Machine Learning

Natural Language Processing

IBM's Watson computer

- Watson is an open-domain question-answering system.
- In 2011 Watson beat the two highest ranked human players in a nationally televised two-game Jeopardy! match.

http://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=6177717

Modeling document similarity

- Vector space models
- Latent semantic indexing

Vector space models

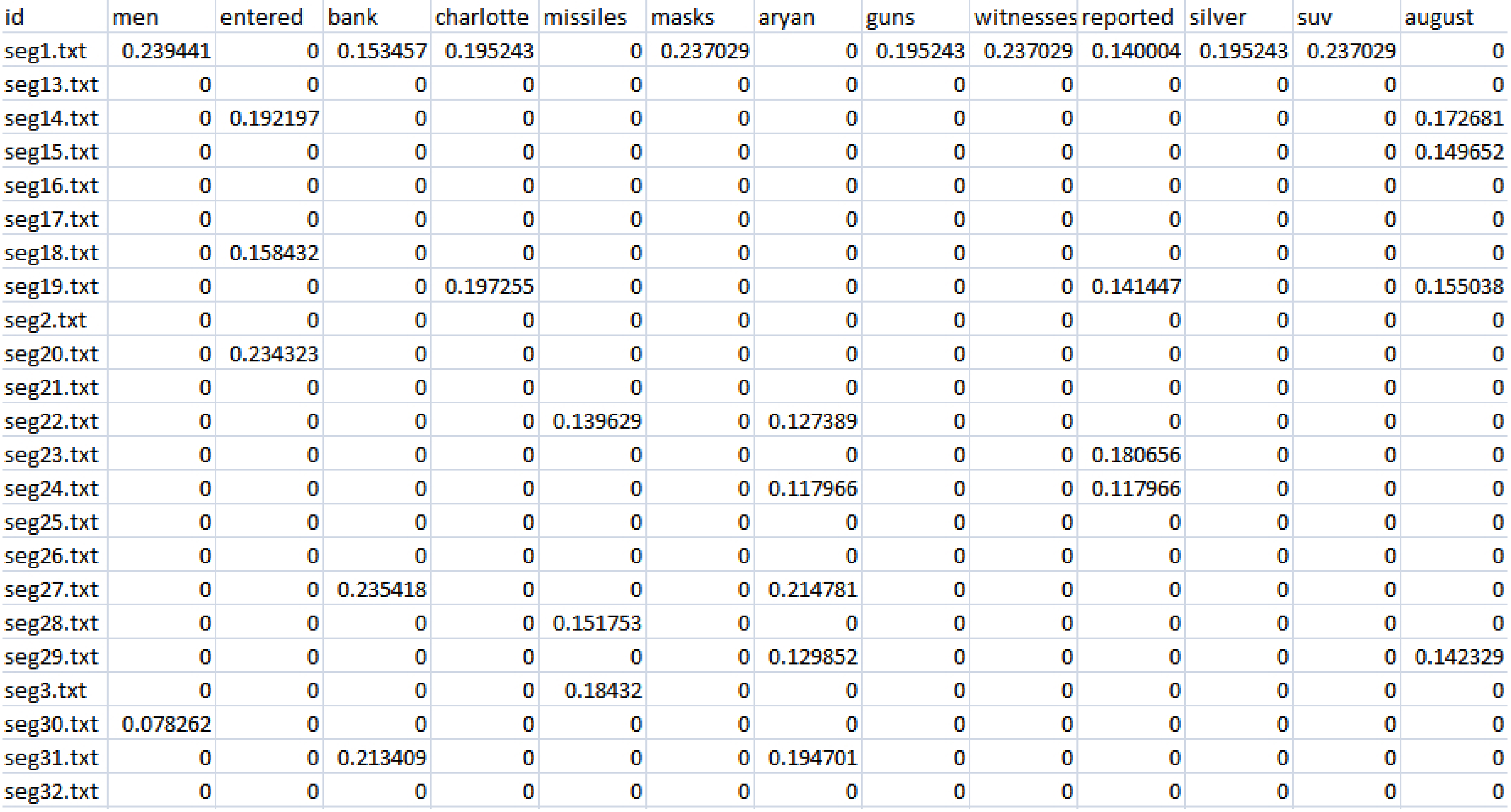
- Vector of features for each document
  - Word frequencies (usually weighted)
  - Meta attributes, e.g. title, URL, PageRank
- Can use vector as document descriptor for classification tasks
- Can measure document relatedness by cosine similarity of vectors
  - Useful for clustering, ranking tasks

Vector space models

- Term frequency-inverse document frequency (tf-idf)
  - Very widely used model
  - Each feature represents a single word (term)
  - Feature value is product of:
    - tf = proportion of counts of that term relative to all terms in document
    - idf = log(total number of documents /
      number of documents that contain the term)

Vector space models

- Example of tf-idf vectors for a collection of documents



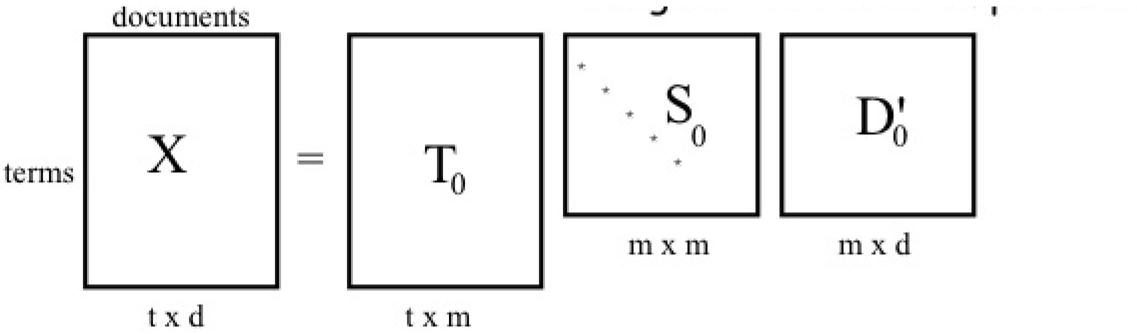
id	men	entered	bank	charlotte	missiles	masks	aryan	guns	witnesses	reported	silver	suv	august
seg1.txt	0.239441	0	0.153457	0.195243	0	0.237029	0	0.195243	0.237029	0.140004	0.195243	0.237029	0
seg13.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg14.txt	0	0.192197	0	0	0	0	0	0	0	0	0	0	0.172681
seg15.txt	0	0	0	0	0	0	0	0	0	0	0	0	0.149652
seg16.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg17.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg18.txt	0	0.158432	0	0	0	0	0	0	0	0	0	0	0
seg19.txt	0	0	0	0.197255	0	0	0	0	0	0.141447	0	0	0.155038
seg2.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg20.txt	0	0.234323	0	0	0	0	0	0	0	0	0	0	0
seg21.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg22.txt	0	0	0	0	0.139629	0	0.127389	0	0	0	0	0	0
seg23.txt	0	0	0	0	0	0	0	0	0	0.180656	0	0	0
seg24.txt	0	0	0	0	0	0	0.117966	0	0	0.117966	0	0	0
seg25.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg26.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg27.txt	0	0	0.235418	0	0	0	0.214781	0	0	0	0	0	0
seg28.txt	0	0	0	0	0.151753	0	0	0	0	0	0	0	0
seg29.txt	0	0	0	0	0	0	0.129852	0	0	0	0	0	0.142329
seg3.txt	0	0	0	0	0.18432	0	0	0	0	0	0	0	0
seg30.txt	0.078262	0	0	0	0	0	0	0	0	0	0	0	0
seg31.txt	0	0	0.213409	0	0	0	0.194701	0	0	0	0	0	0
seg32.txt	0	0	0	0	0	0	0	0	0	0	0	0	0

Vector space models

- Vector space models cannot detect:
  - Synonymy – multiple words with same meaning
  - Polysemy – single word with multiple meanings (e.g. play, table)

Latent semantic indexing

- Aggregate document term vectors into matrix X
  - Rows represent terms
  - Columns represent documents
- Factorize X using singular value decomposition (SVD)
  - $X = T \cdot S \cdot D^T$

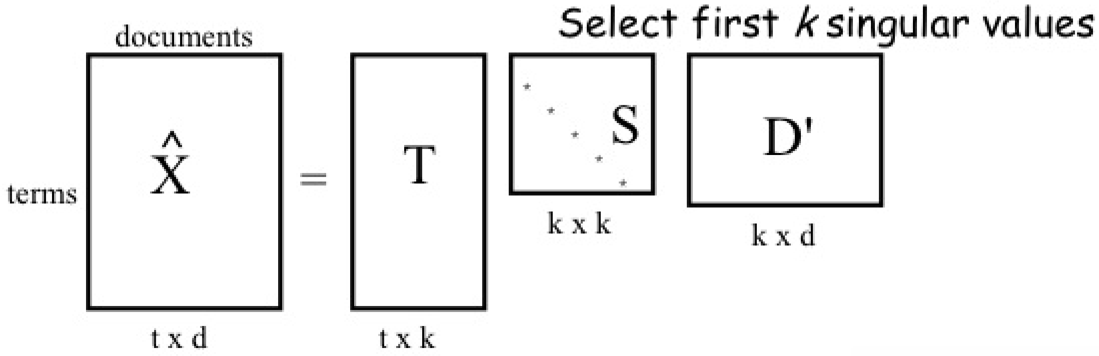


Latent semantic indexing

- Factorize X using singular value decomposition (SVD)
  - $X = T \cdot S \cdot D^T$
    - Columns of T are orthogonal and contain eigenvectors of XX^T
    - Columns of D are orthogonal and contain eigenvectors of X^TX
    - S is diagonal and contains singular values (analogous to eigenvalues)
    - Rows of T represent terms in a new orthogonal space
    - Rows of D represent documents in a new orthogonal space
  - In the new orthogonal spaces of T and D, the correlations originally present in X are captured in separate dimensions
    - Better exposes relationships among data items
    - Identifies dimensions of greatest variation within data

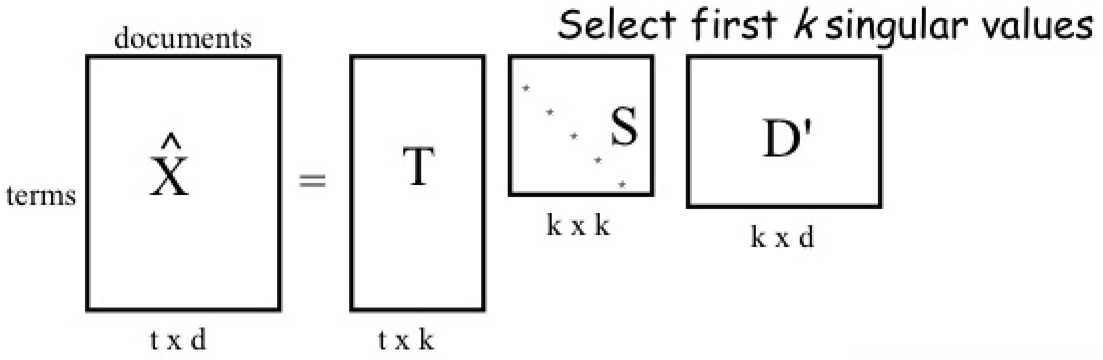
Latent semantic indexing

- Dimensional reduction with SVD
  - Select k largest singular values and corresponding columns from T and D
  - $X_k = T_k \cdot S_k \cdot D_k^T$ is a reduced rank reconstruction of full X



Latent semantic indexing

- Dimensional reduction with SVD
  - Reduced rank representations of terms and documents referred to as “latent concepts”
  - Compare documents in lower dimensional space
    - classification, clustering, matching, ranking
  - Compare terms in lower dimensional space
    - synonymy, polysemy, other cross-term semantic relationships



Latent semantic indexing

- Unsupervised
- May not learn a matching score that works well for a task of interest

Major tasks in NLP (Wikipedia)

- For most tasks on following slides, there are:
  - Well-defined problem setting
  - Large volume of research
  - Standard metric for evaluating the task
  - Standard corpora on which to evaluate
  - Competitions devoted to the specific task

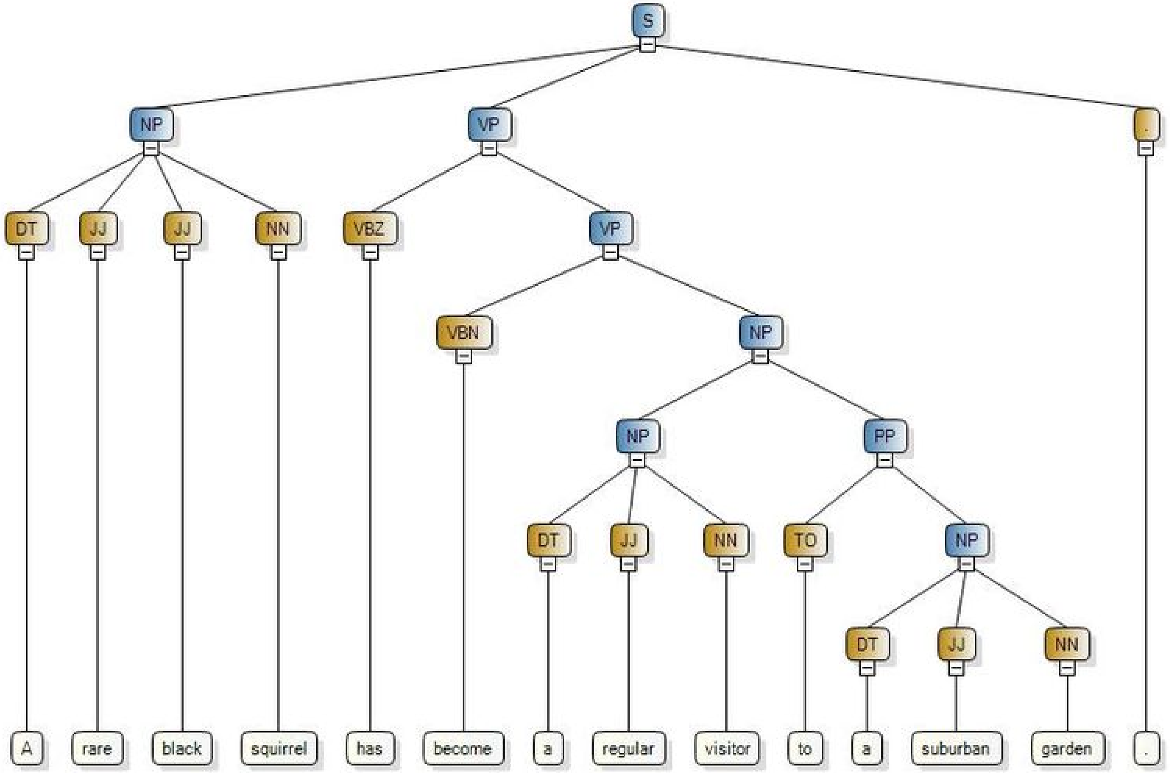
http://en.wikipedia.org/wiki/Natural_language_processing

Machine translation

- Automatically translate text from one human language to another.
  - This is one of the most difficult problems, and is a member of a class of problems colloquially termed "AI-complete", i.e. requiring all of the different types of knowledge that humans possess (grammar, semantics, facts about the real world, etc.) in order to solve properly.

Parse tree (grammatical analysis)

- The grammar for natural languages is ambiguous and typical sentences have multiple possible analyses. For a typical sentence there may be thousands of potential parses (most of which will seem completely non-sensical to a human).



Part-of-speech tagging

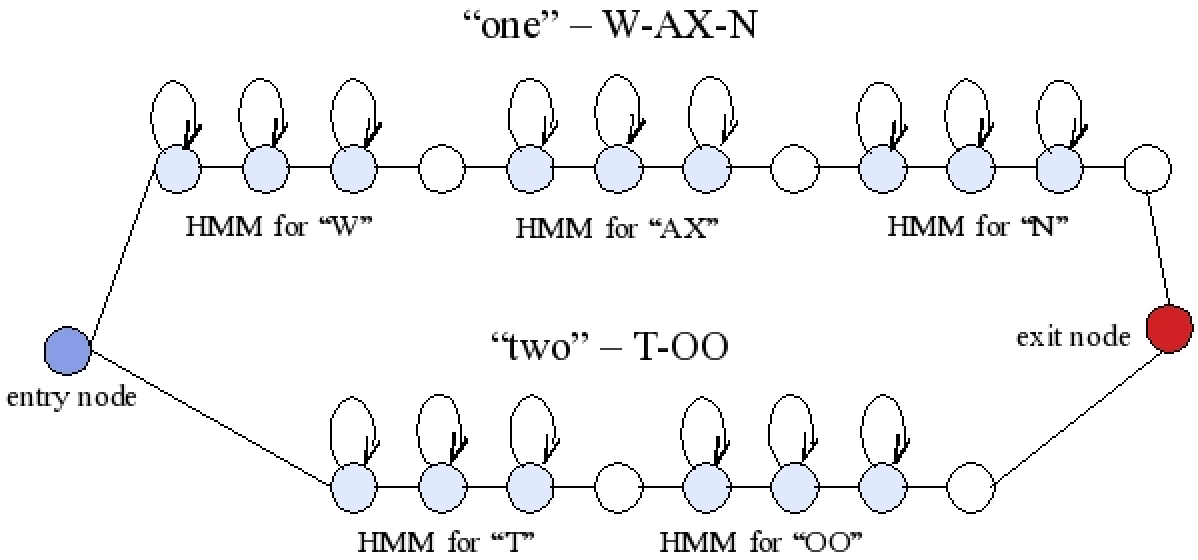
- Given a sentence, determine the part of speech for each word. Many words, especially common ones, can serve as multiple parts of speech. For example, "book" can be a noun or verb, and "out" can be any of at least five different parts of speech.

Speech recognition

- Given a sound clip of a person or people speaking, determine the textual representation of the speech.
  - Another extremely difficult problem, also regarded as "AI-complete".
  - In natural speech there are hardly any pauses between successive words, and thus speech segmentation (separation into words) is a necessary subtask of speech recognition.
  - In most spoken languages, the sounds representing successive letters blend into each other in a process termed coarticulation, so the conversion of the analog signal to discrete characters can be a very difficult process.

Speech recognition

- Hidden Markov model (HMM) for phoneme extraction



https://www.assembla.com/code/sonido/subversion/node/blob/7/sphinx4/index.html

Sentiment analysis

- Extract subjective information, usually from a set of documents like online reviews, to determine "polarity" about specific objects.
  - Especially useful for identifying trends of public opinion in the social media, for the purpose of marketing.

Information extraction (IE)

- Concerned with extraction of semantic information from text.
  - Named entity recognition
  - Coreference resolution
  - Relationship extraction
  - Word sense disambiguation
  - Automatic summarization
  - etc.

Named entity recognition

- Given a stream of text, determine which items in the text map to proper names, such as people or places, and what the type of each such name is (e.g. person, location, organization).
  - In English, capitalization can aid in recognizing named entities, but cannot aid in determining the type of named entity, and in any case is often insufficient. For example, the first word of a sentence is also capitalized, and named entities often span several words.
  - German capitalizes all nouns.
  - French and Spanish do not capitalize names that serve as adjectives.
  - Many languages (e.g. Chinese or Arabic) do not have capitalization at all.

Coreference resolution

- Given a chunk of text, determine which words ("mentions") refer to the same objects ("entities").
  - *Anaphora resolution* is a specific example of this task, concerned with matching up pronouns with the nouns or names that they refer to.
  - The more general task of coreference resolution also includes identifying so-called "bridging relationships" involving referring expressions.
    - For example, in a sentence such as "He entered John's house through the front door", "the front door" is a referring expression and the bridging relationship to be identified is the fact that the door being referred to is the front door of John's house (rather than of some other structure that might also be referred to).

Relationship extraction

- Given a chunk of text, identify the relationships among named entities (e.g. who is the wife of whom).

Word sense disambiguation

- Many words have more than one meaning; we have to select the meaning which makes the most sense in context.
  - For this problem, we are typically given a list of words and associated word senses, e.g. from a dictionary or from an online resource such as WordNet.

Automatic summarization

- Produce a readable summary of a chunk of text. Often used to provide summaries of text of a known type, such as articles in the financial section of a newspaper.

Natural language generation

- Convert information from computer databases into readable human language.

Natural language understanding

- Convert chunks of text into more formal representations such as first-order logic structures that are easier for computer programs to manipulate.
  - Natural language understanding involves the identification of the intended semantic from the multiple possible semantics which can be derived from a natural language expression which usually takes the form of organized notations of natural languages concepts. Introduction and creation of language metamodel and ontology are efficient however empirical solutions. An explicit formalization of natural languages semantics without confusions with implicit assumptions such as closed world assumption (CWA) vs. open world assumption, or subjective Yes/No vs. objective True/False is expected for the construction of a basis of semantics formalization.

Optical character recognition (OCR)

- Given an image representing printed text, determine the corresponding text.

Word segmentation

- Separate a chunk of continuous text into separate words.
  - For English, this is fairly trivial, since words are usually separated by spaces. However, some written languages like Chinese, Japanese, and Thai do not mark word boundaries in such a fashion, and in those languages text segmentation is a significant task requiring knowledge of the vocabulary and morphology of words in the language.

Speech processing

- Speech recognition
- Text-to-speech

Automated essay scoring