosine similarity to query
      3. If a document is similar to the query, it is likely to be relevant



1. Classical Possibility IR:
   1. Compare query to documents base on their relevance
   2. Rank according to probability of document being relevant w.r.t. information need
   3. Bayes’ Rule
2. Language Models:
   1. The goal of a language model is to assign a probability to a sequence of words by means of a probability distribution
   2. Having a way to estimate the relative likelihood of different phrases is useful in many natural language processing applications
   3. Language modeling is used in speech recognition, machine translation, part-of-speech tagging, parsing, handwriting recognition, information retrieval
   4. Estimating a document’s language model:
      1. tokenize/split the document text into terms
      2. count the number of times each term occurs (tft,D)
      3. count the total number of term occurrences (ND)
      4. assign term t a probability equal to: tft,D/ND
3. Inference Networks:
   1. Similar to Bayesian Networks

Elasticsearch: use TF-IDF model

Tf-idf: term frequency-inverse document frequency, numerical statistics of how important a word is

Wumpus:

Indri:

Set-theory models: represent documents as sets of words or phrases

Algebraic models: represent documents and queries usually as vector matrices or tuples

Probabilistic models: treat the process of document retrieval as a probabilistic inference.