Text Analytics 101

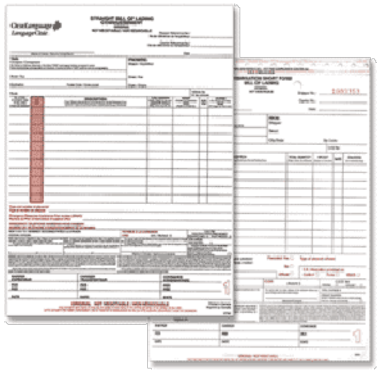
Rayid Ghani

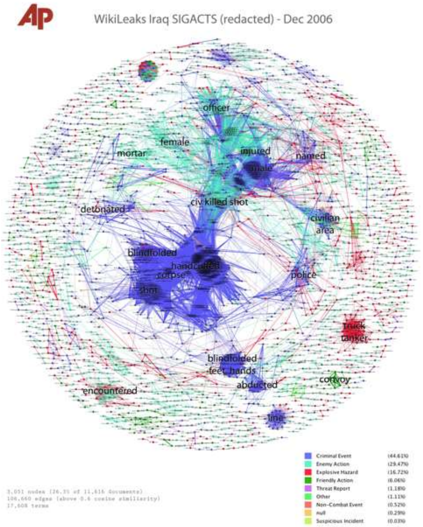
THE UNIVERSITY OF CHICAGO

What we'll cover today

- What kinds of analyses can be done with text data?
- How do to those analyses?
- What are those analyses useful for?

Where does text data come from?





Why is text data different?

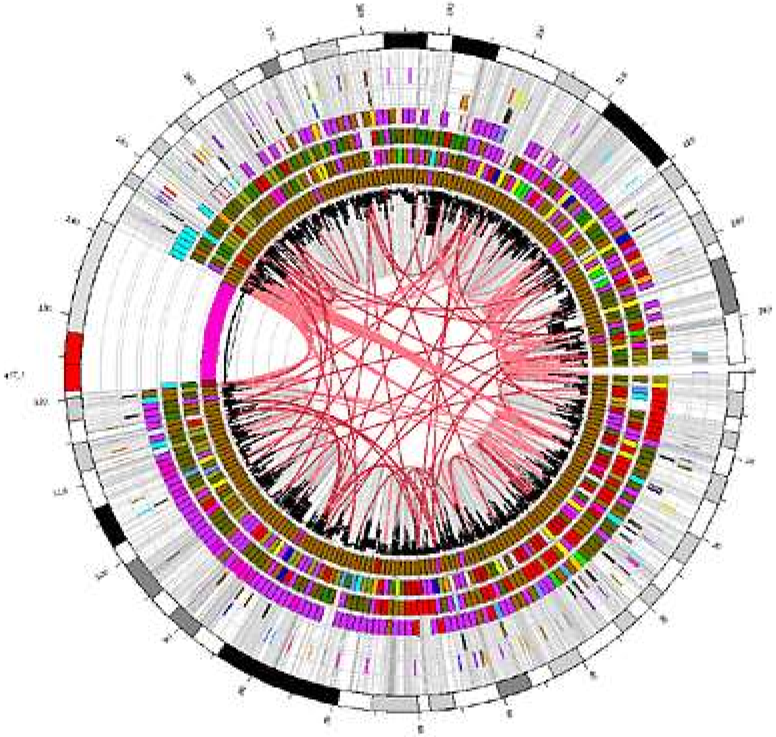
- Structured Data: Humans have already pre-determined the attributes and relations of interest (e.g. relational databases)

- Text data can have too many possible dimensions (millions)

- Text data often reflects human observations that are exceptions to regular business processes.
  - Complaints, Suggestions, (the ubiquitous “other” field)

Text Analytics: Capabilities

- Search
- Classification
- Clustering
- Extraction
- Summarization
- Visualization



Text mining platforms require a variety of capabilities

Capability	Description	Maturity
Classification	Classification of a document into one or more previously defined content categories. (e.g. Classify incoming emails as personal, business, news article, etc.)	High
Word Clustering / Synonyms	Finding groups of words that are similar to each other. Depending on the strictness of the definition of similarity, similar words can be synonyms (IBM, Deep Blue) or companies in similar industry (Oracle, Teradata)	Low
Topic Detection / Clustering	Finding emerging or existing topics in large amounts of text data. Clustering news articles can help detect emerging topics.	High
Opinion & Sentiment Analysis	Detection of sentiment and opinions in different levels of granularity (word, sentence, message).	Medium
Named Entity Extraction (People, Locations, Organizations)	Recognition, tagging, and extraction of named entities of type Person, Location, and Organization. Typically limited to proper nouns and not much customization possible.	High
General Extraction (Entities, Events, Facts, Relationships)	Recognition, tagging, and extraction of specified classes of words /phrases as an entity (client, competitor, company), event (acquisition), relationship (John King works for MSFT)	Low
Search	Ranked retrieval of a document or component based on the presence of one or more supplied search terms.	High
Visualization	Visualization of text data and /or visual mashups combining text with other forms of data (maps, networks, etc).	Medium
Summarization	Summarization of a document, intended to produce a readable abstract of the source document which captures salient points using fewer words.	Low

Different Text analytics paradigms: Best is to use statistical machine learning augmented with lexicons and linguistics

	Do it by hand	Hire linguists	Do it the right way
Description	Rules based on lists of words	Rules using words and linguistic operators (parts of speech for example)	Statistical approaches that can be trained and learn over time. Can incorporate lexicons and linguistics as well
Ease of creation & maintenance	Low	Low	High
Accuracy	Low	Medium	High
Context Sensitiveness	Low	High	High
Interpretability	High (unless the rules get large)	Medium	Medium

Today we’re going to

- Explore a new text corpus to understand what’s in it
- Build classifiers to scale human tagging / classification

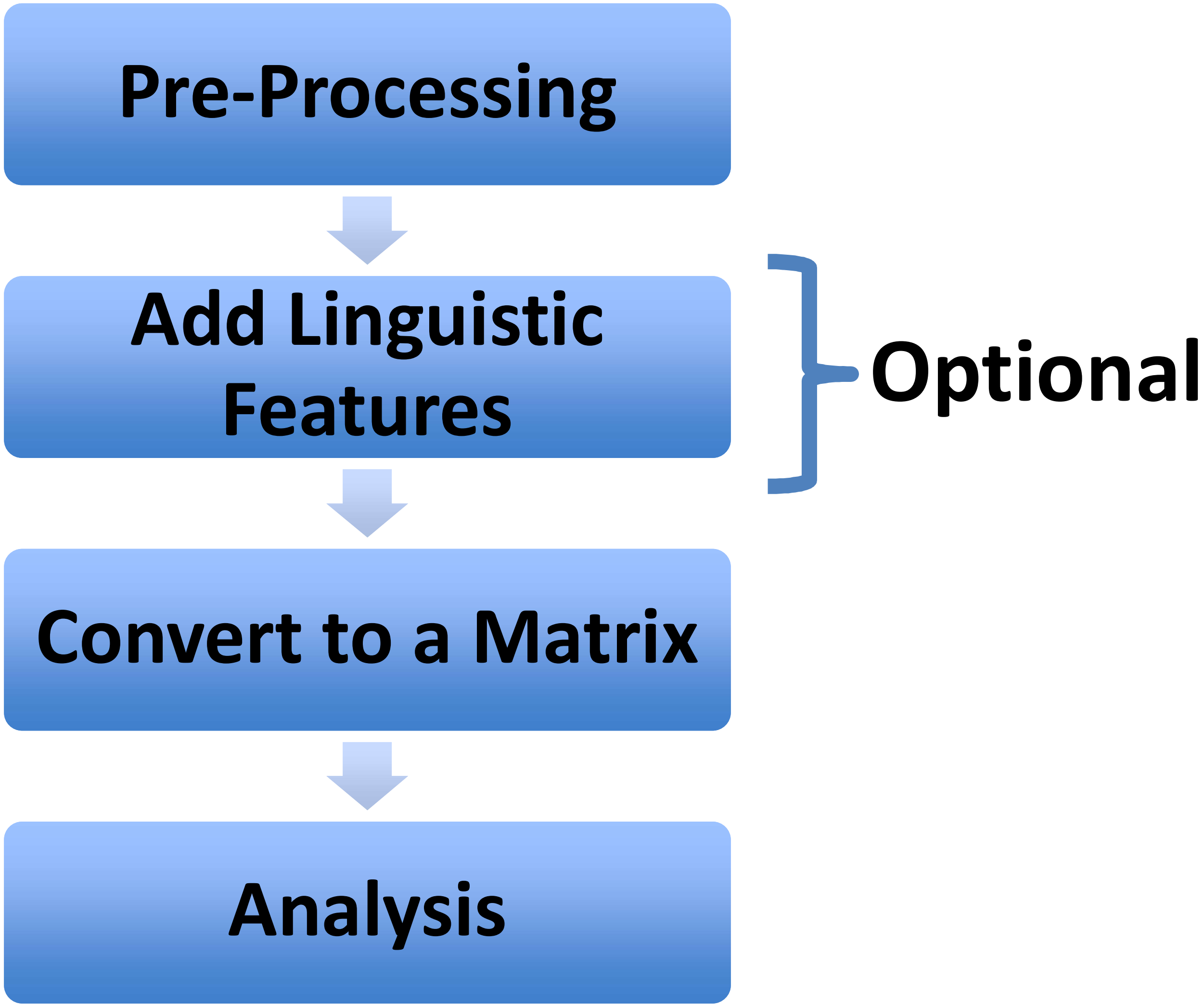
How do we do Machine Learning with Text Data?

- Everything we covered yesterday in ML applies here if we can convert text data into rows and columns

Simple ways of text to “data” conversion

- Treat each word as a column/variable/feature
- Remove, Combine, Transform, Abstract some words

NLP Pipeline



Pre-Processing

Add Linguistic Features

Optional

Convert to a Matrix

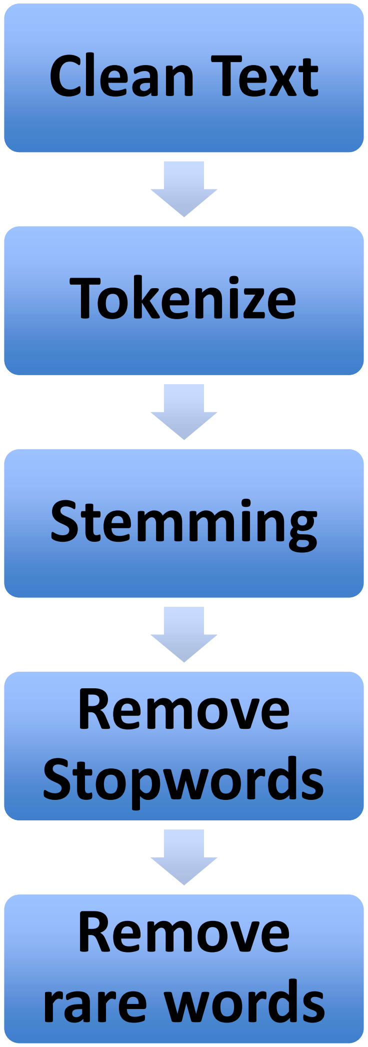
Analysis

Raw data from a webpage

<div><p class=header>Our mission is to provide comprehensive ;nbsp social services to refugees to help them overcome the societal and language barriers and become productive members of the US society. </p></div>

NLP Pipeline – Pre-Processing

<div><p class=header>Our mission is to provide comprehensive ;nbsp social services to refugees to help them overcome the societal and language barriers and become productive members of the US society. </p></div>



Clean Text

Tokenize

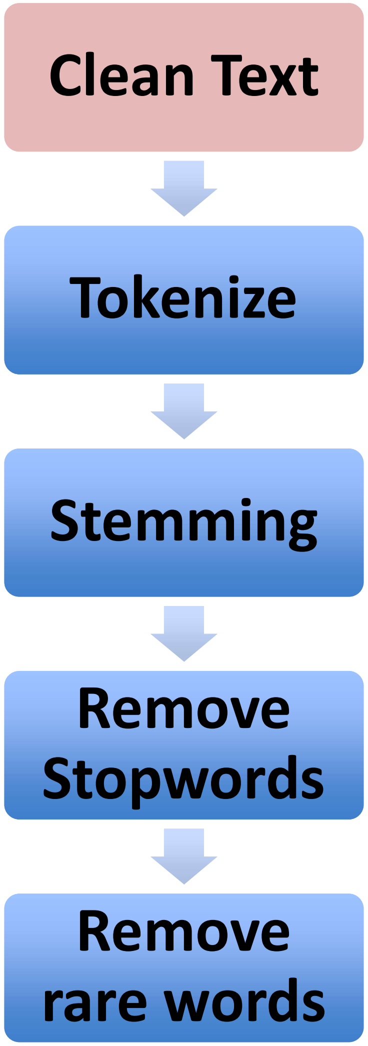
Stemming

Remove Stopwords

Remove rare words

NLP Pipeline – Pre-Processing

Our mission is to provide comprehensive social services to refugees to help them overcome the societal and language barriers and become productive members of the US society.



Clean Text

Tokenize

Stemming

Remove Stopwords

Remove rare words

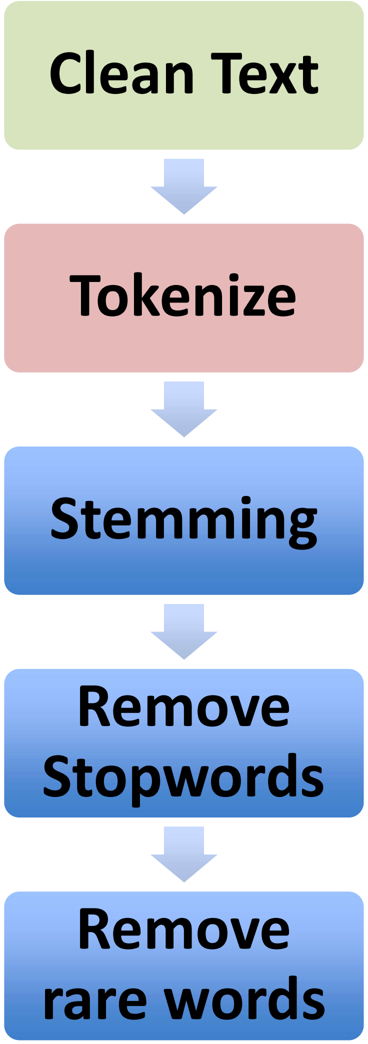
NLP Pipeline – Pre-Processing

Our mission is to provide comprehensive social services to refugees to help them overcome the societal and language barriers and become productive members of the US society.

Our | mission | is | to | provide | comprehensive | social | services | to

refugees | to | help | them | overcome | the | societal | and | language | barriers

and | become | productive | members | of | the | US | society | .



Clean Text

Tokenize

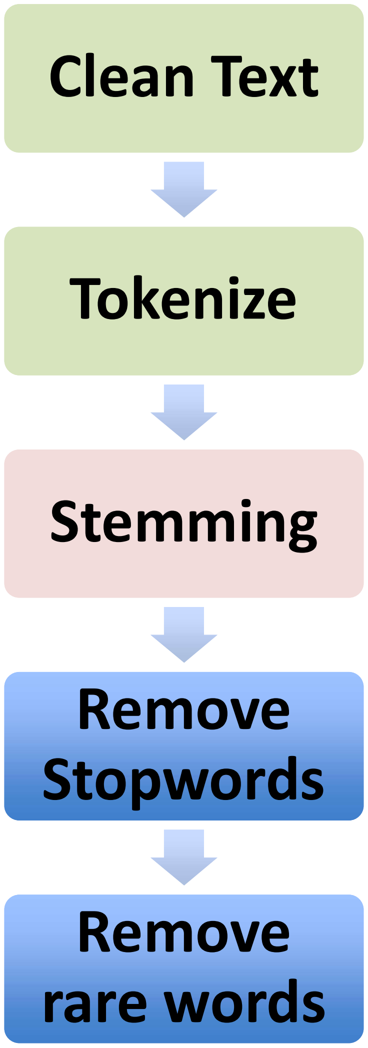
Stemming

Remove Stopwords

Remove rare words

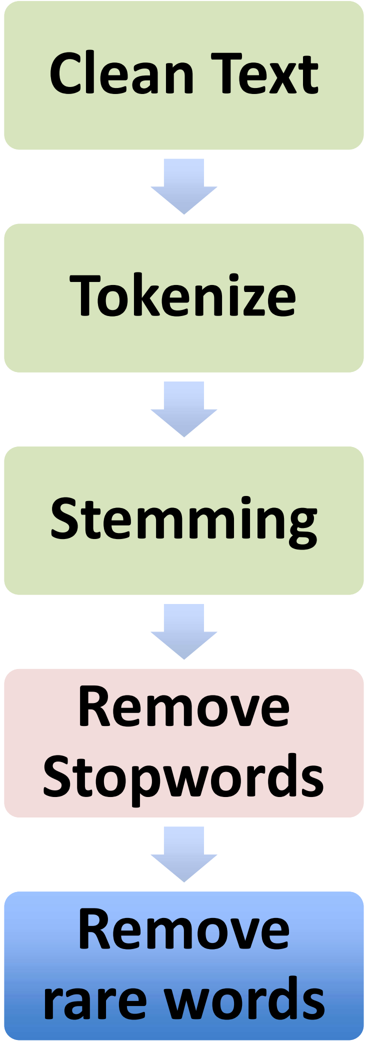
NLP Pipeline – Pre-Processing

Our mission is to provid comprehens social servic to refuge to help them overcom the societ and languag barrier and becom product member of the US societi .



NLP Pipeline – Pre-Processing

mission provid comprehens social servic refuge help overcom societ languag barrier becom product member US societi .



Clean Text

Tokenize

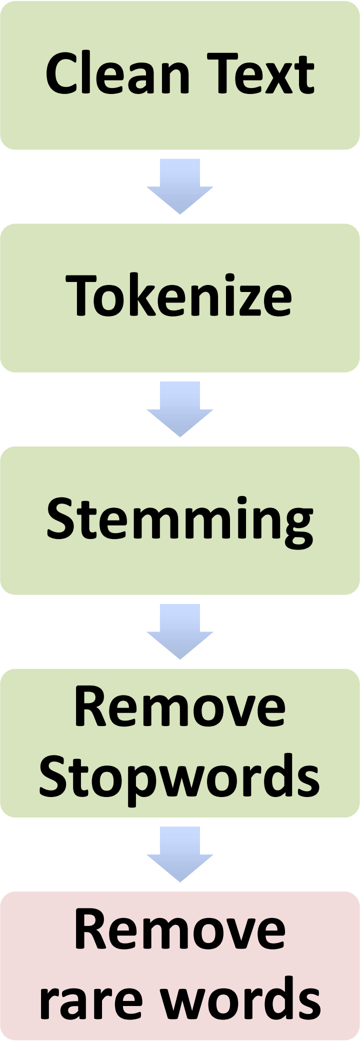
Stemming

Remove Stopwords

Remove rare words

NLP Pipeline – Pre-Processing

mission provid comprehens social servic refuge help overcom societ languag barrier becom product member US societi .



Clean Text

Tokenize

Stemming

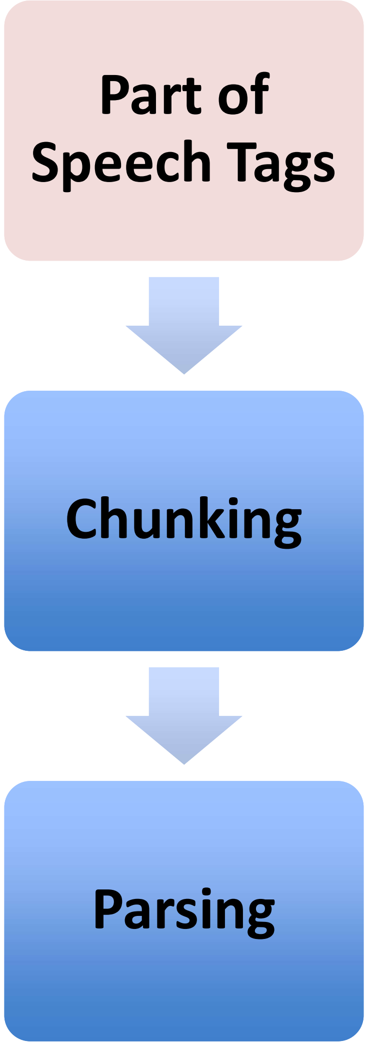
Remove Stopwords

Remove rare words

NLP Pipeline – Pre-Processing

PRP$ Our **NN** mission **VBZ** is **TO** to **VB** provide **JJ** comprehensive **JJ** social **NNS** services **TO** to **NNS** refugees **TO** to **VB** help **PRP** them **VB** overcome **DT** the **NN** societal **CC** and **NN** language **NNS** barriers **CC** and **VB** become **JJ** productive **NNS** members **IN** of **DT** the **NNP** US **NN** society **.** .

PRP Personal Pronoun
IN Preposition
NN Singular Noun
VBZ Verb, 3rd ps. sing. present


Part of Speech Tags

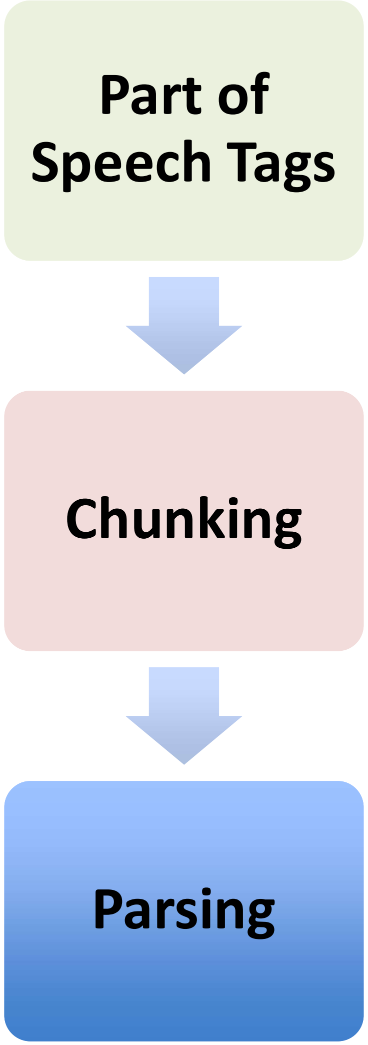
Chunking

Parsing

NLP Pipeline – Pre-Processing

NP Our mission **VP** is **VP** to provide **NP** comprehensive social services **PP** to **NP** refugees **VP** to help **NP** them **VP** overcome **NP** the societal and language barriers and **VP** become **NP** productive members **PP** of **NP** the US society .

NP	Noun Phrase
VP	Verb Phrase
PP	Prepositional Phrase



Part of Speech Tags

Chunking

Parsing

NLP Pipeline – Turning in to a Matrix

- What do you want the columns to be?
  - Words, phrases, POS tags, …
- What value do you put in a cell?
  - If a word appears in a document – binary variable
  - # of times a word in a document – word frequency
  - Words that uniquely characterize this document - tfidf

TFIDF – Term Frequency Inverse Document Frequency

- TF = Term Frequency (word count/ # words in the document)
- IDF = Inverse Document Frequency (how many documents does this word occur in?)

  log (# total documents / # documents this word appears in)

- TFIDF = TF X IDF

- highest when the word occurs many times within a small number of documents (thus lending high discriminating power to those documents)

- lower when the word occurs fewer times in a document, or occurs in many documents

- lowest when the word occurs in virtually all documents.

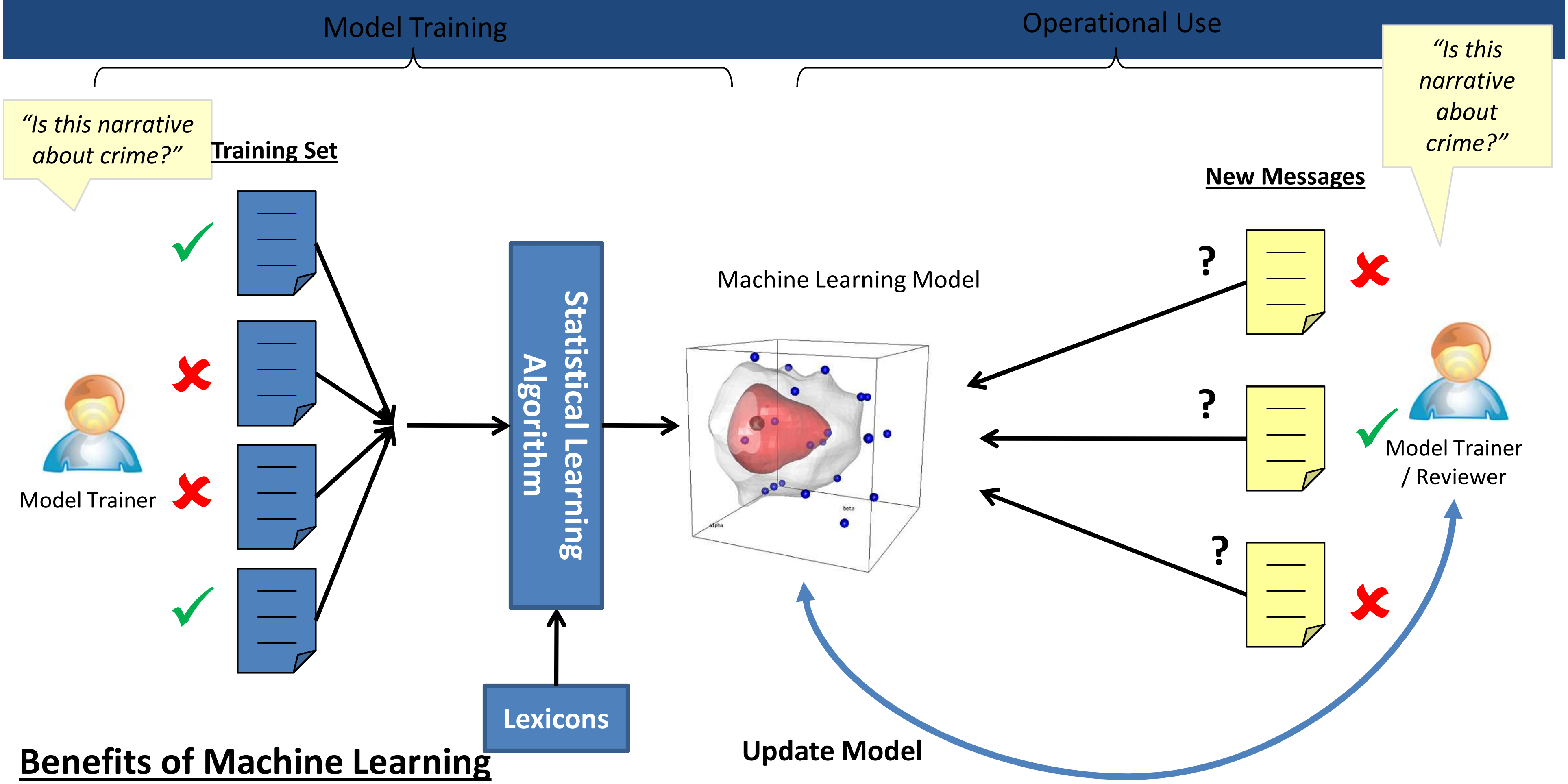
Analysis: Topic Models

- “Soft” clustering for analyzing text corpora
- Topic: probability distribution over words
- Document: probability distribution over topics
- Generative Model - To generate a document:
  - First select a topic
  - Then generate words from that topic’s distribution
  - Repeat
- Latent Dirichlet Allocation (LDA) is the most common method

Topic Models

- Python libraries
  - Gensim
  - Lda
- pyLDAvis is a python libarary for interactive topic model visualization

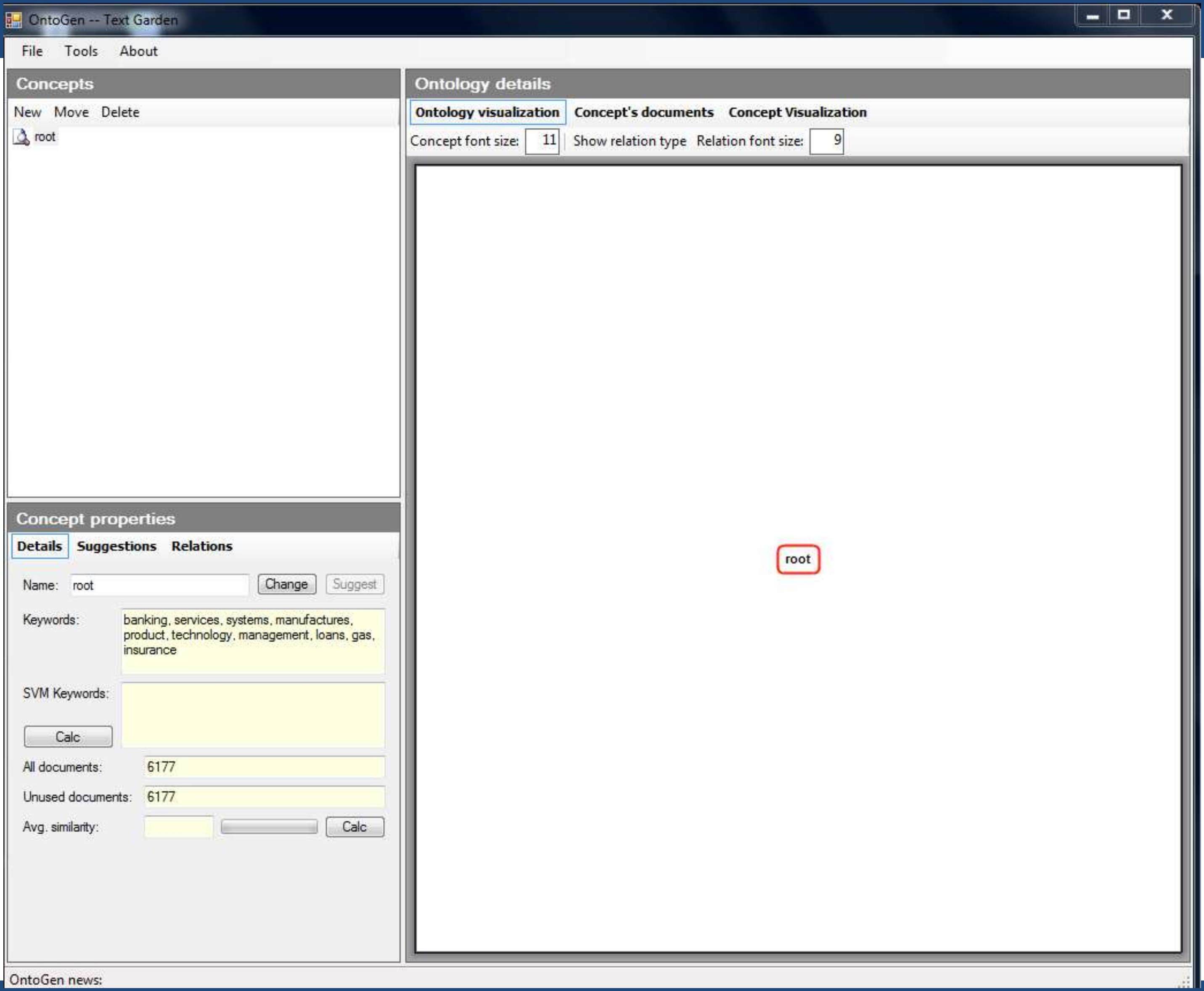
Statistical Machine Learning based approaches compensate for the shortcomings of rule-based systems



Benefits of Machine Learning

- ✓ Significantly cheaper approach to achieve a given level of accuracy compared to manual rule or lexicon creation
- ✓ No advanced linguistic or technical skills to train and maintain the system (business users or analysts are the maintainers)

OntoGen: Interactively discovering topics and themes in large amounts of text data



Concept font size: 11

Relation font size: 9

Name: root

Keywords: banking, services, systems, manufactures, product, technology, management, loans, gas, insurance

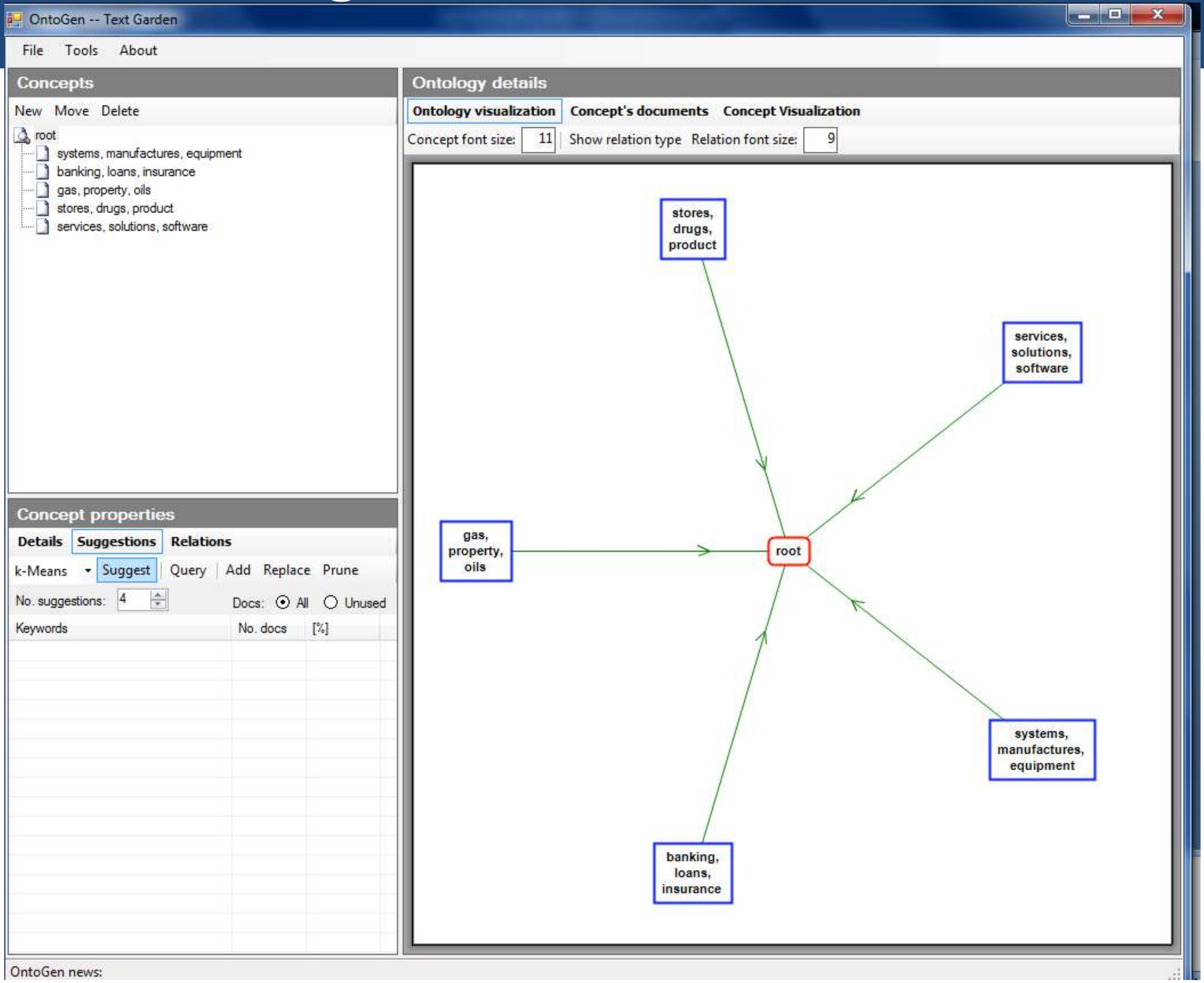
SVM Keywords:

All documents: 6177

Unused documents: 6177

Avg. similarity:

OntoGen: Interactively discovering topics and themes in large amounts of text data



OntoGen -- Text Garden

- root
  - systems, manufactures, equipment
  - banking, loans, insurance
  - gas, property, oils
  - stores, drugs, product
  - services, solutions, software

Concept font size: 11

Relation font size: 9

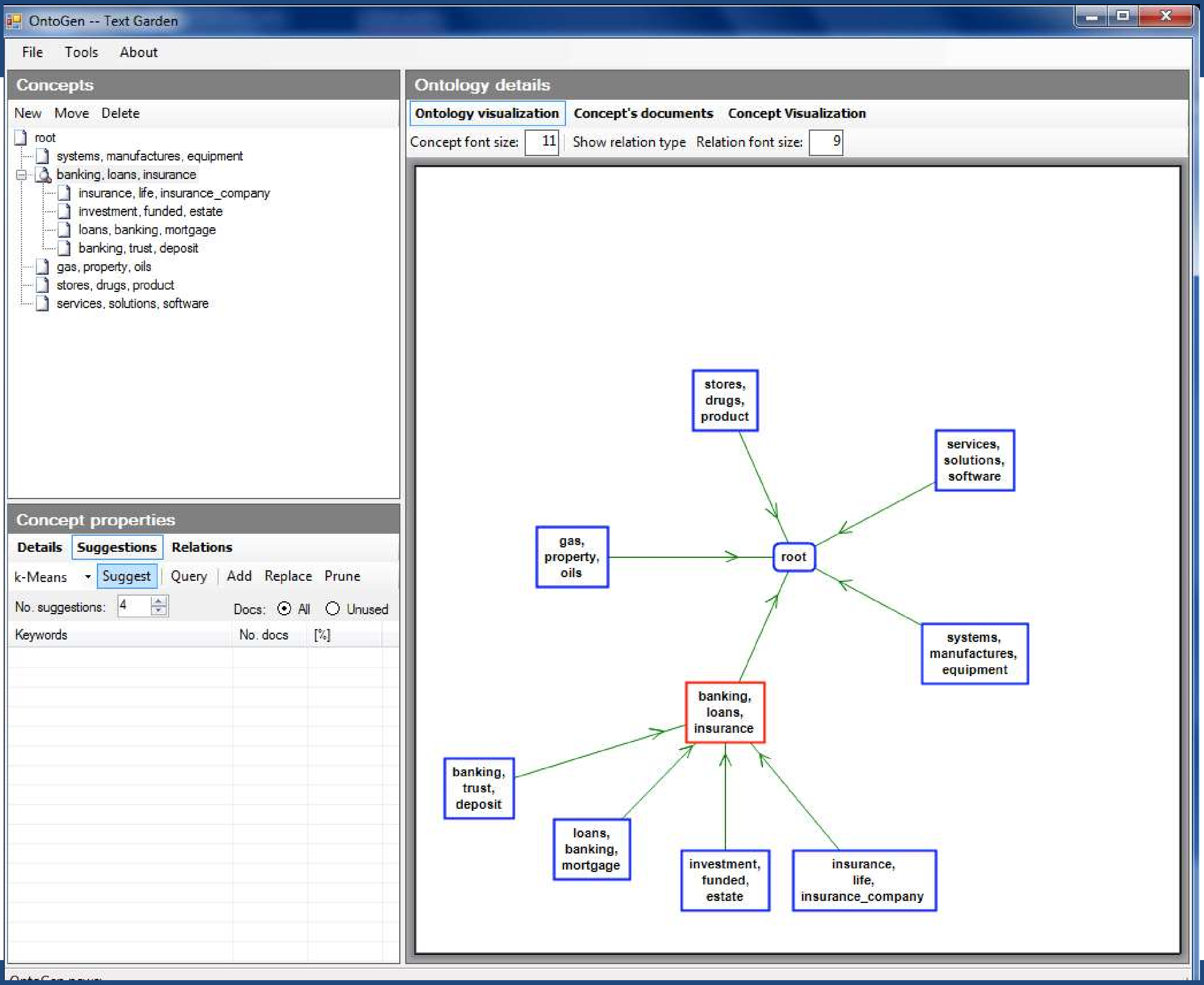
k-Means

No. suggestions: 4

Docs: ● All ○ Unused

Keywords	No. docs	[%]

OntoGen news:

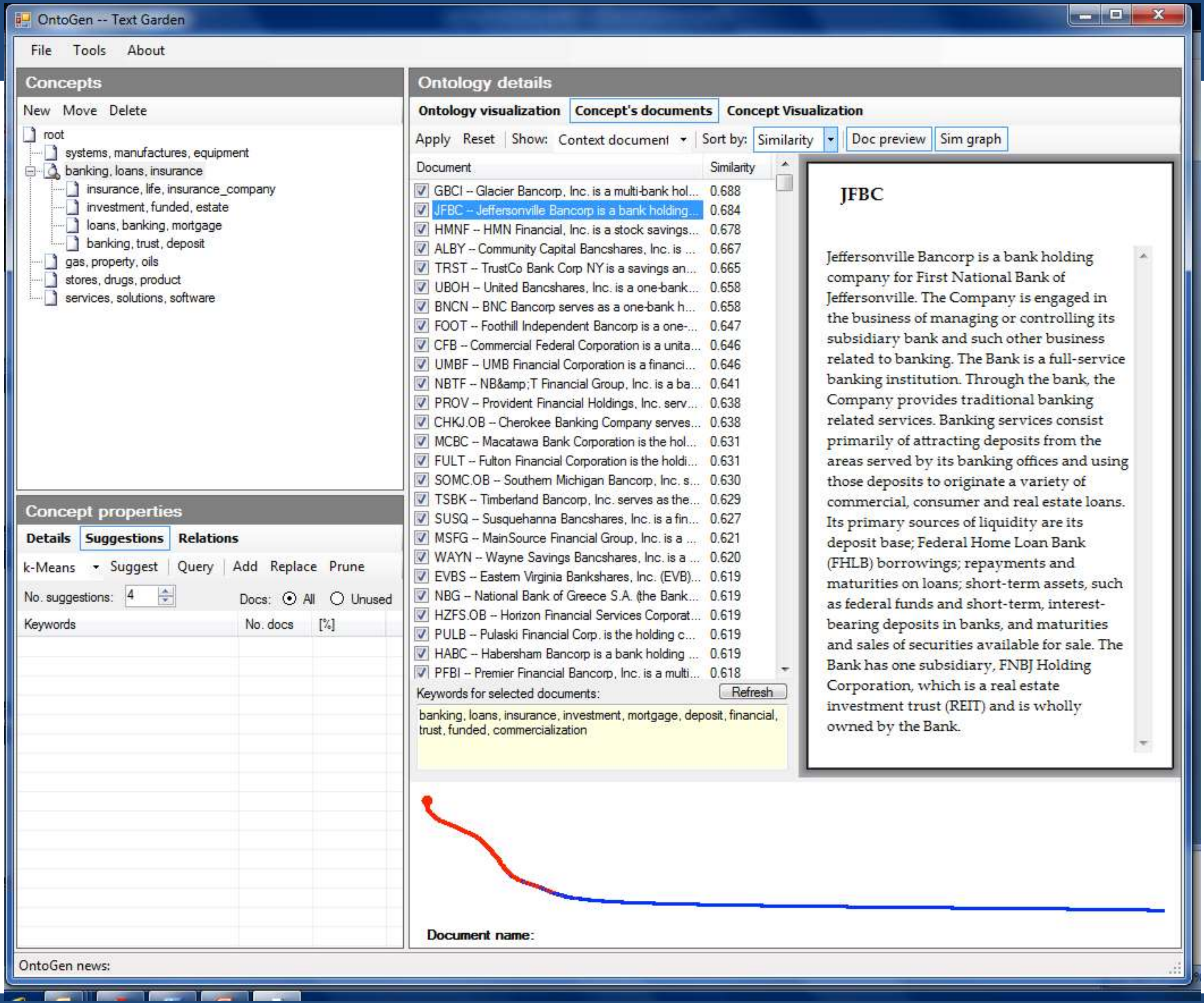


- root
  - systems, manufactures, equipment
  - banking, loans, insurance
    - insurance, life, insurance_company
    - investment, funded, estate
    - loans, banking, mortgage
    - banking, trust, deposit
  - gas, property, oils
  - stores, drugs, product
  - services, solutions, software

Concept font size: 11 Show relation type Relation font size: 9

k-Means

No. suggestions: 4 Docs: ● All ○ Unused



- root
  - systems, manufactures, equipment
  - banking, loans, insurance
    - insurance, life, insurance_company
    - investment, funded, estate
    - loans, banking, mortgage
    - banking, trust, deposit
  - gas, property, oils
  - stores, drugs, product
  - services, solutions, software

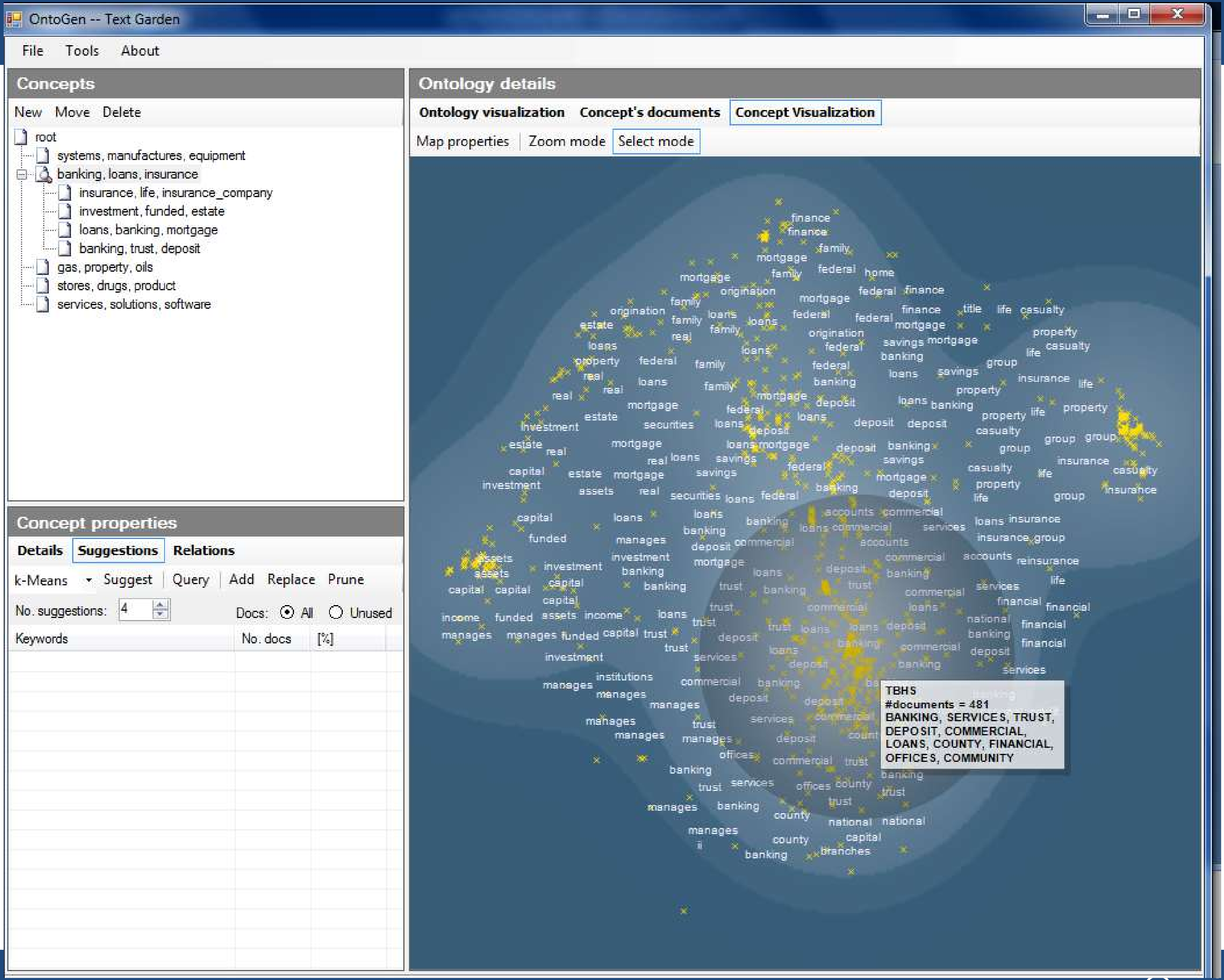
Document	Similarity
☑ GBCI -- Glacier Bancorp, Inc. is a multi-bank hol...	0.688
☑ JFBC -- Jeffersonville Bancorp is a bank holding...	0.684
☑ HMNF -- HMN Financial, Inc. is a stock savings...	0.678
☑ ALBY -- Community Capital Bancshares, Inc. is ...	0.667
☑ TRST -- TrustCo Bank Corp NY is a savings an...	0.665
☑ UBOH -- United Bancshares, Inc. is a one-bank...	0.658
☑ BNCN -- BNC Bancorp serves as a one-bank h...	0.658
☑ FOOT -- Foothill Independent Bancorp is a one-...	0.647
☑ CFB -- Commercial Federal Corporation is a unita...	0.646
☑ UMBF -- UMB Financial Corporation is a financi...	0.646
☑ NBTF -- NB&amp;T Financial Group, Inc. is a ba...	0.641
☑ PROV -- Provident Financial Holdings, Inc. serv...	0.638
☑ CHKJ.OB -- Cherokee Banking Company serves...	0.638
☑ MCBC -- Macatawa Bank Corporation is the hol...	0.631
☑ FULT -- Fulton Financial Corporation is the holdi...	0.631
☑ SOMC.OB -- Southern Michigan Bancorp, Inc. s...	0.630
☑ TSBK -- Timberland Bancorp, Inc. serves as the...	0.629
☑ SUSQ -- Susquehanna Bancshares, Inc. is a fin...	0.627
☑ MSFG -- MainSource Financial Group, Inc. is a ...	0.621
☑ WAYN -- Wayne Savings Bancshares, Inc. is a ...	0.620
☑ EVBS -- Eastern Virginia Bankshares, Inc. (EVB)...	0.619
☑ NBG -- National Bank of Greece S.A. (the Bank...	0.619
☑ HZFS.OB -- Horizon Financial Services Corporat...	0.619
☑ PULB -- Pulaski Financial Corp. is the holding c...	0.619
☑ HABC -- Habersham Bancorp is a bank holding ...	0.619
☑ PFBI -- Premier Financial Bancorp, Inc. is a multi...	0.618

Keywords for selected documents:

banking, loans, insurance, investment, mortgage, deposit, financial, trust, funded, commercialization

JFBC

Jeffersonville Bancorp is a bank holding company for First National Bank of Jeffersonville. The Company is engaged in the business of managing or controlling its subsidiary bank and such other business related to banking. The Bank is a full-service banking institution. Through the bank, the Company provides traditional banking related services. Banking services consist primarily of attracting deposits from the areas served by its banking offices and using those deposits to originate a variety of commercial, consumer and real estate loans. Its primary sources of liquidity are its deposit base; Federal Home Loan Bank (FHLB) borrowings; repayments and maturities on loans; short-term assets, such as federal funds and short-term, interest-bearing deposits in banks, and maturities and sales of securities available for sale. The Bank has one subsidiary, FNBJ Holding Corporation, which is a real estate investment trust (REIT) and is wholly owned by the Bank.



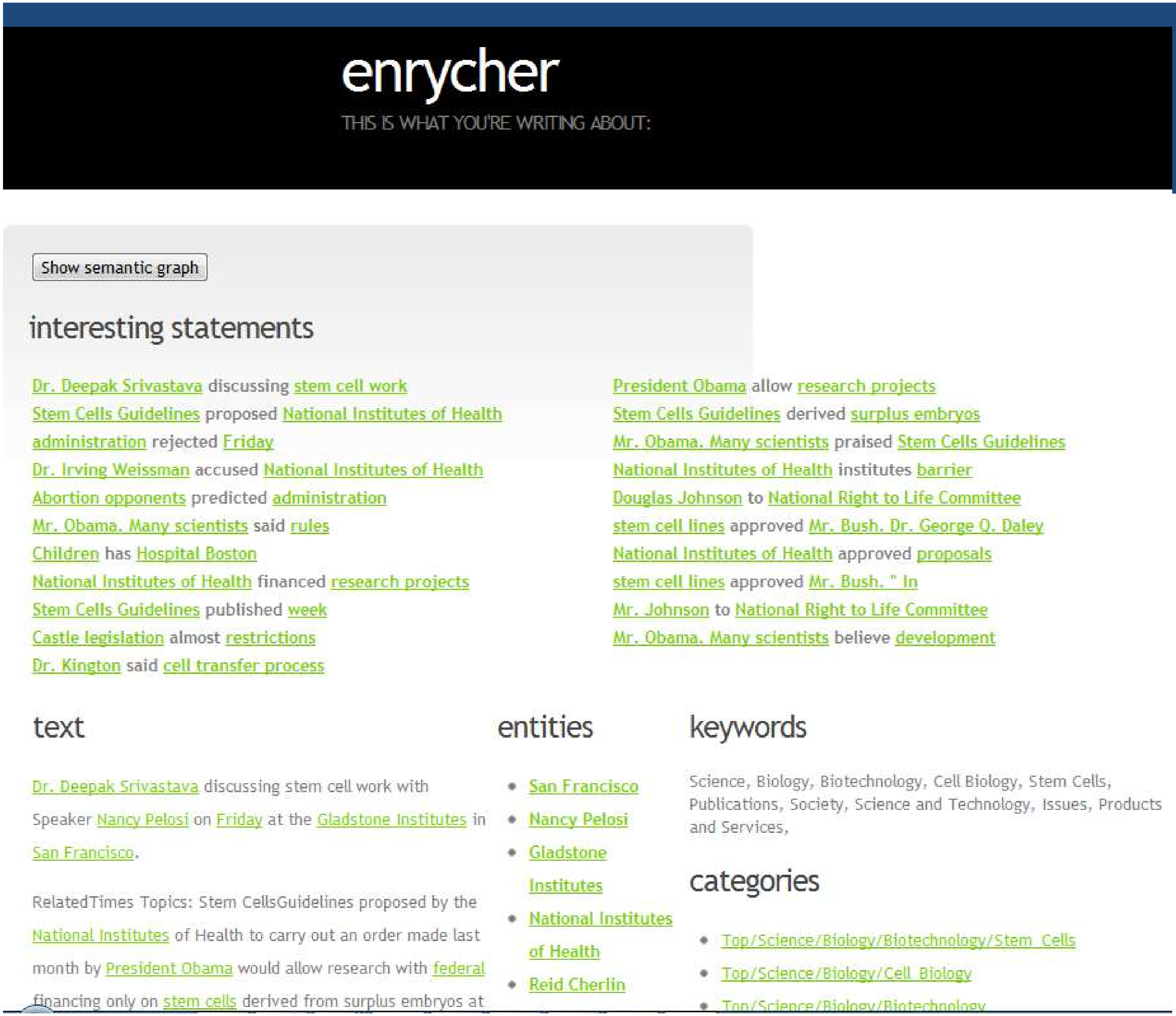
- root
  - systems, manufactures, equipment
  - banking, loans, insurance
    - insurance, life, insurance_company
    - investment, funded, estate
    - loans, banking, mortgage
    - banking, trust, deposit
  - gas, property, oils
  - stores, drugs, product
  - services, solutions, software

No. suggestions: 4

Docs: ● All ○ Unused

Keywords	No. docs	[%]

TBHS
#documents = 481
BANKING, SERVICES, TRUST, DEPOSIT, COMMERCIAL, LOANS, COUNTY, FINANCIAL, OFFICES, COMMUNITY



interesting statements

Dr. Deepak Srivastava discussing stem cell work
Stem Cells Guidelines proposed National Institutes of Health
administration rejected Friday
Dr. Irving Weissman accused National Institutes of Health
Abortion opponents predicted administration
Mr. Obama. Many scientists said rules
Children has Hospital Boston
National Institutes of Health financed research projects
Stem Cells Guidelines published week
Castle legislation almost restrictions
Dr. Kington said cell transfer process

President Obama allow research projects
Stem Cells Guidelines derived surplus embryos
Mr. Obama. Many scientists praised Stem Cells Guidelines
National Institutes of Health institutes barrier
Douglas Johnson to National Right to Life Committee
stem cell lines approved Mr. Bush. Dr. George Q. Daley
National Institutes of Health approved proposals
stem cell lines approved Mr. Bush. " In
Mr. Johnson to National Right to Life Committee
Mr. Obama. Many scientists believe development

text

Dr. Deepak Srivastava discussing stem cell work with Speaker Nancy Pelosi on Friday at the Gladstone Institutes in San Francisco.

RelatedTimes Topics: Stem CellsGuidelines proposed by the National Institutes of Health to carry out an order made last month by President Obama would allow research with federal financing only on stem cells derived from surplus embryos at

entities

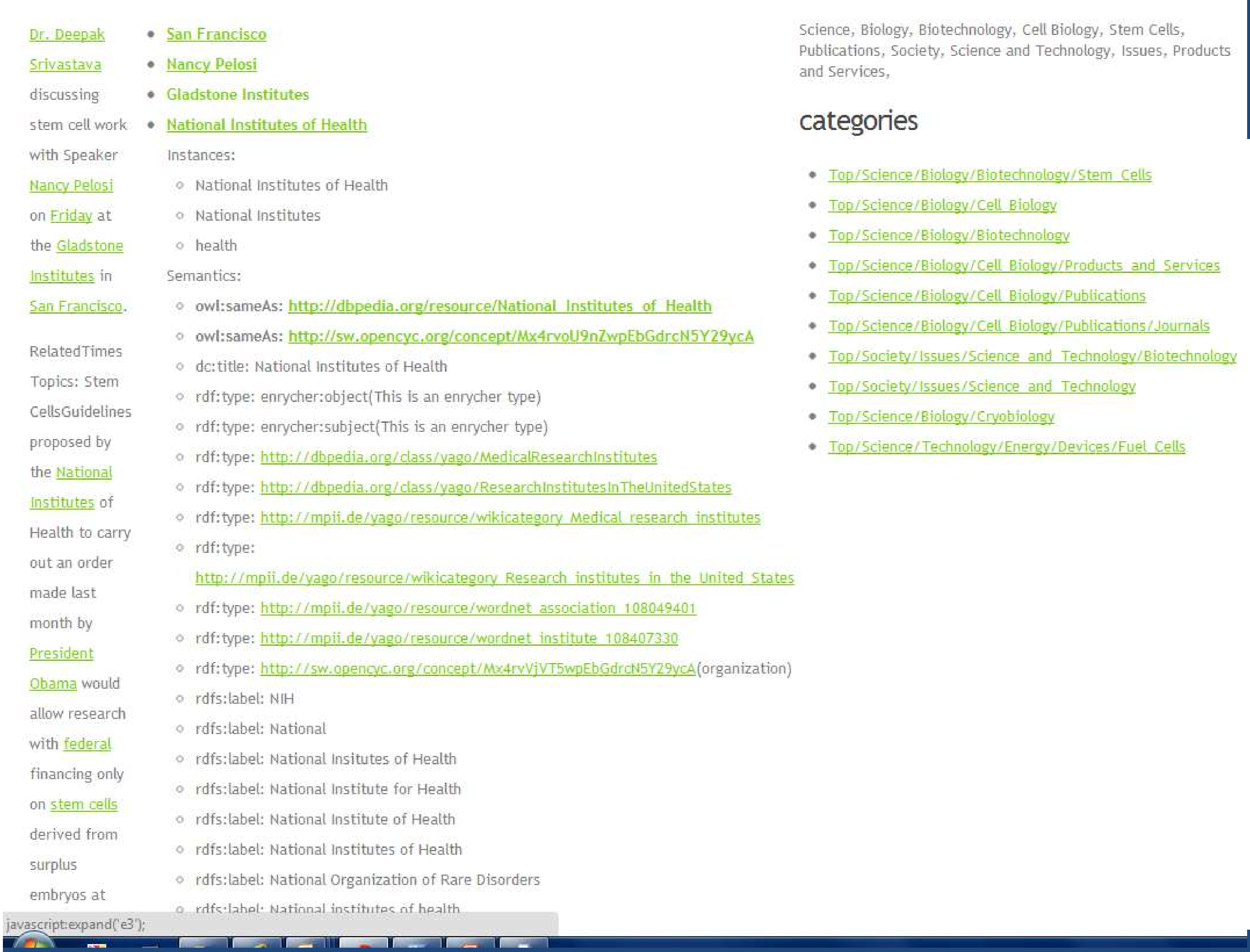
- San Francisco
- Nancy Pelosi
- Gladstone Institutes
- National Institutes of Health
- Reid Cherlin

keywords

Science, Biology, Biotechnology, Cell Biology, Stem Cells, Publications, Society, Science and Technology, Issues, Products and Services,

categories

- Top/Science/Biology/Biotechnology/Stem_Cells
- Top/Science/Biology/Cell_Biology
- Top/Science/Biology/Biotechnology



Dr. Deepak Srivastava discussing stem cell work with Speaker Nancy Pelosi on Friday at the Gladstone Institutes in San Francisco.

RelatedTimes Topics: Stem CellsGuidelines proposed by the National Institutes of Health to carry out an order made last month by President Obama would allow research with federal financing only on stem cells derived from surplus embryos at

- San Francisco
- Nancy Pelosi
- Gladstone Institutes
- National Institutes of Health

Instances:

- National Institutes of Health
- National Institutes
- health

Semantics:

- owl:sameAs: http://dbpedia.org/resource/National_Institutes_of_Health
- owl:sameAs: http://sw.opencyc.org/concept/Mx4rvoU9nZwpEbGdrcN5Y29ycA
- dc:title: National Institutes of Health
- rdf:type: enrycher:object(This is an enrycher type)
- rdf:type: enrycher:subject(This is an enrycher type)
- rdf:type: http://dbpedia.org/class/yago/MedicalResearchInstitutes
- rdf:type: http://dbpedia.org/class/yago/ResearchInstitutesInTheUnitedStates
- rdf:type: http://mpii.de/yago/resource/wikicategory_Medical_research_institutes
- rdf:type: http://mpii.de/yago/resource/wikicategory_Research_institutes_in_the_United_States
- rdf:type: http://mpii.de/yago/resource/wordnet_association_108049401
- rdf:type: http://mpii.de/yago/resource/wordnet_institute_108407330
- rdf:type: http://sw.opencyc.org/concept/Mx4rvVjVT5wpEbGdrcN5Y29ycA(organization)
- rdfs:label: NIH
- rdfs:label: National
- rdfs:label: National Insitutes of Health
- rdfs:label: National Institute for Health
- rdfs:label: National Institute of Health
- rdfs:label: National Institutes of Health
- rdfs:label: National Organization of Rare Disorders
- rdfs:label: National institutes of health

javascript:expand('e3');

Science, Biology, Biotechnology, Cell Biology, Stem Cells, Publications, Society, Science and Technology, Issues, Products and Services,

categories

- Top/Science/Biology/Biotechnology/Stem_Cells
- Top/Science/Biology/Cell_Biology
- Top/Science/Biology/Biotechnology
- Top/Science/Biology/Cell_Biology/Products_and_Services
- Top/Science/Biology/Cell_Biology/Publications
- Top/Science/Biology/Cell_Biology/Publications/Journals
- Top/Society/Issues/Science_and_Technology/Biotechnology
- Top/Society/Issues/Science_and_Technology
- Top/Science/Biology/Cryobiology
- Top/Science/Technology/Energy/Devices/Fuel_Cells

Python Toolkit - NLTK

- Book at http://www.nltk.org/
- Useful scripts in nltk-trainer
- You can also combine nltk and sklearn to do pre-processing in nltk and classification in sklearn
- http://streamhacker.com/tag/classification/

Online Resources

- Interactive clustering tool Ontogen: ontogen.ijs.si
- Tutorial at http://eprints.pascal-network.org/archive/00000017/01/Tutorial_Marko.pdf
- Demos
  - http://cogcomp.cs.illinois.edu/page/demos/
  - https://dandelion.eu/semantic-text/text-classification-demo
- List of commercial software http://www.kdnuggets.com/software/text.html

Python Toolkit - NLTK

- Book at http://www.nltk.org/

- Useful scripts in nltk-trainer

http://brandonrose.org/clustering

Open Source NLP Toolkits

- Apache OpenNLP
- Natural Language Toolkit (NLTK)
- Standford NLP
- MALLET

Online Resources

- Interactive clustering tool Ontogen: ontogen.ijs.si
- Tutorial at http://eprints.pascal-network.org/archive/00000017/01/Tutorial_Marko.pdf
- Online lectures at videolectures.net
- List of commercial software http://www.kdnuggets.com/software/text.html