Text Analytics Using Latent Semantic Analysis

John Martin
Small Bear Technologies, Inc.
<John.Martin@SmallBearTechnologies.com>

www.SmallBearTechnologies.com
Overview

- Text Analytics
- Need for automated methods
- What is LSA
- How LSA works
- Applications of LSA
- Misconceptions
- Conclusion - Q&A
What is Text Analytics?

A set of linguistic, statistical, and machine learning techniques that model and structure the information content of textual sources for business intelligence, exploratory data analysis, research, or investigation (Wikipedia)

Text Analytics $\approx$ Text Mining
Text Analytics

- Derive meaning from (textual) data sources
- Structured data
  - Fixed format
  - Known attributes
- Unstructured data
  - Natural language
Unstructured Text

- News feeds
- Call center logs
- E-mail traffic
- Surveys
- Social network postings
- Publishing
- Observational data
Automated Methods Required

- Volume of information
- Speed of change/production
- Complexity
- Need impartial/consistent analysis
The Goal

Maintaining and increasing the value of information

• Collecting information is not the problem - gaining understanding is the issue

• Too much information is just as useless at too little information
Some Common Methods

- Lexical matching
- Statistical evaluation
- Vector space models
- Rule based systems
- Parts of speech analysis
The Problem

- Failure to capture meaning and provide insight
- Methods not universally applicable
  - Language or domain dependent
- Need for specialized prior knowledge of data
- Require human interaction
  - Tagging, Keyword identification, Categorizations
- Not practical for large data sets
The Cost

Too much information is just as useless at too little information

Failure to understand the information we have leads to:

• Lost opportunities
• Unsatisfied customers
• Inability to fulfill mission
• Financial repercussions
What is LSA?

Latent Semantic Analysis
Latent Semantic Indexing

• The distinction is really one of application, as the same mathematics and computation are employed for both.

• LSA may be considered to refer to a broad collection of application while LSI is more closely associated with information retrieval.
Latent Semantic Analysis

• Theory of meaning[8]
  • Creates a mapping of meaning acquired from the text itself

• Computational model
  • Can perform many of the cognitive tasks that humans do essentially as well as humans [7]
How LSA Works

- LSA processing constructs a mapping of meaning in a semantic space.
- The mapping gives the meaning of words and documents not vice versa.
Compositionality Constraint

• The meaning of a document is the sum of the meaning of its words
• The meaning of a word is defined by the documents in which it appears (and does not appear)
LSA Space Construction

• LSA models a document as a simple linear equation

• A collection of documents (corpus) is a large set of simultaneous equations
Processing a Corpus

• Divide text corpus into units (documents)
  • Typically paragraphs of text
• Raw matrix is constructed from units
  • One row for each word type
  • One column for each unit (document)
  • Cells contain the number of times a particular word appears in a particular document
• Weighting functions may be applied [5]
Sparse Matrix

- The weighted term by document matrix represents a large set of simultaneous equations
- The term by document matrix is sparse
  - Typically less than 1% of the values are nonzero [2]
Solving Simultaneous Equations

- The system of simultaneous equations is solved for the meaning of each word type and document
  - Sparse matrix Singular Value Decomposition
  - Lanczos algorithm is typically used
  - Only solve for a reduced number of dimensions
  - Produces vectors representing the meaning of each term and document
Singular Value Decomposition \[^{[10]}\]

- The rows of matrix \( U \) are the vectors for the word types.
- Columns of \( U \) are the eigenvectors defining the axes for word type space.

\[
A = U \Sigma V^T
\]
Singular Value Decomposition \([10]\)

- The rows of matrix \(V\) are the text unit (document) vectors
- Columns of \(V\) are eigenvectors defining the axes for document space

\[
A = U \Sigma V^T
\]
Dimensional Reduction

• Typically solve for 300 - 500 dimensions \[^{10}\]
• Dimensional reduction allows comparison of all terms and all documents with each other
• In the sparse matrix comparison was not possible
• Dimensions are orthogonal
Dimensional Reduction

- With enough variables, every object is different
- With too few variables, every object is the same

Consider a geographic map
Semantic Space

- Vectors represent the meaning of a document (or term)
- Items similar in meaning are near each other in the semantic space
Computational Issues

• Nontrivial computation
  • Large sparse symmetric eigenproblem

• Scalability concerns [11]
  • Size of document set
  • Speed of processing

• Accuracy issues
  • Finite arithmetic introduces significant error
Operations

- Retrieval
- Clustering
- Comparison
- Interpretation
- Completion
Applications of LSA

Library Illustration

- Retrieval
- Content analysis
- Evaluation of “fit” into an existing collection
- Comparison of multiple collections
- Indexing of multilingual collections
Applications of LSA

- Repairing/cleaning data
- Education
  - Grading
  - Summarizing
- Non-textual applications
  - Bio-informatics
  - Personality profiles/compatibility analysis
Misconceptions and Misunderstandings

- Driven by term co-occurrence
- Word order issues
- Data collection size
- Content and Meaning
Co-occurrence

- LSA starts with a kind of co-occurrence
- Appearing in the same document does not make words similar
- Similarity is determined by the effect of the word meaning on the system of equations
Word Order

• Word order effects almost entirely within single sentences

• Research indicates only around 10% of meaning is word order dependent (for English)

[“order syntax? Much. Ignoring word Missed by is how”][8]
Data Collection Size

Beware of small data collections

• Generally - Use at least 100,000 documents
Content

LSA builds its notion of meaning from the content of the data collection
The Problem Revisited

- Failure to capture meaning and provide insight
- Methods not universally applicable
- Need for specialized prior knowledge of data
- Require human interaction
- Not practical for large data sets
Conclusion

● LSA offers powerful capabilities for gaining insight and understanding the contents of a data collection

● LSA provides analysis techniques not available with other Text Analytic methods

● Small Bear Technologies provides the core technology, tools, and support for performing Latent Semantic Analysis
Suggested Reading


References


References (cont.)


