An Introduction to Text Mining

Nature Inspired Technologies Group

Outline of the presentation

- What is text mining? Why is it useful?
- What are main goals and tasks of text mining?
- Why does text differ from other kinds of data?
- Document collections, documents and their representation
- How to ensure basic text mining tasks:
  - Classification
  - Clustering
  - ...
- Text mining methods

What Is Text Mining?

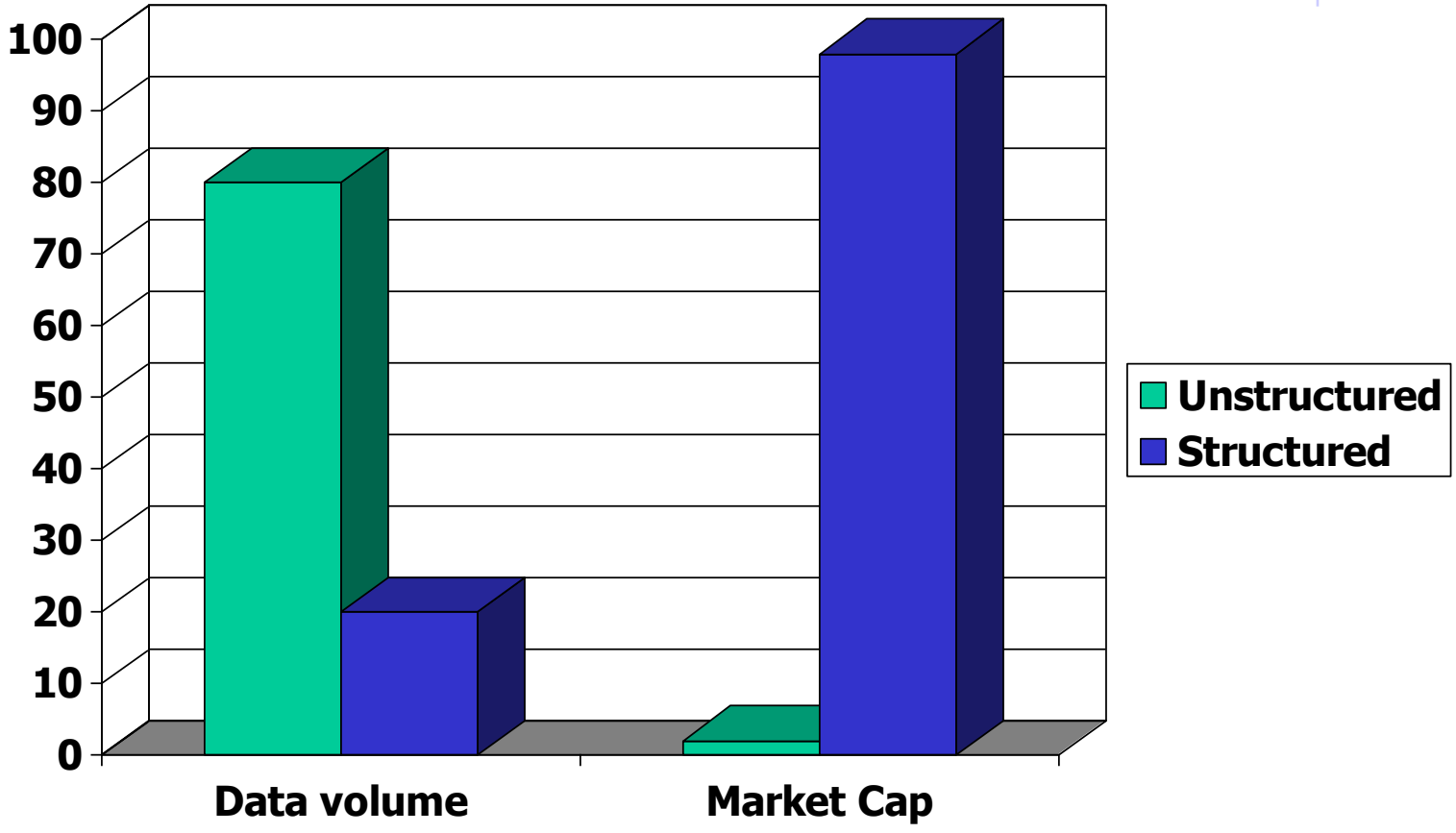
"The objective of Text Mining is to exploit information contained in textual documents in various ways, including …discovery of patterns and trends in data, associations among entities, predictive rules, etc." (Grobelnik et al., 2001)

"Another way to view text data mining is as a process of exploratory data analysis that leads to heretofore unknown information, or to answers for questions for which the answer is not currently known." (Hearst, 1999)

Challenges in Text Mining

- Data collection is "free text" (not well-organized): **Semi-structured or unstructured data**
- **Complex structure can hide the real content** : e.g. problems with negation call for using syntactic analysis!
- Natural language text contains **ambiguities** on many levels :
  - lexical – a.of words (*diamond*, ..)
  - syntactical – "*Read about the problem in newspapers* !"
  - referential - "*The boy was passing a man with a dog. **He** stroked **him**.* " - anaphora
  - …

The business opportunity in text mining…



Where is Corporate Knowledge "Gold"?

Most often in stuff not very accessible via standard data-mining

- Email
- Insurance claims
- News articles
- Web pages
- Patent portfolios
- IRC
- Scientific articles
- Customer complaint letters
- Contracts
- Technical documents
- Transcripts of phone calls with customers

Text Mining

- How does it relate to data mining in general?
- How does it relate to computational linguistics?
- How does it relate to information retrieval?

	Finding Patterns	Finding "Nuggets"	
		Novel	Non-Novel
Non-textual data	General data-mining	Exploratory Data Analysis	Database queries
Textual data	Computational Linguistics		Information Retrieval

Text Mining Tasks

- Exploratory Data Analysis
  - Using text to form hypotheses about diseases (Swanson and Smalheiser, 1997).
- Information Extraction
  - (Semi)automatically create (domain specific) knowledge bases, and then use standard data-mining techniques.
    - Bootstrapping methods (Riloff and Jones, 1999).
- Text Classification
  - Useful intermediary step for information extraction
    - Bootstrapping method using EM (Nigam et al., 2000).

Challenges in Data Exploration

- How can valid inference links be found without succumbing to combinatorial explosion of possibilities?
  - Need better models of lexical relationships and semantic constraints (very hard)
- How should the information be presented to the human experts to facilitate their exploration?

Information Extraction (IE)

- Extract domain-specific information from natural language text using
  - A dictionary of extraction patterns (e.g., "traveled to <x>" or "presidents of <x>")
    - Constructed by hand
    - Automatically learned from hand-annotated training data
  - A semantic lexicon (dictionary of words with semantic category labels)
    - Typically constructed by hand

Challenges in IE

- Automatic learning methods are typically supervised (i.e., need labeled examples)
- But annotating training data is a time-consuming and expensive task.
- Can we develop better unsupervised algorithm?
- Can we make better use of a small set of labeled examples?

Text classification (TC)

- Tag a document as belonging to one of a set of pre-defined classes
  - "This does not lead to discovery of new information..." (Hearst, 1999).
  - Many practical uses
    - Group documents into different domains (useful for domain specific information extraction)
    - Learn reading interests of users
    - Automatically sort e-mail
    - On-line New Event Detection

Challenges in TC

- Like IE, also needs lots of labeled examples as training data
  - After a user has labeled 1000 UseNet news articles, the system was only right ~50% of the time at selecting articles interesting to the user.
- What other sources of information can reduce the need for labeled examples?

Documents and Document Collections

- **Document collection =** a grouping of text based documents which is either *static* or *dynamic* (change in time).
  - e.g. **PubMed** "on-line repository of abstracts for > 13 million papers (about 40.000 new abstracts are added each month)
- **Document =** a unit of discrete textual data within a collection.
  - e.g. a business report, research paper, news story, e-mail, ..
- Single document can be included in different document collections (e.g. e-health papers can appear both in health document collections and in those on ICT)

Document Representation

- **How can we search** for a particular document in PubMed? *Keyword search* is not very useful:
  - *protein* or *gene* return more than 3 million documents
  - Even very specific term *epidermal growth factor receptor* →10.000 doc.!
- What **features** can we use to represent a document (so that some DM algorithms can be applied, e.g. clustering) ?
- **Content would be the best choice!** But natural language processing (NLP) and understanding is one of the big AI challenges (Turing test). Why?
  - Many ambiguities which have to be resolved using context information: very demanding!
  - Negation, …

Features for Doc. Representation

- **Characters**
  - enabling to recognize e.g.morphological issues (e.g.for text prediction)
  - bigrams (trigrams) represent sequences of 2 (or 3) characters
- **Words**
  - Often the term word-level tokens is used instead
  - Tokens can be annotated (e.g. with labels representing noun, verb,..)
  - bag-of-words representation ignores the order of tokens
  - word stem represents a group of words stripped of a suffix
- **Terms** may represent single words or multiword units, e.g. "White house"
- **Concepts**
  - represent groups of entities to resolve problems with synonyms, ..
  - Identifier "car" can represent different words in a text: *automobile*, *truck*, *Lightning McQueen*

Problems of High Dimensionality

- These features make it possible to represent a document as a vector of words (or terms)
  - Each component of the vector represents some "***quantity***" related to a single word.

- Very useful simplification, BUT:
  - Words or terms result **in large number of features**:
    - Small Reuters collection of 15.000 documents contains 25.000 non-trivial features (word stems)
    - Most algorithms do not deal with large numbers of features → need to employ feature reduction techniques
  - **Feature sparsity**: Each document contains only very limited number of all potential features

Basic Text Mining Tasks

- Feature extraction (from a document) finds a good subset of words that represent this document best.
- Document classification (categorization)
- Information retrieval
- Clustering / organization of documents
- Information extraction
- ...

Feature Extraction:Task

While more and more textual information is available on-line, effective retrieval is difficult without good indexing of text content.

20

16

While-more-and-textual-information-is-available-online-effective-retrieval-difficult-without-good-indexing-text-content

Feature Extraction

Number of words

5

Text-information-online-retrieval-index

2 1 1 1 1

term frequency in the original text

Nature Inspired Technologies Group

Further refinements to the DV representation

- Not all words are equally important
  - *the, is, and, to, he, she, it (*Why?*)*
  - Of course, these words could be important in certain contexts
- We have the option of scaling the components of these words, or completely removing them from the corpus
- In general, we prefer to remove the **stopwords** and scale the remaining words
  - ***TF*** stands for *Term Frequency*
  - Important words should be scaled upwards, and vice versa: One widely used scaling factor – **TF_IDF**
  - where **IDF** is *Inverse Document Frequency* product, for a word**.**

2.2 Feature Extraction:Indexing(1)

Training documents

Identification all unique words

Removal stop words

- non-informative word
- ex.{the,and,when,more}

Word Stemming

- Removal of suffix to generate word stem
- grouping words
- increasing the relevance
- ex.{*walker,walking*}→*walk*

Term Weighting

- Naive terms
- Importance of term in Doc

Nature Inspired Technologies Group

FE: Indexing(2)+Weighting Model

- Document representations: vector space models

$$d=(w_1,w_2,\ldots w_t)\in\mathbf{R}^t$$

w_i is the weight of ith term in the document d.

- **tf - Term Frequency weighting**

  w_{ij} = $Freq_{ij}$

  $Freq_{ij}$: := the number of times jth term occurs in d_i.

× **Drawback**: no reflection of importance
low factor for document discrimination.

ABRTSAQWA
XAO

RTABBAXA
QSAK

	A	*B*	*K*	*O*	*Q*	*R*	*S*	*T*	*W*	*X*
D1	4	1	0	1	1	1	1	1	1	1
D2	4	2	1	0	1	1	1	1	0	1

FE:Weighting Model(2)

tf×idf - Inverse Document Frequency weighting

$$w_{ij} = Freq_{ij} * \log(N / DocFreq_j)$$

N := number of documents in the training document collection.

$DocFreq_j$:= the number of documents where the *j*th term occurs.

✓Advantage with reflection of importance factor for document discrimination.

Assumption:terms with low DocFreq are better discriminator than ones with high DocFreq in document collection

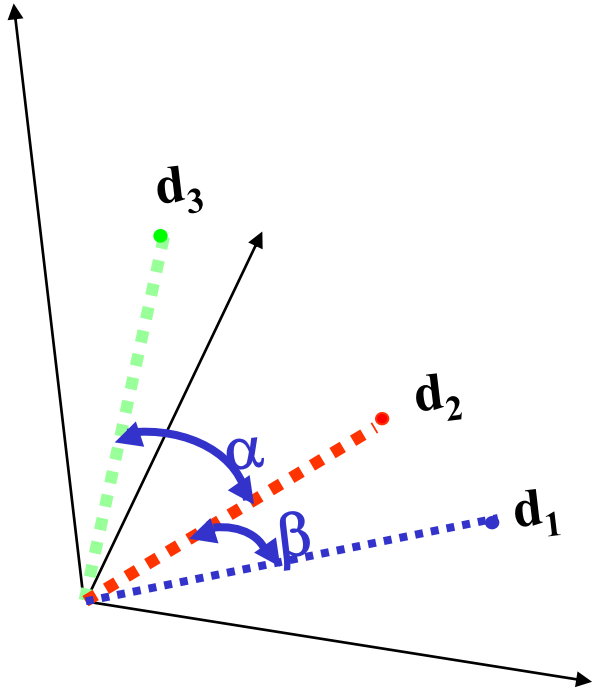
ABRTSAQWA
XAO

RTABBAXA
QSAK

	A	*B*	*K*	*O*	*Q*	*R*	*S*	*T*	*W*	*X*
D1	0	0	0	0.3	0	0	0	0	0.3	0
D2	0	0	0.3	0	0	0	0	0	0	0

Similarity of Document Vectors?

- docA "*Java Programming Language*" <1, 1, 0, 0, 1, 0, 0>
- docB "*Barcelona beats Real Madrid*" <0, 0, 1, 1, 0, 1, 0>
- docC "*Barcelona beats Slavia*" <0, 0, 1, 1, 0, 0, 1>
- What vector operation can you think of to find two *similar documents*?

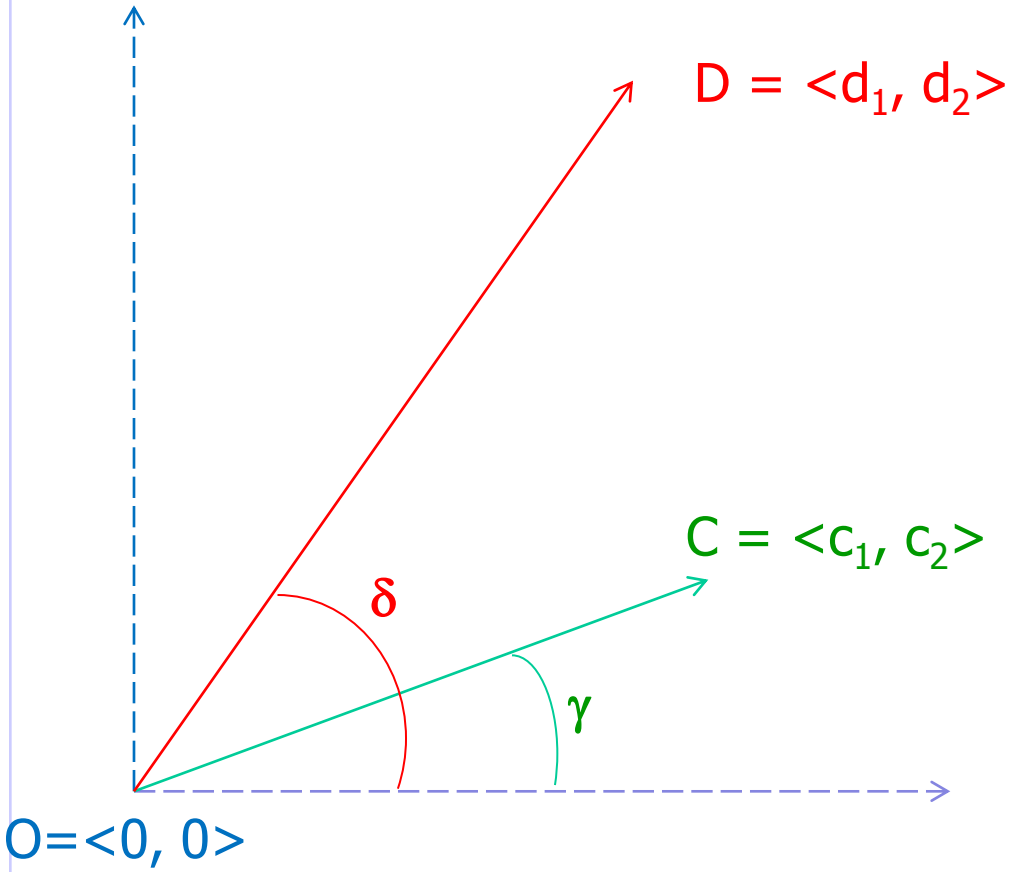


If $\alpha > \beta$ than $\cos(\alpha) < \cos(\beta)$

d_2 is closer to d_1 than d_3

Cosine-based similarity model can reflect the *relations between features*.

Distance of 2 vectors v_1 and v_1 in a plane?



What is the cosine distance of 2 vectors **c** (= OC) and **d** (= OD) in a plane? $\cos(\delta - \gamma)$

Let $|\mathbf{c}|$ be the length of the vector **c**:

$|\mathbf{c}| = \sqrt{(c_1 * c_1 + c_1 * c_2)}$

$\cos(\delta - \gamma) = \cos\delta * \cos\gamma + \sin\delta * \sin\gamma$

$= c_1 * d_1 / |c| * |d| + c_2 * d_2 / |c| * |d|$

$= (c_1 * d_1 + c_2 * d_2) / |c| * |d|$

=(scalar product of **c** and **d**)/$|\mathbf{c}| * |\mathbf{d}|$

Model:Centroid-Based Classifier

1.Input:new document ***d*** $=(w_1, w_2,\ldots,w_n)$;

2.Predefined categories: $C=\{c_1,c_2,\ldots,c_l\}$ specified by the corresponding sets of documents.

3.//Compute centroid vector $\vec{C}_i$ for each category c_i

4.//Similarity model - cosine function

$$Simil(d_i,d_j)=\cos(d_i,d_j)=\frac{d_i \bullet d_j}{\|d_i\|_2 \times \|d_j\|_2}=\frac{\sum(w_{ik}\times w_{jk})}{\sqrt{\sum w^2{}_{ik}}\times\sqrt{\sum w^2{}_{jk}}}$$

5.//Compute similarity

$$Simil(\vec{c}_i,d)=\cos(\vec{c}_i,d)$$

6.//Output:Assign to document ***d*** the **category** c_{max}

$$Simil(\vec{c}_i,d)\le Simil(c_{\max},d)$$

Text Mining Applications

- Document/Term Clustering
  - Given a large set, group similar entities
- Text Classification
  - Given a document, find what topic does it talk about
- Information Retrieval
  - Search engines
- Information Extraction
  - Question Answering

Classification

The problem statement

- Given a set of documents, each with a *label* called the class label for that document
- Given, a classifier which ***learns*** from the above data set
- For a new, unseen document, the classifier should be able to "predict" with a high degree of accuracy the correct class to which the new document belongs

Solutions?

- **Decision tree learning**: Good for text mining. But doesn't scale
- Bayessian classification
- **Support Vector Machine** finds the **best** discriminant (`*plane*` between two classes
- Neural Networks, **Case-based reasoning**

Web Text Mining

- The WWW is a huge, directed graph, with documents as nodes and hyperlinks as the directed edges
- Apart from the text itself, this graph structure carries a lot of information about the "usefulness" of the "nodes"
- For example
  - 10 random, average people on the streets say Mr. T. Ache is a good dentist
  - 5 reputed doctors, including dentists, recommend Mr. P. Killer as a better dentist
  - Who would you choose?

Some more topics *we haven't touched*

- Using external dictionaries, e.g.**WordNet**
- Using language specific techniques
  - Computational linguistics
  - Use grammar for judging the "sense" of a query in the "information retrieval" scenario
- Question answering techniques
- Available applications of TM, see e.g.

http://core.kmi.open.ac.uk/search

- Some "purists" do not consider most of the current activities in the text mining field as *real text mining* – they search for something really innovative!

Some more comments

PubMed http://www.ncbi.nlm.nih.gov/pubmed/ is a free resource developed and maintained by the Nat. Center for Biotechnology Information (NCBI) at the USA National Library of Medicine® (NLM).

It comprises more than 21 million citations for biomedical literature from MEDLINE, life science journals, and online books.

Example of "innovative" TM in PubMed documents will be described in the next lecture.