Text Analytics and Graph Theory
By Victoria Loewengart

Advanced.

Analytical.

Intelligence.
Agenda

• Introductions
• Brief Overview of Natural Language Processing
  – Definitions and terms
  – Basic concepts – NE, Topics, Sentiment, Semantic, Ontology
  – Basic processes – Sentence detection, tokenization, stemming, POS tagging, chunking, parsing
  – Higher level processes – clustering, classification, importance, co-references
• Usage of graph theory
  – Using named entity and semantic decomposition to create a graph
  – Semantic relationships and ontologies
    • Overview
    • Navigation and inference
  – Clustering and topic extraction
    • Distance based clustering
    • Other methods
      – Term frequency – inverse document frequency (tf-idf)
      – Eigenvectors and Singular Value Decomposition
• Summary and Conclusion
Definitions

- **Natural Language Processing (NLP)** is understanding, analysis, manipulation, and/or generation of natural (spoken) languages.

- **Computational Linguistics** is the study of the applications of computers in processing and analyzing language, as in automatic machine translation and text analysis.

- **Text analytics** is the process of deriving high-quality information from text.
Basic Concepts

• **Information Retrieval (IR)** refers to the human-computer interaction (HCI) that happens when we use a machine to search a body of information for information objects (content) that match our search query. Depending on the sophistication of the algorithm, a person's query is matched against a set of documents to find a subset of 'relevant' documents.

• **Information Extraction (IE)** is extraction of specific information such as Named Entities, Events, and Facts.

• Metrics are **Precision, Recall, and F-Measure**
IR – Hubs and Authorities

- **Hubs** are index pages that provide lots of useful links to relevant content pages (or authorities).

- **Authorities** are pages that are recognized as providing significant, trustworthy, and useful information on a topic.

- Together they form a bipartite graph.
Basic Operations

- Sentence detection
  - A bit more difficult than it seems
    - Bob gave Mr. Bill a pencil.
- Tokenizing
- Stemming – singing -> sing
- Part of speech tagging
  - Penn Treebank, Treebank II
    - Tagged copra
    - Gold standards
  - More complex methods
- Chunking and parsing
  - Finding phrases
- Semantic analysis
  - Subject, object, other relationships

<table>
<thead>
<tr>
<th>Table 2</th>
<th>The Penn Treebank POS tagset.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. CC</td>
<td>Coordinating conjunction</td>
</tr>
<tr>
<td>2. CD</td>
<td>Cardinal number</td>
</tr>
<tr>
<td>3. DT</td>
<td>Determiner</td>
</tr>
<tr>
<td>4. EX</td>
<td>Existential there</td>
</tr>
<tr>
<td>5. FW</td>
<td>Foreign word</td>
</tr>
<tr>
<td>6. IN</td>
<td>Preposition/subordinating</td>
</tr>
<tr>
<td></td>
<td>conjunction</td>
</tr>
<tr>
<td>7. JJ</td>
<td>Adjective</td>
</tr>
<tr>
<td>8. JJR</td>
<td>Adjective, comparative</td>
</tr>
<tr>
<td>9. JJ$</td>
<td>Adjective, superlative</td>
</tr>
<tr>
<td>10. LS</td>
<td>List item marker</td>
</tr>
<tr>
<td>11. MD</td>
<td>Modal</td>
</tr>
<tr>
<td>12. NN</td>
<td>Noun, singular or mass</td>
</tr>
<tr>
<td>13. NNS</td>
<td>Noun, plural</td>
</tr>
<tr>
<td>14. NNP</td>
<td>Proper noun, singular</td>
</tr>
<tr>
<td>15. NNPS</td>
<td>Proper noun, plural</td>
</tr>
<tr>
<td>16. PDT</td>
<td>Predeterminer</td>
</tr>
<tr>
<td>17. POS</td>
<td>Possessive ending</td>
</tr>
<tr>
<td>18. PRP</td>
<td>Personal pronoun</td>
</tr>
<tr>
<td>19. PR$</td>
<td>Possessive pronoun</td>
</tr>
<tr>
<td>20. RB</td>
<td>Adverb</td>
</tr>
<tr>
<td>21. RBR</td>
<td>Adverb, comparative</td>
</tr>
<tr>
<td>22. RBS</td>
<td>Adverb, superlative</td>
</tr>
<tr>
<td>23. RP</td>
<td>Particle</td>
</tr>
<tr>
<td>24. SYM</td>
<td>Symbol (mathematical or</td>
</tr>
<tr>
<td></td>
<td>scientific)</td>
</tr>
<tr>
<td>25. TO</td>
<td>to</td>
</tr>
<tr>
<td>26. UH</td>
<td>Interjection</td>
</tr>
<tr>
<td>27. VB</td>
<td>Verb, base form</td>
</tr>
<tr>
<td>28. VBD</td>
<td>Verb, past tense</td>
</tr>
<tr>
<td>29. VBG</td>
<td>Verb, gerund/present participle</td>
</tr>
<tr>
<td>30. VBN</td>
<td>Verb, past participle</td>
</tr>
<tr>
<td>31. VB$</td>
<td>Verb, non-3rd ps. sing. present</td>
</tr>
<tr>
<td>32. VBG</td>
<td>Verb, 3rd ps. sing. present</td>
</tr>
<tr>
<td>33. WDT</td>
<td>w/determiner</td>
</tr>
<tr>
<td>34. WP</td>
<td>wh-pronoun</td>
</tr>
<tr>
<td>35. WP$</td>
<td>Possessive wh-pronoun</td>
</tr>
<tr>
<td>36. WRB</td>
<td>wh-adverb</td>
</tr>
<tr>
<td>37. #</td>
<td>Pound sign</td>
</tr>
<tr>
<td>38. $</td>
<td>Dollar sign</td>
</tr>
<tr>
<td>39. .</td>
<td>Sentence-final punctuation</td>
</tr>
<tr>
<td>40. ,</td>
<td>Comma</td>
</tr>
<tr>
<td>41. :</td>
<td>Colon, semi-colon</td>
</tr>
<tr>
<td>42. (</td>
<td>Left bracket character</td>
</tr>
<tr>
<td>43. )</td>
<td>Right bracket character</td>
</tr>
<tr>
<td>44. &quot;</td>
<td>Straight double quote</td>
</tr>
<tr>
<td>45. '</td>
<td>Left open double quote</td>
</tr>
<tr>
<td>46. `</td>
<td>Right open double quote</td>
</tr>
<tr>
<td>47. '</td>
<td>Right close single quote</td>
</tr>
<tr>
<td>48. &quot;</td>
<td>Right close double quote</td>
</tr>
</tbody>
</table>
Nouns

- Proper nouns – James Woods
- Pronouns – she, he, it, they ...
- Singular or plural - trees
- Possessive – Jimmy’s
- Conjunctions – Bill and Bob
- Noun phrase – A big green tree

The monkey ate a banana.
Named Entities

• Start by tagging
  
  Jim Black was the CEO of Medco Enterprises.

  [NNP NNP VBD DT EX IN NNP NNP]

• How do we do this?
  – Rules
  – Stochastic methods – machine learning, MEMM

• How do we assign more specific meaning?
  – Dictionaries, thesauri, ML disambiguation, etc.

• How about this one: Jim got Ed James’s book.
Rules and Stochastic Methods

• Rules specify “productions”
  Noun phrase -> Noun
    -> Adjective Noun
    -> Determiner Adjective Noun
    -> Noun “Prepositional Phrase”
  Prepositional Phrase -> Preposition “Noun Phrase”
  and so on.....

• Stochastic methods specify relationship probabilities
  – We start by examining a very large number of sentences (from a corpus)
  – Given a new sentence, if we have found a Determiner, then an adjective, what is the probability that the next word is a noun?
Text Analytics Processes
Named Entity Extraction

• Named Entities applicable to most domains:
  — People names
  — Organization names
  — Dates
  — Locations (Countries, Cities, Continents/geographic terms)
  — Currency

• Domain specific named entities:
  — Diseases, diagnoses, procedures, body parts
  — Drugs, dosages, and usage
  — Identifiers – SSN, Driver’s license, Claim number, Domain name, URL
The National Information Exchange Model (NIEM)

<table>
<thead>
<tr>
<th>Person</th>
<th>Patient</th>
<th>Police Report</th>
<th>User ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Medical Provider</td>
<td>Coroner Report</td>
<td>IP Address</td>
</tr>
<tr>
<td>Address</td>
<td>Hospital or Facility</td>
<td>Arrest Record</td>
<td>Network Origination</td>
</tr>
<tr>
<td>Phone</td>
<td>Pharmaceutical</td>
<td>Charge</td>
<td>Online postings</td>
</tr>
<tr>
<td>Identification</td>
<td>Diagnosis / Injury</td>
<td>Conviction</td>
<td>Social Media Pages</td>
</tr>
<tr>
<td>License</td>
<td>Procedure</td>
<td>Enforcement Agency</td>
<td>Email</td>
</tr>
<tr>
<td>Company</td>
<td>Pharmacy</td>
<td>Alias</td>
<td>Text Messages</td>
</tr>
<tr>
<td>Vehicle</td>
<td>Medical Report</td>
<td>Observation</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>Biometrics</td>
<td>Weapon</td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>Criminal Method</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Account</th>
<th>Info Bearing Entity</th>
<th>Security log</th>
<th>Financial Instrument</th>
</tr>
</thead>
<tbody>
<tr>
<td>Credit Card</td>
<td>Document</td>
<td>Web log</td>
<td>Event</td>
</tr>
<tr>
<td>Policy</td>
<td>URL</td>
<td>Asset</td>
<td>Task</td>
</tr>
<tr>
<td>Claim</td>
<td>Term</td>
<td>Asset class</td>
<td>Language</td>
</tr>
<tr>
<td>Lien</td>
<td>Concept</td>
<td>HR Report</td>
<td>Prediction</td>
</tr>
<tr>
<td>Title</td>
<td>Sentiment</td>
<td>Encryption Method</td>
<td>Inference</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Security log</th>
<th>Web log</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asset</td>
<td>Asset class</td>
<td>Task</td>
</tr>
<tr>
<td>Title</td>
<td>HR Report</td>
<td>Language</td>
</tr>
<tr>
<td>Lien</td>
<td>Encryption Method</td>
<td>Prediction</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>Inference</td>
</tr>
</tbody>
</table>
Named Entities and More Complex Graphs

Bob was Bill’s friend and he drove a white truck.

Bob drove Bill’s white truck and he wanted one of his own.

Bob was Bill’s friend and he drove his white truck and wanted one of his own.
Simple Relationships

Semantic relations among words can be extracted from their textual context in natural languages.

- **Relationships** may occur through communication, friendship, advice, influence, or exchange. The two basic elements of a relationship network are links and nodes.
- **Relationship analysis** is the mapping and measuring of relationships and flows between people, groups, organizations, computers or other information/knowledge processing entities.

*Graphs allow us to store the relationships between entities, and algorithms allow us to interrogate these connections.*
Simple Relationships - Example
Simple Relationships - Techniques

• Simple relationships are identified through co-reference.
• Co-reference is the instance of occurrence within a unit of text
  • Sentence
  • Paragraph
  • Document
• Metadata is relevant too – coauthors.
• Topics are words that are assigned to a document that relate “concepts.”
Text Analytics Processes

Semantic Named Entity Extraction

• Nouns are parsed into sentence structures
  – Yields <subject> <verb> <object> relationships
  – Can usually detect compound subjects and various verb inflective forms
  – Captures modifiers (adjectives and adverbs) that can be used in sentiment or inversion

• Graph analysis and graph theory now comes into play
  – When documents and document sets are processed, typically creates a very large graph

Clusters of terms

Graph structures

Central terms
Text Analytics Processes
The Importance of Ontologies

• An Ontology is “a description of things that exist and how they relate to each other” (Chris Welty).

• An Ontology Model is:
  – the classification of entities and
  – modeling the relationships between those entities.
What is an ontology?

- An ontology is a specification of a conceptualization (Thomas Gruber)
Why create ontologies?

...And that, in simple terms, is how you increase your ranking on search engines.”
Why you should think “semantically” even if you never create a formal ontology

```xml
<xs:element name="Date">
  <xs:complexType>
    <xs:choice>
      <!-- for those unparsed dates -->
      <xs:element name="Date_str" type="xs:string"/>
      <!-- use this if you can -->
      <xs:element name="Date_std">
        <xs:complexType>
          <xs:sequence>
            <xs:element ref="Date-std"/>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
    </xs:choice>
  </xs:complexType>
</xs:element>
<!-- NOTE: this is NOT a unix tm struct -->
<xs:element name="Date-std">
  <xs:complexType>
    <xs:sequence>
      <!-- full year (including 1900) -->
      <xs:element name="Date-std_year" type="xs:integer"/>
      <!-- month (1-12) -->
      <xs:element name="Date-std_month" type="xs:integer" minOccurs="0"/>
      <!-- day of month (1-31) -->
      <xs:element name="Date-std_day" type="xs:integer" minOccurs="0"/>
      <!-- for "spring", "may-june", etc -->
      <xs:element name="Date-std_season" type="xs:string" minOccurs="0"/>
      <!-- hour of day (0-23) -->
      <xs:element name="Date-std_hour" type="xs:integer" minOccurs="0"/>
      <!-- minute of hour (0-59) -->
      <xs:element name="Date-std_minute" type="xs:integer" minOccurs="0"/>
      <!-- second of minute (0-59) -->
      <xs:element name="Date-std_second" type="xs:integer" minOccurs="0"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
```
**Why you should create ontologies**

- Ontologies model their domain and models provide:
  - A common understanding of the structure of information among people and software
  - Explain and make and make predictions
  - Enable reuse of domain knowledge
  - Make domain assumptions explicit and mediates between multiple viewpoints
  - Aids in separating domain knowledge from operational knowledge
Is it a lexicon, a thesaurus or an ontology?

- As you can see, we're already in trouble with our semantics...

- The first rule of an ontology is to begin to define the specification of the conceptualizations.
  - Conceptualizations: Lexicon, Thesaurus, Ontology
  - Specifications:
    - **Lexicon**: Set of words used within a domain of discourse. The meaning of the words rely on their “linguistic meaning”. Good luck getting two people to agree to the linguistic meaning even in relatively simple domains.
    - **Thesaurus**: A set of words used within a domain of discourse and may contain “is-a” relationships between the words. The meaning of the words still rely on their “linguistic meaning”.
    - **Ontology**: A formal explicit specifications of concepts in the domain of discourse. You will still disagree, but formalizing a specification forces a decision and an understanding.
Spectrum of Knowledge and Reasoning

Strong Semantics

Composite applications and info/knowledge-intensive processes share information at these levels

Data models set the bar here

Conceptual Model

Logical Theory

Axiomology

Modal Logic

First Order Logic

Description Logic

Semantic Interoperability

Structural Interoperability

Syntactic Interoperability

Increasing Metadata

Thesaurus

Taxonomy

ER Model

Database Schema, XML Schema

Relational Model, XML

Glossary

Controlled Vocabulary

Most applications do not enable data discovery and sharing

Increasing Search Capability

List

Discovery

Intelligence

Question Answering

Reasoning

Source: Dr. Leo Obst, Mines: Mils Davis, Project 10X
Moving up the knowledge and reasoning spectrum

- SKOS (Simple Knowledge Organization System)
  - Migrate thesauri into a SKOS structure
  - Uses RDF to define thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web

- Begin development of specifications within OWL
  - OWL Species
    - OWL Full – Superset of RDF, no constraints in expression, has undecidability
    - OWL DL – Description Language sublanguage provides decidability and inferencing capabilities
    - OWL Lite - provides a minimal set of features
Building blocks of the semantic web

- **RDF** – Resource Description Framework
  - Provides a basic mechanism to allow anyone to make a basic statement about anything

- **RDFS** – The RDF Schema Language
  - Provides the capability to describe classes, subclasses, and properties

- **RDFS-Plus**
  - A subset of OWL, more expressive than RDFS, but less complex than OWL

- **OWL** – Ontology Web Language
  - Allows modelers to express detailed constraints between classes, entities and properties
RDF basics

- The building block for RDF is the *triple*
  - `<subject><predicate><object>`
- RDF triples can be viewed as a directed graph
  - RDF can represent the semantic graph as well as the ontological graph and their exists a mapping between the ontological graph and the semantic graph
- Semantic identify of nodes are resolved through the use of URIs (Uniform Resource Identifiers)
- RDF allows you to make statements about other statements (reification)
RDFS basics

- Provides the capability to describe classes, subclasses and properties:
  - `rdfs:Class`
    - Describes things
  - `rdfs:subClassOf`
    - Describes things that have an “is-a” relation to other things
  - `rdfs:superClassOf`
    - Describes things that have an “is-a” relation to other others
  - `rdfs:subPropertyOf`
    - Describes the relationships that one thing has with another
RDFS-Plus (SKOS)

- SKOS (Simple Knowledge Organization System) is implemented in RDFS-Plus
  - Contains thesauri-like entities implemented in RDFS-Plus
    - Preferred label (skos:prefLabel)
    - Broader terms (skos:broader)
    - Narrow terms (skos:narrow)
    - Synonyms (skos:altLabel)
    - Related terms (skos:related)
    - Scope Notes (skos:scopeNote)
**OWL basics**

- **owl:Thing**
  - The most general class in OWL, provides the superclass for all classes

- **owl:Class**
  - Defines basic concepts

- **owl:Nothing**
  - Subclass of all classes

- **Properties**
  - Related subject individuals to either a datatype value or another individual object.
    - **owl:DatatypeProperty** (associates with a datatype)
    - **owl:ObjectProperty** (relation to another individual)
    - **owl:AnnotationProperty** (annotation of individuals, “rdfs:comment”)
    - **owl:OntologyProperty** (relates two ontologies)
Classes, instances and individuals

- Generally in OWL an instance is considered synonymous with an individual
  - Although one speaks of Bernie as an individual, or Bernie as an instance of Person

- Choosing between a class and an instance/individual is not always clear cut.
  - Is Bernie a class of all people named Bernie on the planet, or is Bernie an individual who lives at a specific address.
    - OK..I'm named after my Dad so what if there are two Bernie's living at the same address. Is it Bernie at an address born in 1936, or Bernie at the same address born in 1955, or the Bernie with this fingerprint... or...
Extending beyond a single ontology to building common knowledge across domains

- The concept of a Semantic Web has resulted in the explosion of ontologies describing overlapping domains.
- This semantic fragmentation has driven the need to align similar ontologies.

Bipartite graph mapping between Staff's and Starbucks' ontologies.
Using the ontology graph to accelerate path finding in semantic graphs

- Ontology graphs can be used to prioritize the search space in large semantic graphs
  - Ability to “reason” over smaller graphs more efficiently.
- Algorithms such as A* use heuristics to estimate the future cost of exploring a search space.
- Frequency statistics about vertex types and edge types can be used to define a probability model for measuring an edge's occurrence in the semantic graph.
- These heuristics can be embedded in algorithms such as A*
The best advice I can give you, if you are interested in building ontologies, is to read this book...

Semantic Web for the Working Ontologist

By

Dean Allemang
Jim Hendler
Classification and Clustering - IR

• Classification / Categorization
  – The task is to assign a document to one or more classes or categories. This may be done "manually" or algorithmically.

• Clustering
  – Document clustering is a more specific technique for unsupervised document organization, automatic topic extraction and fast information retrieval or filtering.

• Machine learning is used
  – Supervised uses “known results”
  – Unsupervised finds results from the unknown
Clustering Example

Top 148 results of at least 2,348,524 retrieved for the query cyber attack (details)

Mitigate Cyber Attacks
Secure Your Sensitive Assets from Cyber Attack. View Free eBook!
www.cyber-ark.com/eBook

Cyber Security Testing
Ensure that your security device can prevent cyber security attacks
www.sprentifederal.com

IT Perimeter Defense
Easy-to-use, Low cost solutions, Protect critical digital assets
www.ovlcti.com/energy

Critical U.S. Infrastructure Vulnerable to Cyber Attack - YouTube
Since the internet has become so critical to daily life, systems like the electrical grid and water supply have become vulnerable to cyber attacks...
www.youtube.com/watch?v=Vnh3eEuKVM - [cache] - Additional Sources, Yippy Sources

Cyber Attack | eBay - Electronics, Cars, Fashion ...
Visit eBay for great deals on a huge selection Cyber Attack. Shop eBay!
www.ebay.com/sch/i.html?_nkw=cyber+attack - [cache] - Additional Sources, Yippy Sources

WikiLeaks plugged? Cyber attack on whistleblower enters 2nd ...
Whistleblower site WikiLeaks says it is enduring a week-long attack on its sites, which is rendering them sluggish or inaccessible.
Sentiment

• An opinion is a binary expression that consists of two key components:
  – A target (which we shall call “topic”, as referred to by most social analytics tools);
  – A sentiment on the target/topic, often accompanied by a probability.

• Sentiment analysis on content means discerning the opinions in content and picking the mood (attitude) within those opinions.

• A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level — whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral.
Sentiment Analysis
Israel / Gaza conflict

Sentiment Analysis Approaches

- **Sentiment analysis**
  - **Lexical Approach**
    - Use dictionaries and thesauri with sentiment words and given semantic orientation
  - **Supervised Machine Learning**
    - Learn from annotated corpus
  - **Semi-supervised and Unsupervised machine learning**
    - Small amount of annotated data
    - Graph-based algorithms
  - **Cross-domain sentiment classification**
    - Ensemble of classifiers
    - Graph-based algorithms
Text Analytics processes
Extracting topics and sentiment

• Sentiment dictionaries – anger, threat, mood (depression, happiness, etc.)
• Use term frequency-inverse document frequency (tf-idf) to calculate frequency of terms
• Assign weight to sentiment terms
• Use Latent Dirichlet Allocation (LDA), Collapsed Variational Bayes (CVB), and K-Means (KM) for topic clustering and classification
  – Can be supervised (topic list or controlled vocabulary is provided) or unsupervised (topic list is created by the algorithms)
Text Analytics Processes
Extracting topics

Documents

LDA and CVB

Clustering

Topic Term

Cluster Title

Topic Term

Cluster Title

Topic Term

Cluster Title

Topic Term

Cluster Title

Topic Term

Cluster Title

Topic Term

Cluster Title

Analytics Inside™ - 2014
Text Analytics processes

Document Term matrix

<table>
<thead>
<tr>
<th>Terms</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
<th>M6</th>
<th>M7</th>
<th>M8</th>
<th>M9</th>
<th>M10</th>
<th>M11</th>
<th>M12</th>
<th>M13</th>
<th>M14</th>
</tr>
</thead>
<tbody>
<tr>
<td>abnormalities</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>age</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>behavior</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>blood</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>close</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>culture</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>depressed</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>discharge</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>disease</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>fast</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>generation</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>oestrogen</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>patients</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>pressure</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>rats</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>respect</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>rise</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>study</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Note that this looks just like a graph adjacency matrix!

Space reduction, Latent Semantic Indexing, and eigenvectors

Reveals the most important terms in a set of documents
RelExtract™ Text Analytics

NLP Toolkit Components

- Named Entity Extraction
- Classification/Clustering
- Relationship Extraction

High Level Implementation

- Generic entities (people, organizations, places)
- Domain-specific entities
- Topic extraction
- Similarity
- TFIDF
- General Sentiment
- Domain specific Sentiment
- Sentence/Document level dependency
- Geocode/Proximity
- Sentence level dependency

High Level Usage

- Collection-level discovery
- Domain annotation
- Sentiment analysis
- Document Collection content discovery
- Domain specific/ general trend discovery
- Document summarization
- Inputs for graph analysis
- Domain-specific event discovery

August 6, 2015
Copyright © 2014-2015 Analytics Inside
Conclusion

![Diagram](image)

- **Information:**
- **Knowledge:**
- **?**
Questions & Answers

http://www.AnalyticsInside.us