Text Analytics

Module 1: Introduction to Text Analytics

Dr. Fan Zhenzhen
NUS-ISS
National University of Singapore

Objectives of this module

At the end of this module, you can:

- Describe the tasks in text analytics with examples
- Define the process to perform text analytics based on the business requirements and text analytics artifacts

Outline

- What is text analytics?
- What can text analytics do?
- The text analytics process
- Tools & solutions for text mining
- Exercise

© 2013-2024 NUS. All rights reserved.

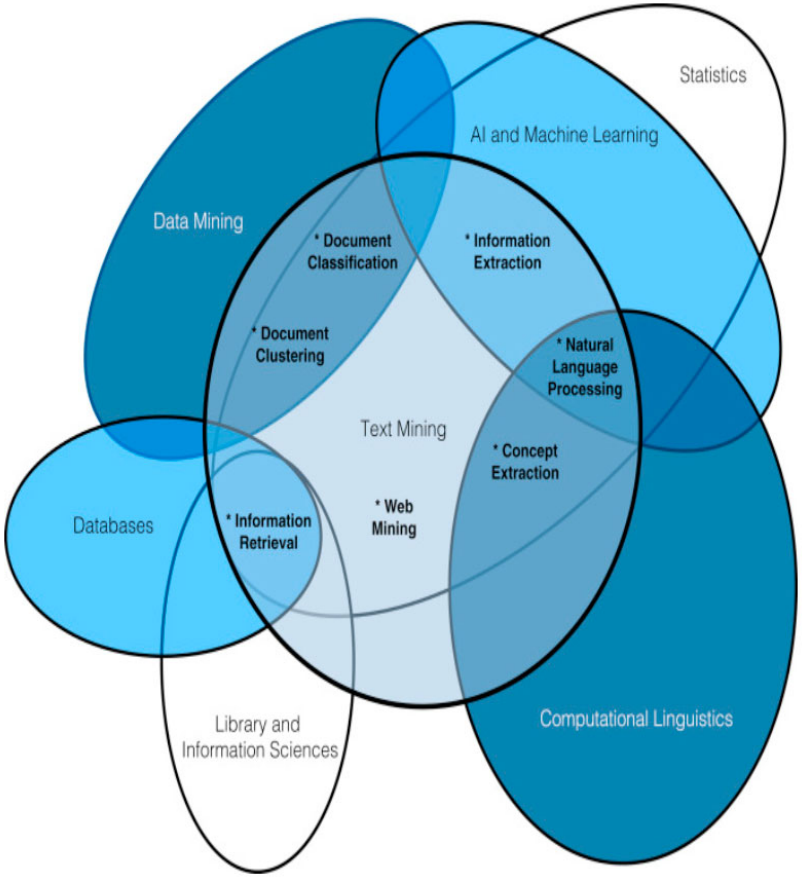
What is Text Analytics

What is Text Analytics?

- Also known as **Text Mining**

- The task of deriving insights, trends and patterns from large amounts of unstructured text data to answer questions in a business context

- The process of transforming unstructured text data into structured (numerical) data so that automatic algorithms can be applied to extract out information in an understandable structure for further use.



Practical Text Mining and Statistical Analysis for non-Structured Text Data Applications

© 2013-2024 NUS. All rights reserved.

Unstructured to structured

Cust ID	Date:time	Model	Comments
00010	201212 03:2201	8560	Doesn't work – may have dropped. Out of warranty. Sent to svc.
00023	201212 03:1034	8850	Cannot roam. System is enabled. Reset settings on phone. Done.
00025	201212 03:1640	2338	No sound. Rebooted many times. Sent to svc. 3 months old.
01003	201212 03:1030	6000-1	Bought 2 weeks back. Gift. No receipt. Wants to upgrade. Sent to svc.
20456	201212 03:1025	6000-1	Out of space. 4GB uSD. Set default save to uSD for songs. Done.

Cust ID	Date:time	Model	Svc	Closed	Issue	. . . etc
00010	201212 03:2201	8560	1	1	99	
00023	201212 03:1034	8850	0	1	45	
00025	201212 03:1640	2338	1	1	12	
01003	201212 03:1030	6000-1	1	1	99	
20456	201212 03:1025	6000-1	0	1	28	

Text Mining converts unstructured text fields into one or more columns of easily processed numeric data

© 2013-2024 NUS. All rights reserved.

Why text analytics?

- Data analytics works
  - Most information in the world is not in structured data form
  - The information in text needs to be unlocked
- Text is being created in digital format and available
  - Formal documents: word processing
  - Semi-structured text documents: patents, websites, …
  - Informal text: email, social media, sms, tweets, …
- Analyzing text, by itself or in conjunction with data, has become an important part of many business intelligence and analytics processes, for better data-driven decision making

© 2013-2024 NUS. All rights reserved.

What can Text Analytics Do?

Common applications of text analytics

- **Customer experience management**: to analyze customer feedback, measure customer satisfaction, identify issues and improve customer experience.
- **Brand monitoring**: to monitor brand mentions on social media and other online platforms.
- **Market research:** to analyze customer feedback and identify trends in the market and discover new opportunities.
- **Fraud detection**: to detect fraudulent activities by analyzing text data in claims/reports.
- **Healthcare**: to analyze medical records, clinical notes, and patient feedback, and improve healthcare outcomes.
- Etc.

© 2013-2024 NUS. All rights reserved.

What can text analytics do?

- Sentiment analysis
  - Determine the sentiment of a piece of text, commonly applied to analyze customer feedback, social media posts, and other types of text data.
  - What are my customers saying about me? What are the areas of concern to a target group? Analyzing open-ended responses to survey questions
- Text classification
  - Automatically put text into predefined categories for subsequent processing
  - Email routing, spam filtering; news filtering; fraud detection, normally flag for human intervention
- Improve accuracy in predictive modeling or unsupervised learning
  - Derive useful text features to be used with structured data for better “lift”
- Document search
  - Identify specific or similar/relevant documents
- Extract specific information from the text
  - For tasks like QA (given a question, get the answer from the text)
  - Named Entity Recognition (NER) for dates, money sums, organizations, stock symbols, etc.

© 2013-2024 NUS. All rights reserved.

Example 1

From: http://www.youtube.com/watch?v=PSq7hZ0shLs

© 2013-2024 NUS. All rights reserved.

Things to Note

This does the hard work

New Sheet: Macro

Sheet Name: Sheet1

LW - Sentiment Analysis

This is a Languageware UDF for sentiment analysis.

Fill in parameters:

content*

text

type*

com.ibm.DictTrigger
com.ibm.Watchlist
com.ibm.NegativeIndicator
com.ibm.PositiveIndicator
com.ibm.QuestionIndicator
com.ibm.TwitterID
com.ibm.URL

Parameters | Carry Over

A	B		C
id	name		created_at
628	????		0 19:55:06 +0000 2010
029	waldheins		15 22:34:43 +0000 2010
126	Curt Hall		1 17:08:03 +0000 2010
352	Alan Phillips		0 19:55:05 +0000 2010
165	Bryan Hammond		15 22:34:43 +0000 2010
355	Curt Hall		1 17:08:03 +0000 2010
765	Martin Richard		0 19:55:06 +0000 2010
394	?????? Bieber		15 22:34:43 +0000 2010
624	Curt Hall		1 17:08:04 +0000 2010
155	KickPost		0 19:55:06 +0000 2010
709	Leonidas Koustimpis		15 22:34:45 +0000 2010
855	Curt Hall		1 17:08:04 +0000 2010
244	Black.Mamba		0 19:55:06 +0000 2010
290	Jennifer ?		15 22:34:46 +0000 2010
679	Tweets Espa□a		1 17:08:04 +0000 2010
341	Hairulnizar		0 19:55:06 +0000 2010
785	J.As	R_Angel_9	Wed Sep 15 22:34:46 +0000 2010

© 2013-2024 NUS. All rights reserved.

Things to Note

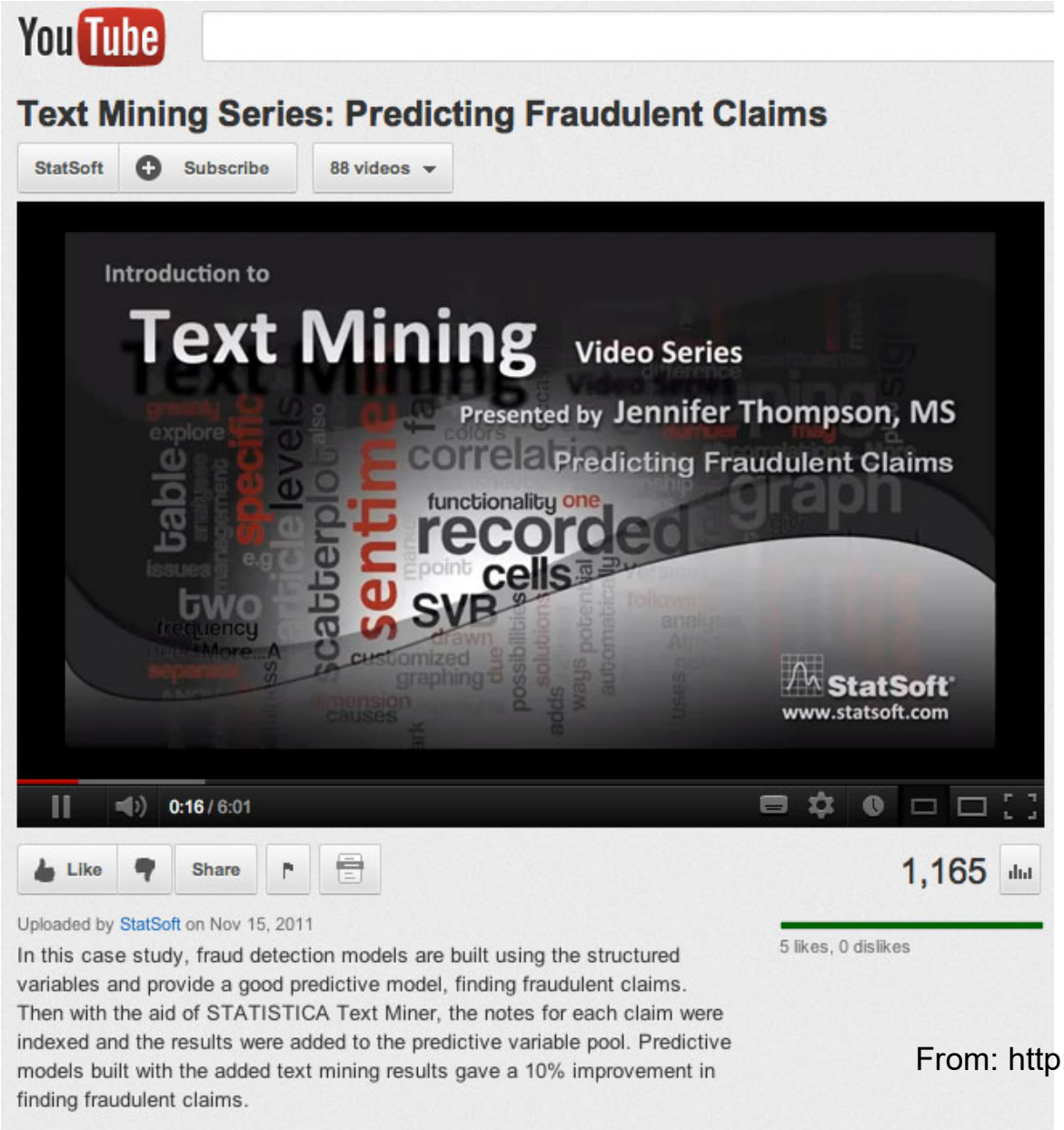
Result Data: Ready

	type	name	screen_name	
1	com.ibm.en.PositiveIndicator	Special	Marlen1929	@m
2	com.ibm.en.PositiveIndicator	fantastic	Curtsiphone	Mak
3	com.ibm.en.PositiveIndicator	glorious	sharding	@Ke
4	com.ibm.en.PositiveIndicator	Amazing	thaibisz	Top
5	com.ibm.en.PositiveIndicator	like	MagsJB	My a
6	com.ibm.en.PositiveIndicator	Cool	Cellphonez	Iph
7	com.ibm.en.PositiveIndicator	best	THE_Efram	@HI
8	com.ibm.en.PositiveIndicator	Quick	Leesa19043	@ilo
9	com.ibm.en.PositiveIndicator	First	paladigarisbiz	(HO
10	com.ibm.en.PositiveIndicator	wow	sjbuchanan007	@Ar
11	com.ibm.en.PositiveIndicator	Fast	Leesa19043	@ilo
12	com.ibm.en.PositiveIndicator	great	grattonboy	I'm
13	com.ibm.en.NegativeIndicator	doubt	gadgetinn	App
14	com.ibm.en.NegativeIndicator	hurt	msluvmylife	Girl
15	com.ibm.en.PositiveIndicator	like	POSTMODERNISM_	Rec
16	com.ibm.en.PositiveIndicator	like	gadgetinn	App

Decides sentiment based on lists of built-in keywords

© 2013-2024 NUS. All rights reserved.

Example 2



From: http://www.youtube.com/watch?v=OlQpm8qTog4

© 2013-2024 NUS. All rights reserved.

Things to Note

Predicting Fraudulent Claims

- Can predictability of fraudulent claims be improved by adding Text Mining results?
- Variables for analysis include
  - Delay in report
  - Policy age
  - Unusual injury
  - Minor injury
  - Text notes

Data: Fraud.sta (11v by 2732c)

	1 Fraud	2 Delay _in_ Report	3 Date _of_ Loss	4 Vehicle _Year	5 Age _In_ Days	6 Vehicle _Age	7 Unusual _Injury	8 Minor Injury	9 Text Notes	10 Cause	11 stratified sample
1	Yes	5	1/17/07	2010	1510	1	1	0	Car had right-of-way	'B' cited but 'A' wa	train
2	Yes	4	3/1/05	2009	2196	2	1	0	Car had right-of-way	Accident per inve	train
3	No	0	7/23/09	2006	591.1	5	0	0	Left turn other than	Attempting to pas	exclude
4	No	0	4/10/09	1998	695.1	13	0	0	Left turn other than	Attempting to pas	exclude
5	Yes	0	4/26/06	2009	1775	2	1	0	Car had right-of-way	Attempting to pas	train
6	Yes	1	12/3/04	1995	2285	16	1	1	Car had right-of-way	Attempting to pas	train
7	Yes	2	2/23/06	2008	1838	3	0	1	Hit by other car - no	Attempting to pas	train

Data: fraud with TM results.sta* (39v by 2732c)

Text Miner results: may be used as Input spreadsheet for subsequent analyses

	1 Fraud	2 Delay _in_ Report	3 Age _In_ Days	4 Unusual _Injury	5 Minor Injury	6 stratifi samp	7 appli	8 better	9 call	10 car	11 client	12 descript	13 hit	14 injuri	15 intersect	16 lawyer	17 legal	18 move	19 neck	20 park	21 rear	22 right	23 right-of -way	24 stop	25 way
1	Yes	5	1509.6	1	0	train			1.596	1.465	1.596			1.596		1.5938			1.596				1.1363		
2	Yes	4	2196.03	1	0	train			1.596	1.465	1.596			1.596		1.5938			1.596				1.1363		
3	No	0	591.108	0	0	exclude																			
4	No	0	695.129	0	0	exclude																			
5	Yes	0	1775.05	1	0	train			1.596	1.465	1.596			1.596		1.5938			1.596				1.1363		
6	Yes	1	2284.79	1	1	train			1.596	1.465	1.596			1.596		1.5938			1.596				1.1363		
7	Yes	2	1837.84	0	1	train	4.357	4.357		1.465		4.30187	0.85												
8	Yes	3	880.781	0	0	train			1.596	1.465	1.596			1.596		1.5938			1.596				1.1363		
9	Yes	0	376.028	1	1	train			1.596	1.465	1.596			1.596		1.5938			1.596				1.1363		
10	Yes	4	339.661	1	1	train			1.596	1.465	1.596			1.596		1.5938			1.596				1.1363		
11	No	0	684.083	0	0	exclude							0.85				0.902				0.86			0.9	
12	No	0	408.183	0	0	test			1.596	1.465	1.596			1.596		1.5938			1.596				1.1363		
13	Yes	0	1922	1	0	test			1.596	1.465	1.596			1.596		1.5938			1.596				1.1363		
14	Yes	1	2560.34	0	0	test			1.596	1.465	1.596			1.596		1.5938			1.596				1.1363		
15	Yes	7	2043.96	1	0	train			1.596	1.465	1.596			1.596		1.5938			1.596				1.1363		
16	Yes	0	1463.17	1	1	train			1.596	1.465	1.596			1.596		1.5938			1.596				1.1363		
17	Yes	2	1818.81	1	1	train			1.596	1.465	1.596			1.596		1.5938			1.596				1.1363		
18	No	0	998.468	0	0	exclude			1.596	1.465	1.596			1.596		1.5938			1.596				1.1363		
19	No	6	573.433	0	1	train							0.85				0.902				0.86			0.9	
20	No	2	571.677	0	1	train			1.596	1.465	1.596			1.596		1.5938			1.596				1.1363		
21	Yes	0	2486.18	1	1	test			1.596	1.465	1.596			1.596		1.5938			1.596				1.1363		
22	Yes	2	1484.84	1	1	train			1.596	1.465	1.596			1.596		1.5938			1.596				1.1363		

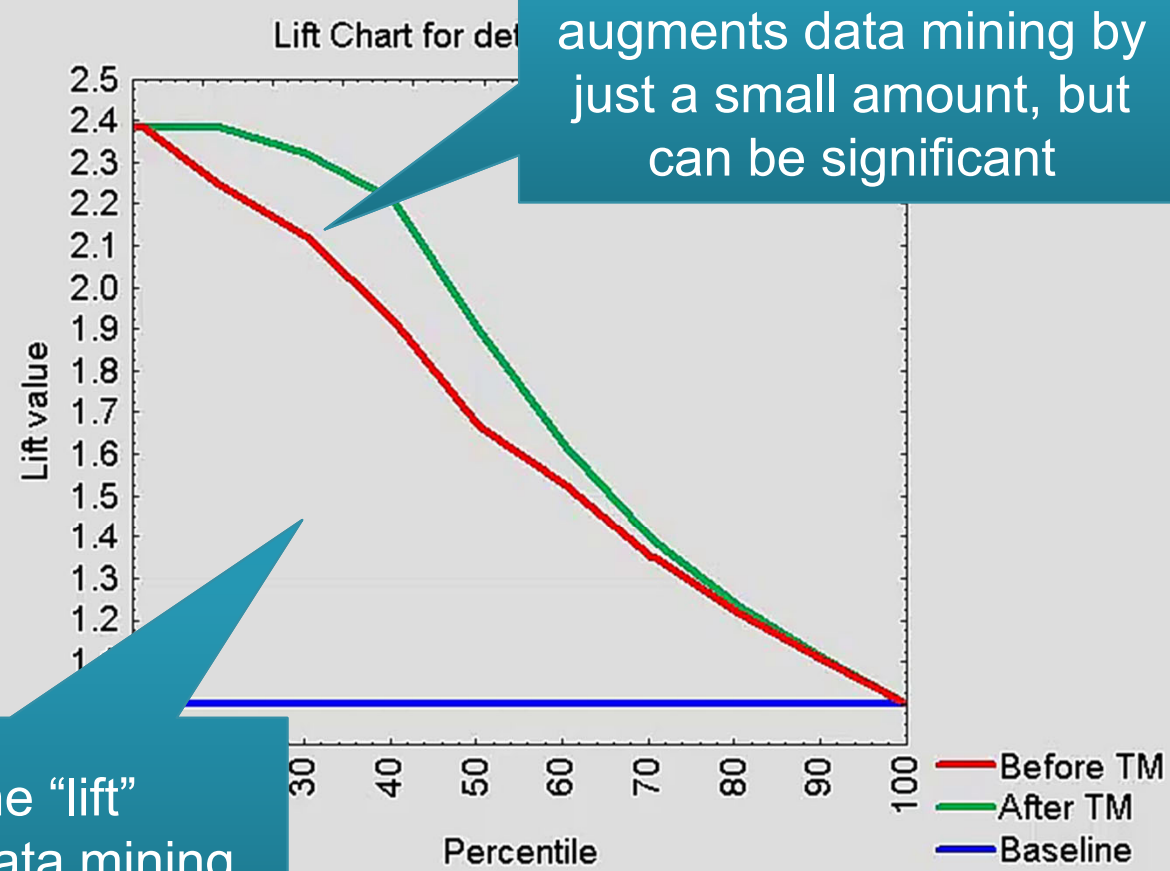
Text converted into numbers

© 2013-2024 NUS. All rights reserved.

Things to Note

Predicting Fraudulent Claims – Project Steps

1. Build predictive models using the structured data
2. Index the text notes for accident claims
3. Build predictive models using the structured data and text mining results
4. Compare model performance



Text mining normally augments data mining by just a small amount, but can be significant

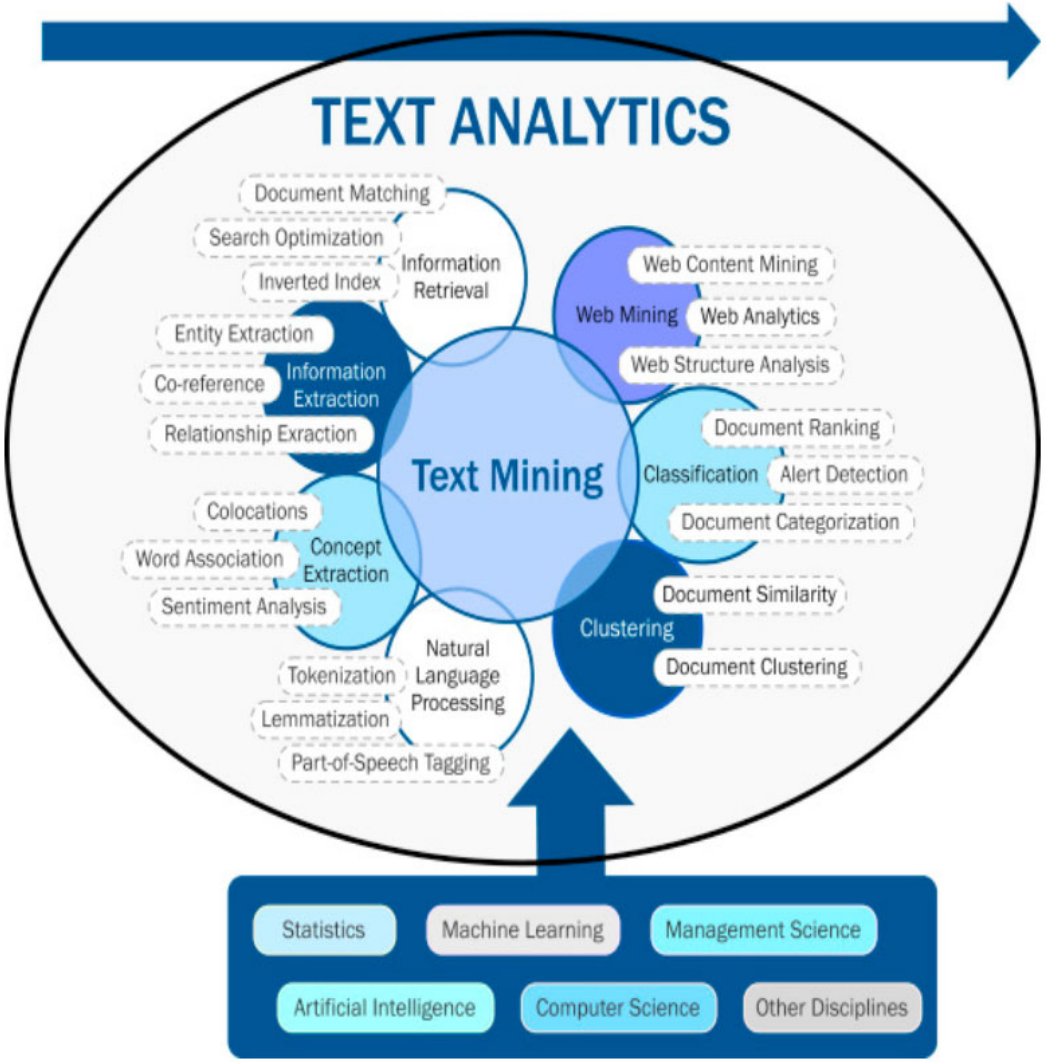
Most of the "lift" provided by data mining on structured data

© 2013-2024 NUS. All rights reserved.

Other language processing tasks

Many more tasks dealing with natural language text

- Information extraction
- Question answering
- Name entity recognition
- Parsing
- Fill in the blank
- Text generation
- Summarization
- Machine translation
- Etc.

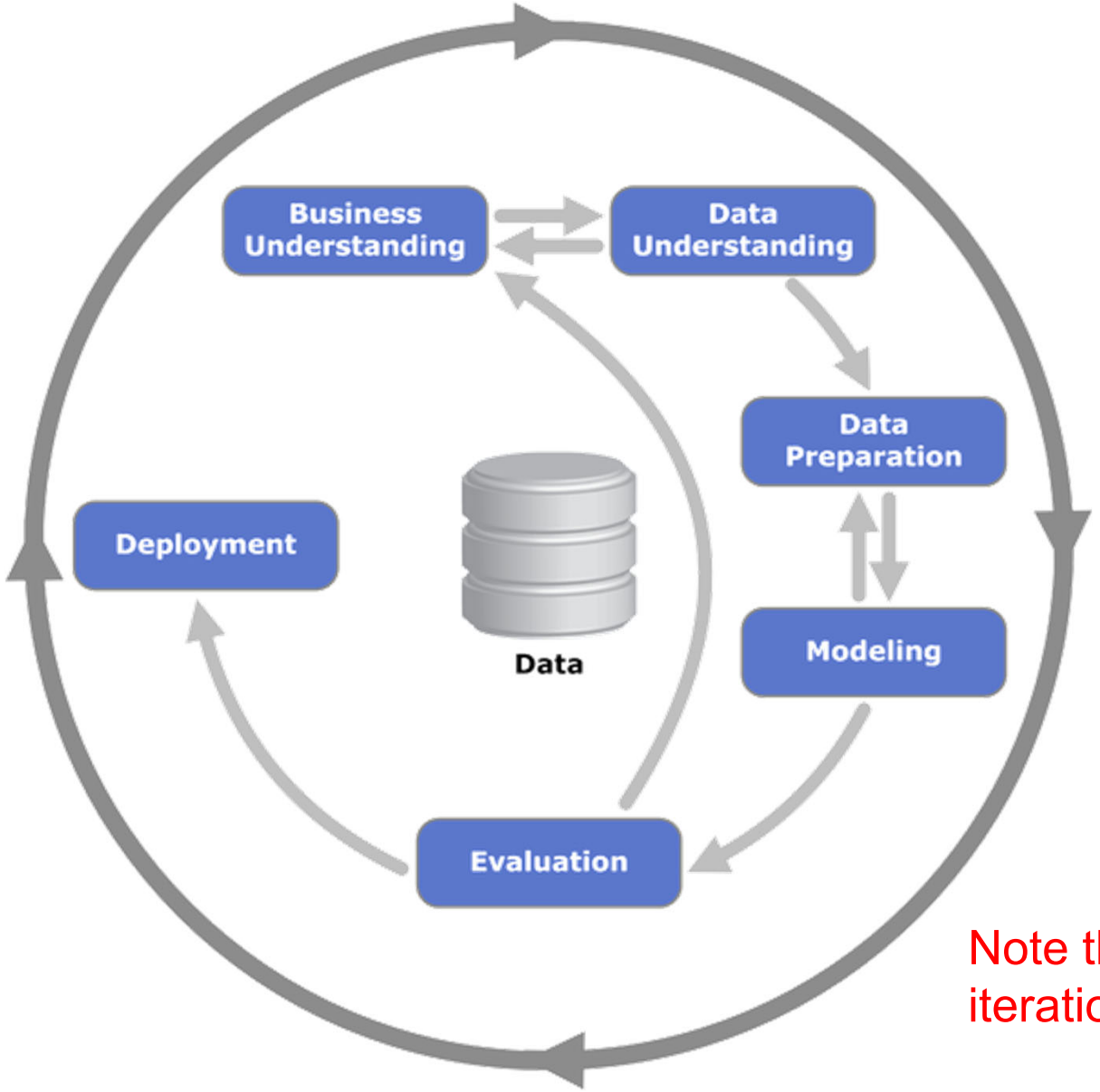


© 2013-2024 NUS. All rights reserved.

The Text Analytics Process

CRISP-DM Process Diagram

Cross-Industry Standard Process for Data Mining



Note the Arrows – iteration & sequence

From: http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

© 2013-2024 NUS. All rights reserved.

6 Phases of CRISP-DM

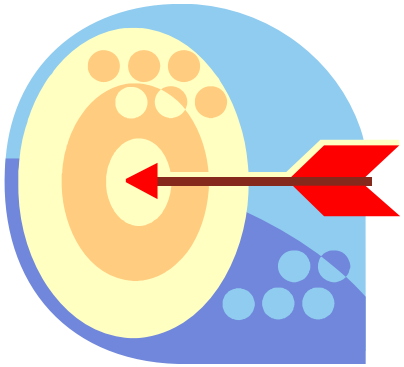
- **Business Understanding**: This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.
- **Data Understanding**: The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.
- **Data Preparation**: The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.
- **Modeling**: In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.
- **Evaluation**: At this stage in the project you have built a model (or models) that appear to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.
- **Deployment**: Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the customer, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front the actions which will need to be carried out in order to actually make use of the created models.

From: http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

© 2013-2024 NUS. All rights reserved.

1. Business Understanding

- Determine business objectives
- Assess situation
- Determine data mining goals
- Produce project plan



Don't find answers to the wrong questions!

- Understand the business case
- Determine the purpose of the study
- Inventory of available text data
- Text data alone or … ?

© 2013-2024 NUS. All rights reserved.

2. Data Understanding

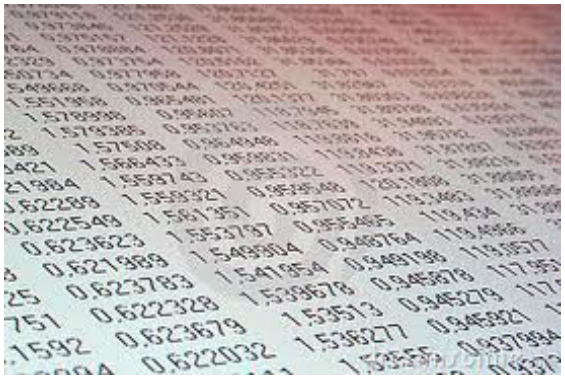
- Collect Initial data
- Describe data
- Explore data
- Verify data quality

- Identify the text data sources (digitized or paper-based; internal or external to the organization)
- Assess the accessibility and usability of the data
- Collect an initial set of data
- Explore the richness of the data (e.g., does it have the information content needed?)
- Assess the quantity and quality of the data (any errors?)

© 2013-2024 NUS. All rights reserved.

3. Data Preparation

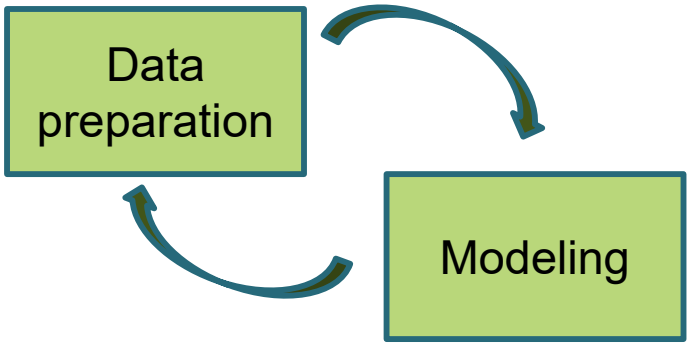
- Select data
- Clean data
- Construct new data
- Integrate data
- Format data



- Establish the text corpus
- Clean the text data
  - formatting, removal of irrelevant sections, combine text, etc.
- Preprocess the data
  - build stopword/include-word list (aka *dictionaries*, and other linguistic resources),
  - create TDM, simplify TDM,
  - label text data for ML models (document with category labels, or words/phrases with specific information – aka *corpus annotation*), etc.

© 2013-2024 NUS. All rights reserved.

4. Modeling



- Select modeling techniques
- Generate test design
- Build model
- Assess model

- Modeling techniques to use depend on the text mining task
- Develop categorization model that can be used to classify/score text
- Other techniques like clustering and association analysis may also be used here
- The output of categorization model may be input to other prediction models (using structured data)

© 2013-2024 NUS. All rights reserved.

5. Evaluation

- Evaluate results
- Review process
- Determine next steps

- Verify and validate the proper execution of all the activities
- Ensure that the models developed and verified are addressing the business problem and satisfying the defined objectives
- Anything left out?

© 2013-2024 NUS. All rights reserved.

6. Deployment

- Plan deployment
- Plan monitoring and maintenance
- Produce final report and presentation
- Review project

- Deployment ranges from writing a report detailing the findings for the decision makers, to integrating the model into BI system
- Models should be updated periodically with new data

© 2013-2024 NUS. All rights reserved.

About Linguistic Resources

© 2013-2024 NUS. All rights reserved.

Dictionaries

- Machine readable resources to improve TA systems on various processing steps
  - Valid term dictionary – terms to be used in TDM
  - Filter dictionary (stopword list) - terms to be excluded
  - Synonym dictionary (substitution dictionary) – group similar words under one standard term (e.g. *U.S.A.* vs *USA*, *dislike* vs *detest*)
  - Terminology dictionary for specific domains (e.g. IT, CRM, biomedical) or tasks (e.g. sentiment mining)

© 2013-2024 NUS. All rights reserved.

Corpora

- Often labelled or annotated, to provide examples for statistical methods and machine learning algorithms to work
- Quality of corpora is critical for the resulting models
  - Validity – correct (need "ground truth" to measure)
  - Reliability – consistent (measured by coefficients of agreement)

© 2013-2024 NUS. All rights reserved.

Corpus Annotation

- Objective tasks - easier, subjective tasks - much harder
- Define task and guidelines
  - The source text? What to annotate? What are the labels for them?
  - Criteria, decision rules and examples
- Train the annotators (humans)
  - Each annotator tries annotating the same set of articles.
  - Check self-agreement rate
    - Individual annotator annotates some randomly selected samples again
    - To eliminate poor performers, or improve their work
  - Check inter-annotator agreement (IAA) rate
    - To resolve conflicts, refine guidelines with example of boundary cases
    - To improve IAA rate before large scale annotation
- In actual annotation, still assign a portion of items to be annotated by two or more annotators for monitoring quality

© 2013-2024 NUS. All rights reserved.

Inter-Annotator Agreement

- Agreement between multiple annotators (assumption: *consistency implies validity*.)
- Used to evaluate and monitor the annotation quality
- Common measures:
  - Cohen's Kappa
  - Fleiss's Kappa
  - Scott's Pi
  - Krippendorff's Alpha
  - Etc.

© 2013-2024 NUS. All rights reserved.

Cohen's Kappa

- The agreement rate for qualitative items taking into account the possibility of chance agreement

$$k = \frac{p_o - p_e}{1 - p_e}$$

  - p_o : observed agreement
  - p_e : expected agreement

$$p_e = \frac{1}{N^2} \sum_k n_{k1}\, n_{k2}$$

  - k : number of categories
  - N : number of observations
  - n_{ki}: number of times annotator i predicted category k

© 2013-2024 NUS. All rights reserved.

Example of Kappa

		B	
		Yes	No
A	Yes	a	b
	No	c	d

		B	
		Yes	No
A	Yes	20	5
	No	10	15

$$p_o = \frac{a+d}{a+b+c+d} = \frac{20+15}{50} = 0.7$$

$$p_{\text{Yes}} = \frac{a+b}{a+b+c+d} \cdot \frac{a+c}{a+b+c+d} = 0.5 \times 0.6 = 0.3$$

$$p_{\text{No}} = \frac{c+d}{a+b+c+d} \cdot \frac{b+d}{a+b+c+d} = 0.5 \times 0.4 = 0.2$$

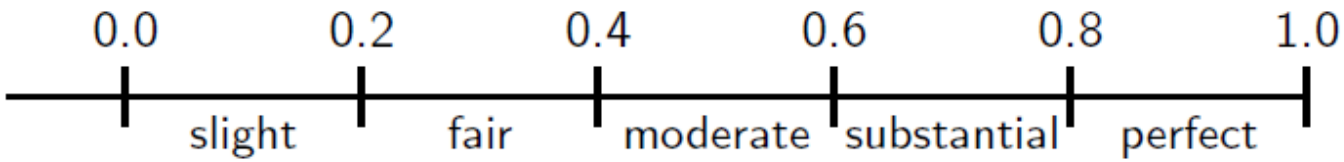
$$p_e = p_{\text{Yes}} + p_{\text{No}} = 0.3 + 0.2 = 0.5$$

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$$

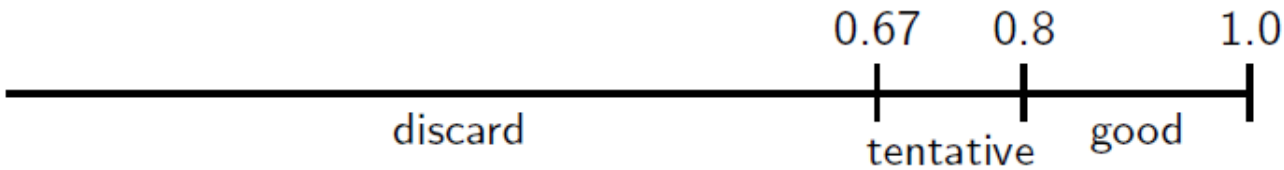
© 2013-2024 NUS. All rights reserved.

Interpretation of Kappa

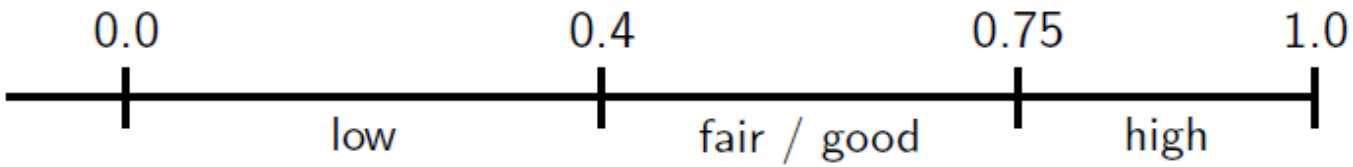
- Landis and Koch, 1977



- Krippendorff, 1980



- Green, 1997



- Artstein and Poesio, 2008: "*If a threshold needs to be set, 0.8 is a good value*."

© 2013-2024 NUS. All rights reserved.

F-Measure

- In some tasks, the number of negative cases can be unknown, undefined or very large
- Then IAA is commonly quantified using the average positive specific agreement - the *average pairwise F-measure*

		B	
		Yes	No
A	Yes	a	b
	No	c	d

$$F = \frac{(1+ß^2) \times \text{recall} \times \text{precision}}{(ß^2 \times \text{precision}) + \text{recall}}$$

$$F = \frac{2[a/(a+b)][a/(a+c)]}{a/(a+b)+a/(a+c)}$$

$$= 2a/(2a+b+c)$$

What will be the value of F for the example on p28?

© 2013-2024 NUS. All rights reserved.

Tools & Solutions for Text Mining

Various Commercial and Open-Source Tools

- You can explore the list of software available here:
  - https://www.kdnuggets.com/software/text.html
  - General tool, as well as tools specialised for industry-specific or task-specific use cases
  - Many as Cloud APIs

SAS® Text Miner

LEXALYTICS

rapidminer
Real Data Science, Fast & Simple

IBM
Watson

Amazon Comprehend
Discover insights and relationships in text

Google Cloud
Cloud Natural Language

IBM SPSS Modeler

orange

© 2013-2024 NUS. All rights reserved.

Factors to decide on the solution

A Business Question is one, where if you knew the answer, you would act differently

- What are your business outcomes?
  - Remember the **Business Question**
- How much skill do you have?
  - In the process, in the tool, in the linguistic ability
- How much time do you have?
  - Skill + time = capacity to improve the linguistic extraction models
  - No time OR no skill = stay with the defaults
- How often are you going to do this?
  - Ad hoc usage? Unique data?
  - Regular runs over similar data?
- What's your budget? Computing resources available?
- Is your data allowed to be sent to cloud?

© 2013-2024 NUS. All rights reserved.

Factors to decide on the tool(s)

- How general/flexible a tool do you want?
  - Does one thing only (e.g., text survey) or very flexible?
  - Domain specific (dictionaries) or open domain?
  - Monolingual, true-multilingual, or translated multilingual?
- How much help do you need?
  - Vendor tools are expensive, but you get training and support
    But…is the support local or overseas?
  - Courses can be tool specific, or tool independent
  - Open source tools are powerful and free, but be prepared to learn on your own (it's a skill you'll employ often ☺)
- Do you need to do it yourself?
  - If you know **what** you need, can you outsource the **how**?

© 2013-2024 NUS. All rights reserved.

What about ChatGPT?

- ChatGPT from OpenAI is a very powerful large language model (LLM) trained to interact with user in a conversational manner.
- It can follow instructions in a prompt and give detailed response.
- Very versatile for language related tasks with proper prompting
- APIs available at a cost
- Tools wrapping around it for specific applications are also flourishing
- Caution:
  - very heavy tool (computationally, memory-wise) and costly
  - Data privacy and security

© 2013-2024 NUS. All rights reserved.

Reference & Resources

- Chris Manning & Hinrich Schutze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999
- Christopher Potts (Stanford University), Sentiment Symposium Tutorial, http://sentiment.christopherpotts.net/index.html
- John Elder, Gary Miner, Bob Nisbet. *Practical Text Mining and Statistical Analysis for non-Structured Text Data Applications*, Academic Press, 2012
- Roger Bilisoly. *Practical Text Mining with PERL*, John Wiley & Sons, 2008
- Hripcsak, G., & Rothschild, A. S. (2005). *Agreement, the f-measure, and reliability in information retrieval*. *Journal of the American medical informatics association*, *12*(3), 296-298.

  https://academic.oup.com/jamia/article-pdf/12/3/296/2429751/12-3-296.pdf
- NLP resources: http://nlp.stanford.edu/links/statnlp.html
- CRISP-DM 1.0 Step-by-step data mining guide (ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf)

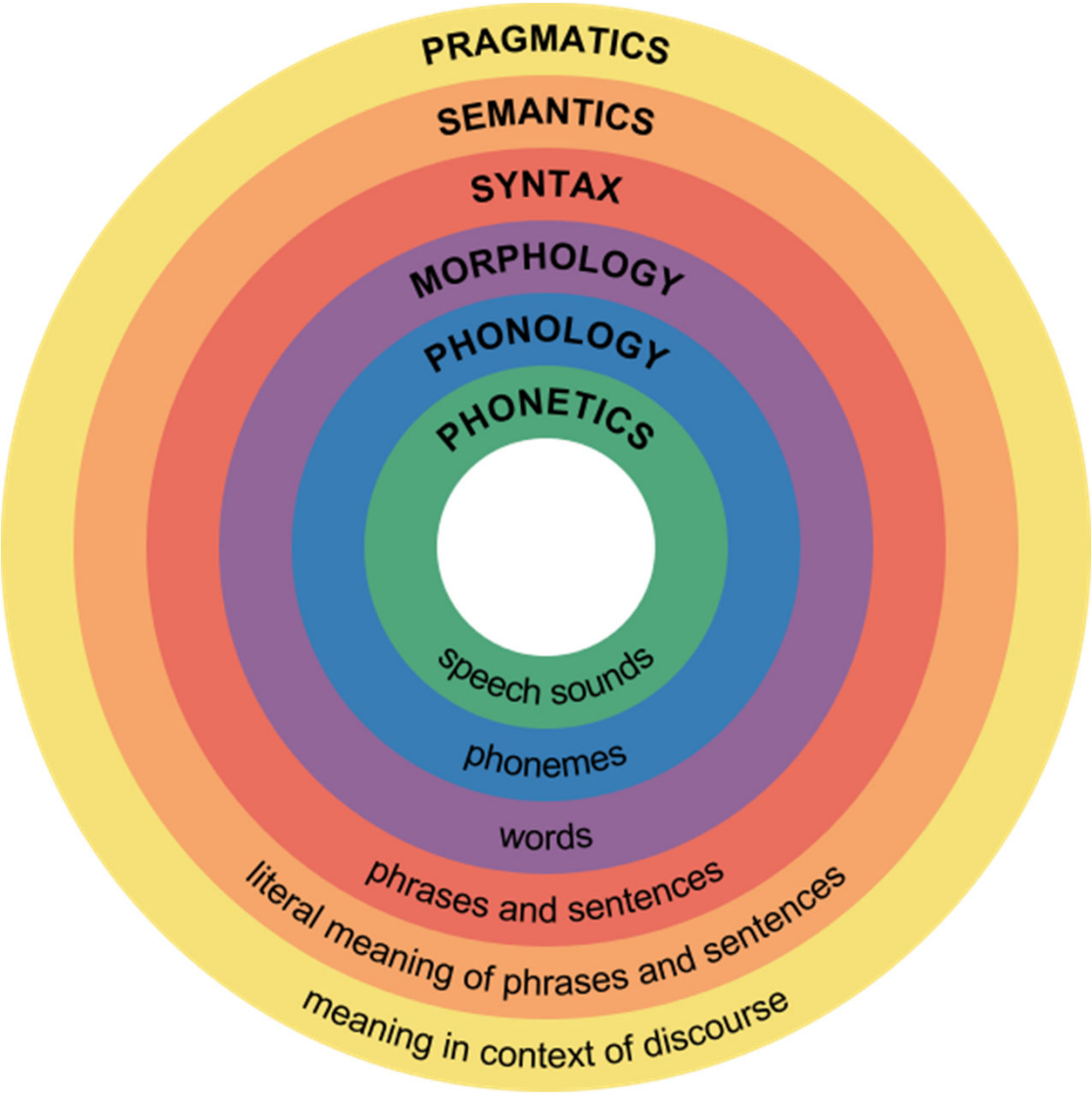
© 2013-2024 NUS. All rights reserved.

Text Analytics

Module 2: Essential linguistics

Dr. Fan Zhenzhen
NUS-ISS
National University of Singapore

Linguistics – the scientific study of language



https://courses.lumenlearning.com/boundless-psychology/chapter/introduction-to-language/

© 2013-2024 NUS. All rights reserved.

1. Morphology

- The structure of words and their part-of-speech (POS, major lexical syntactic categories)
  - Open-class, or content words: nouns, verbs, adjectives, adverbs,
  - Closed-class, or functional words: pronouns, determiners, prepositions, conjunctions, pronouns, numerals, auxiliary verbs, etc.
- Parts of speech group words that have similar neighbouring words (their distributional properties) or take similar affixes (their morphological properties).
- Many words are ambiguous between multiple lexical categories (with >1 POS) E.g. "*book*" can be a noun ("*my* ***book***") or a verb ("*to* ***book*** *a room*")

© 2013-2024 NUS. All rights reserved.

POS Tags

- LDC Penn Tree Bank has 36 POS tags + 12 other tags with detailed information, e.g.

```
1. CC   Coordinating conjunction   25.TO   to
2. CD   Cardinal number            26.UH   Interjection
3. DT   Determiner                 27.VB   Verb, base form
4. EX   Existential there          28.VBD  Verb, past tense
5. FW   Foreign word               29.VBG  Verb, gerund/present participle
6. IN   Preposition/subord.        30.VBN  Verb, past participle
218z      conjunction
7. JJ   Adjective                  31.VBP  Verb, non-3rd ps. sing. present
8. JJR  Adjective, comparative     32.VBZ  Verb, 3rd ps. sing. present
9. JJS  Adjective, superlative     33.WDT  wh-determiner
10.LS   List item marker           34.WP   wh-pronoun
11.MD   Modal                      35.WP   Possessive wh-pronoun
12.NN   Noun, singular or mass     36.WRB  wh-adverb
13.NNS  Noun, plural               37. #   Pound sign
14.NNP  Proper noun, singular      38. $   Dollar sign
15.NNPS Proper noun, plural        39. .   Sentence-final punctuation
16.PDT  Predeterminer              40. ,   Comma
17.POS  Possessive ending          41. :   Colon, semi-colon
18.PRP  Personal pronoun           42. (   Left bracket character
19.PP   Possessive pronoun         43. )   Right bracket character
20.RB   Adverb                     44. "   Straight double quote
21.RBR  Adverb, comparative        45. `   Left open single quote
22.RBS  Adverb, superlative        46. "   Left open double quote
23.RP   Particle                   47. '   Right close single quote
24.SYM  Symbol                     48. "   Right close double quote
        (mathematical or scientific)
```

© 2013-2024 NUS. All rights reserved.

Affixation

- Word stems (lemmas) + affixes (prefixes, suffixes)
  - May involve spelling changes, e.g. *able -> ability*
  - Can be productive, e.g. *unreprogramability*
- Inflectional suffixes – to create variants of the same POS as the stem:
  - *+s, +es* for plural nouns – e.g. *noun -> nouns, class -> classes, story -> stories*
  - *+s, +ed, +ing* for verbs in different tenses and aspects – e.g. *like -> lik**es**, lik**ed**, lik**ing***
  - *+er, +est* for adjectives and adverbs
- Derivational affixes - often change the inherent meaning of the word and/or its POS
  - Suffixes: e.g. *teach* (V) -> *teach**er*** (N), *produce* (V) -> *produc**tion*** (N)
  - Prefixes: e.g. *apply -> **re**apply, happy -> **un**happy*
- There are irregular forms and ambiguities
  - "*corpus*" vs. "*corpora*", "*seek*" vs. "*sought*"
  - Is "*bore*" the present tense of "*bore*" or past tense of "*bear*"?

© 2013-2024 NUS. All rights reserved.

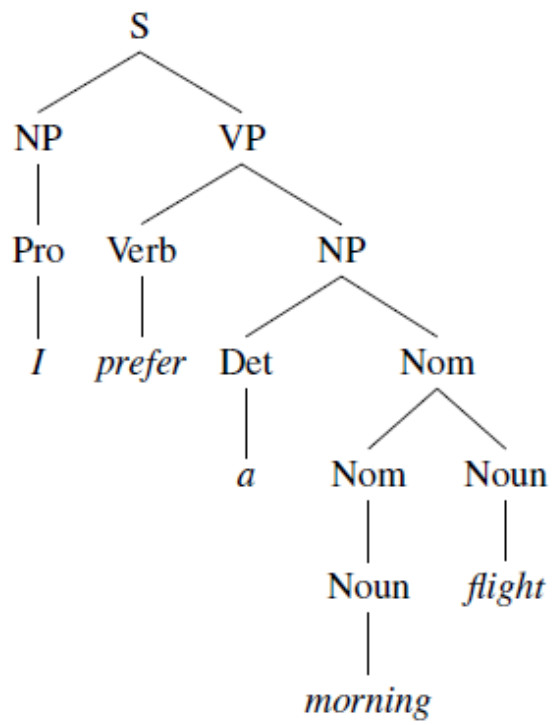
Lexical Features

- Words are described with lexical features based on their syntactic categories and variant forms.
  - Number (num): sg/pl e.g. *word/words*
  - N-type: mass, count, name
  - Person (per): 1, 2, 3 e.g. *I*(1sg), *you*(2sg), *he*(3sg), *they*(3pl)
  - Case: nom, acc e.g. *he, him*
  - Valence: intransitive, transitive, ditransitive, scomp, etc. e.g. *smile, eat, give, believe*, …
  - A-type: base/comparative/superlative e.g. *old, older, oldest*

© 2013-2024 NUS. All rights reserved.

2. Syntax

- Words go together to form syntactic units of various kinds called constituents – words, phrases, clauses
- Parse trees represent the syntactic structure of sentences, showing the constituents.



© 2013-2024 NUS. All rights reserved.

Phrases

- Phrasal Categories (with the corresponding **head** word)
  - **NP** (noun phrases) – e.g. "*all the non-stop morning flights from Denver to Tampa leaving before 10*"
    - Head noun
    - Before head noun: determiners, cardinal/ordinal numbers, quantifiers, adjectives
    - Postmodifiers: prepositional phrases, non-finite clauses, relative clauses
  - **VP** (verb phrases) – e.g. *"book a flight that goes from Denver to Tampa"*
    - Head verb
    - Other constituents: NPs, PPs, Sentential Complements, VP
  - **AP** (adjectival phrases) – head adjective, may be preceded by adverbs. E.g. "*very early*"
  - **PP** (prepositional phrases) – a preposition followed by an NP, e.g. "from Denver"

© 2013-2024 NUS. All rights reserved.

Clauses

- Clausal Categories
  - **Declarative** clauses (e.g. *The taxi arrived early*.)
  - **Interrogative** clauses
    - yes-no questions (e.g. *Is he coming?*)
    - wh-questions (e.g. *When will the taxi arrive?*)
  - **Imperative** clauses (e.g. *Close the door.*)
  - **Relative** clauses (e.g. *Here's the taxi* ***that you called***.)
  - **Complement** clauses (e.g. *I know* ***that the taxi is here***.)
  - **Passive** clause (e.g. *The building was bought by a tycoon*.)

© 2013-2024 NUS. All rights reserved.

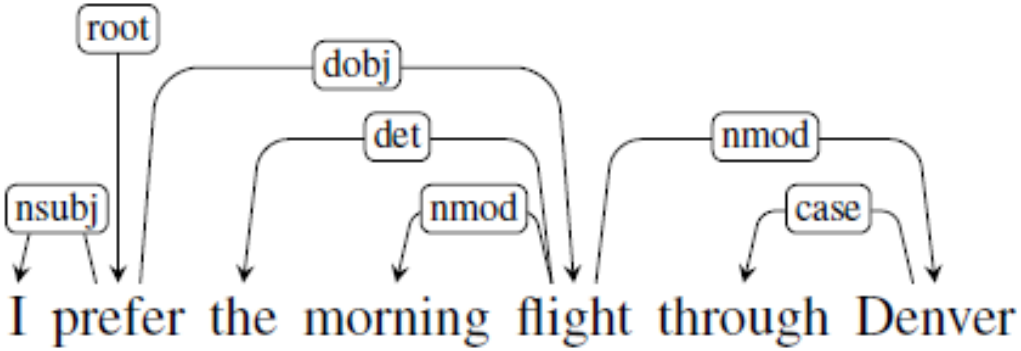
Grammatical Relations of the Constituents

- Subject – ***Alice*** *smiled*.
- Direct object – *Alice ate* ***a burger***.
- Indirect object – *Alice gave* ***me*** *a book*.
- Infinitive complement – *Alice wanted* ***to dance***.
- Specifier/modifier – *Alice is a* ***very clever*** *student*. *She studies* ***diligently***.

© 2013-2024 NUS. All rights reserved.

Dependency Relations

- Typed dependency structures, encoding important information in the sentences
- Illustrated as labelled arcs from heads to dependents



- Approximating the semantic relations between predicates and their arguments

© 2013-2024 NUS. All rights reserved.

Dependency Relations

- Selected dependency relations from the Universal Dependency set

Clausal Argument Relations	Description
NSUBJ	Nominal subject
DOBJ	Direct object
IOBJ	Indirect object
CCOMP	Clausal complement
XCOMP	Open clausal complement
Nominal Modifier Relations	**Description**
NMOD	Nominal modifier
AMOD	Adjectival modifier
NUMMOD	Numeric modifier
APPOS	Appositional modifier
DET	Determiner
CASE	Prepositions, postpositions and other case markers
Other Notable Relations	**Description**
CONJ	Conjunct
CC	Coordinating conjunction

© 2013-2024 NUS. All rights reserved.

3. Semantics

- Linguistic expressions -> meaning representation (knowledge representation, e.g. FOL, frames)
  - Propositions (predicates, referring expressions)
  - Correctness (true/false), contradiction
  - E.g. *I ate a turkey sandwich for lunch at my desk*.

$$\exists e\ Eating(e) \ \wedge\ Eater(e, Speaker) \wedge Eaten(e, TurkeySandwich) \\ \wedge\ Meal(e, Lunch) \wedge Location(e, Desk)$$

- Ambiguity – some sentences can convey more than one proposition.
- Entailment – The assertion of some propositions implies the truth of other propositions. =>Inference!

© 2013-2024 NUS. All rights reserved.

4. Pragmatics

- The use of language in context (both linguistic and situational)
- Utterances and speech acts (to achieve some effect on hearer)
  - Locution
    - Physical utterance with context and reference, i.e., who is the speaker and the hearer, which is the object, etc.
  - Illocution
    - The act of conveying intentions, i.e., the speaker wants the hearer to do something or to think something as a consequence of its utterance
  - Perlocutions
    - Actions that occur as a result of the illocution
  - Example: **Open the window!**
    - Locution: Monique is the speaker, Steve is the hearer, the window is the last left one
    - Illocution: Monique wants Steve to open the window
    - Perlocution: Steve opens the window
- Discourse, coherence

© 2013-2024 NUS. All rights reserved.

Reference & Resources

- *Introduction to Linguistics for Natural Language Processing*, by Ted Brisco (https://www.cl.cam.ac.uk/teaching/1314/L100/introling.pdf)

- Marcus, Mitchell, Beatrice Santorini, and Mary Ann Marcinkiewicz. "Building a large annotated corpus of English: The Penn Treebank." (1993). (https://catalog.ldc.upenn.edu/docs/LDC95T7/cl93.html)

© 2013-2024 NUS. All rights reserved.

Text Analytics

Module 3: Get the Text Data Ready for Analysis

Dr. Fan Zhenzhen
NUS-ISS
National University of Singapore
Email: zhenzhen@nus.edu.sg

© 2015-2024 NUS. The contents contained in this document may not be reproduced in any form or by any means, without the written permission of ISS, NUS, other than for the purpose for which it has been supplied.

Outline

Overview of task

Vectorization - Term-Document Matrix

Vectorization - Embeddings

© 2015-2023 NUS. All rights reserved.

What is needed…

Documents(Corpus)

Lost glamor

High tea at Raffles!

Not what it was, but still a place to go to.

Amazing service
Rated 5 by travel-gini on Feb 26, 2013

Great location with a little bit of history, the staff make this hotel though
Have a drink in the Long Bar, throw your nutshells on the floor, then go to the Tiffin Room for the best curry in the world. About £40a head for food but the choice is brilliant and when my wife mentioned it was her birthday at the end of the meal a cake was presented, what amazing service.

Stayed February 2013

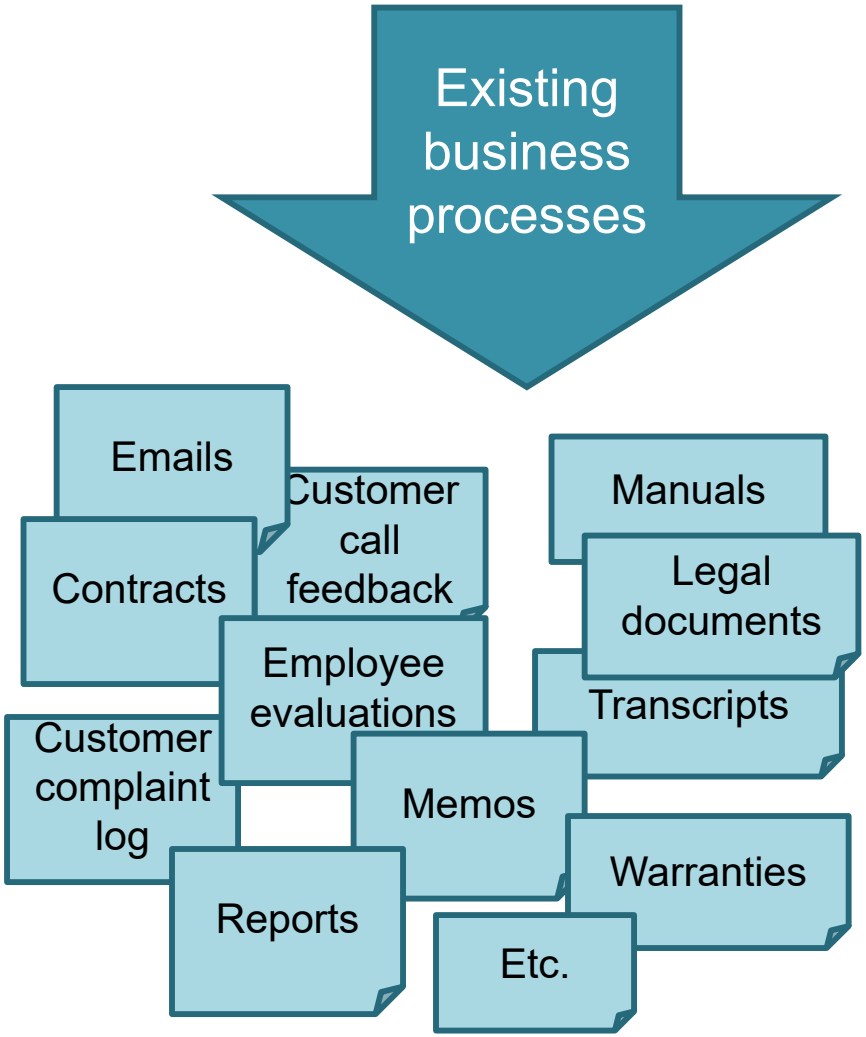
Numeric Vectors

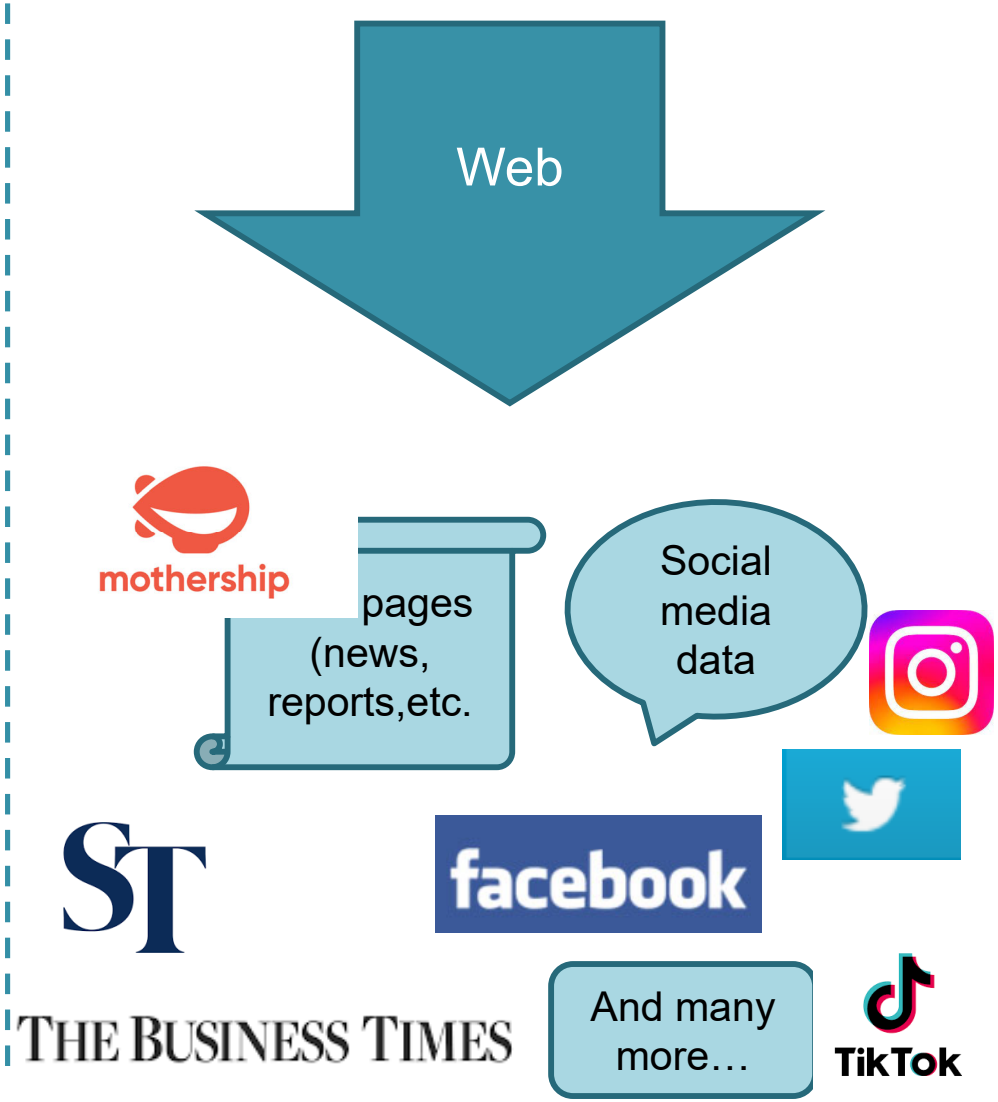
	v1	v2	V3	v4	v5	v6	v7	v8	v9	…
Doc1	1	1	0	0	0	1	0	0	0	
Doc2	0	0	1	1	1	0	0	1	0	
Doc3	0	0	0	1	0	0	1	0	0	
Doc4	0	0	0	0	2	0	0	1	1	
…										

Suitable for tasks considering a document as an object, e.g., document classification, document clustering

© 2015-2023 NUS. All rights reserved.

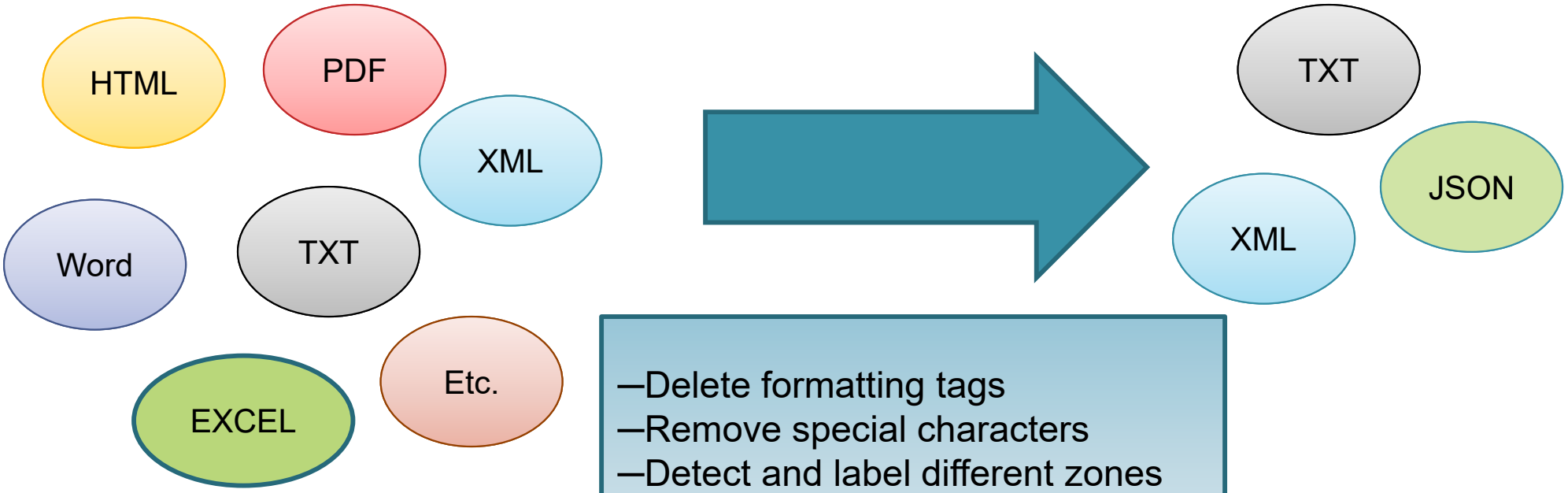
Sources of Text Data





© 2015-2023 NUS. All rights reserved.

File Preprocessing



—Delete formatting tags
—Remove special characters
—Detect and label different zones
—Determine sentence/paragraph boundaries

Most TA tools provide functionality of importing text from some common formats.

© 2015-2023 NUS. All rights reserved.

In a pure text file…

Amazing service

Rated 5 by travel-gini on Feb 26, 2021

Great location with a little bit of history, the staff make this hotel though
Have a drink in the Long Bar, throw your nutshells on the floor, then go to the Tiffin Room for the best curry in the world. About £40a head for food but the choice is brilliant and when my wife mentioned it was her birthday at the end of the meal a cake was presented, what amazing service.

Stayed February 2021

…

© 2015-2023 NUS. All rights reserved.

XML as Standard Exchange Format

- Popular in industry and text-processing community

- With XML, we can insert tags onto a text to identify its parts.
  - Eg. <*DOC*>, <*SUBJECT*>, <*TOPIC*>, <*TEXT*>, etc.
  - Such tags are very useful as they allow selection/extraction of the parts to generate features for subsequent mining.

- Many word processors allow documents to be saved as XML format

© 2015-2023 NUS. All rights reserved.

What would an XML doc look like?

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<REVIEWS>
    <REVIEW>
        <TITLE>Amazing service</TITLE>
        <RATING>5</RATING>
        <DATE>26/02/2021</DATE>
        <BY>travel-gini</BY>
        <CONTENT>Great location with a little bit of history, the staff make this hotel
    though
    Have a drink in the Long Bar, throw your nutshells on the floor, then go to the Tiffin Room
    for the best curry in the world. About £40a head for food but the choice is brilliant and
    when my wife mentioned it was her birthday at the end of the meal a cake was presented,
    what amazing service. Stayed February 2021 </CONTENT>
    </REVIEW>
    ...
</REVIEWS>
```

© 2015-2023 NUS. All rights reserved.

JSON Format

- JavaScript Object Notation (JSON) is a standard text-based format for representing structured data based on JavaScript object syntax
- Uses human-readable text to store and transmit data objects consisting of attribute-value pairs and arrays
- Common data format with diverse uses in electronic data interchange, including that of web applications with servers

```
{
 "FirstName"              : "Sam",
 "LastName"               : “Jackson”,
 "employeeID"             : 5698523,
 "Designation"            : "Manager",
 “LanguageExpertise”      : [“Java”, “C#”, “Python”]
 “Car”                    : “ “
}
```

© 2015-2023 NUS. All rights reserved.

Our example in JSON format

```
{
  "title": "Amazing service",
  "rating": 5,
  "date": "26/02/2021",
  "by": "travel-gini",
  "content": "Great location with a little bit of history, the staff make this hotel though Have a drink in the Long Bar, throw your nutshells on the floor, then go to the Tiffin Room for the best curry in the world. About £40a head for food but the choice is brilliant and when my wife mentioned it was her birthday at the end of the meal a cake was presented, what amazing service. Stayed February 2021"
}
```

© 2015-2023 NUS. All rights reserved.

Vectorization – Term Document Matrix

TDM – Pre-processing

- Term Document Matrix (TDM) uses terms (keywords) as features for the vectors to represent the documents.

	amazing	service	lost	glamour	disappoint	brilliant	super	expensive	noisy	...
Doc1	1	1	0	0	0	1	0	0	0	
Doc2	0	0	1	1	1	0	0	1	0	
Doc3	0	0	0	1	0	0	1	0	0	
Doc4	0	0	0	0	2	0	0	1	1	
...										

- Common pre-processing steps (for English, may be different for other languages)
  - Tokenization
  - Case lowering
  - Stemming/lemmatization
  - Stopword removal
  - Punctuation removal
  - Numeric removal

© 2015-2023 NUS. All rights reserved.

Tokenization

- To break a stream of characters into tokens

Great location with a little bit of history.

Great | location | with | a | little | bit | of | history | .

- This is known as unigram model – every token is a single word.
- There are also bigram, trigram models, where each token is composed of two/three words.

Great location | location with | with a | a little | little bit

bit of | of history | history .

© 2015-2023 NUS. All rights reserved.

Tokenization Challenges

- Tokenization is done by identifying <u>token delimiters</u>
  - Whitespace characters such as ***space, tab, newline***
  - Punctuation characters like ***( ) < > ! ? “ ”***
  - Other characters ***. , : - ‘ ’*** etc.
- It seems simple, but…
  - ***. , :*** between numbers are part of the number

    12.34 | 12,345 | 12:34
  - **.** can be part of an abbreviation or end of a sentence

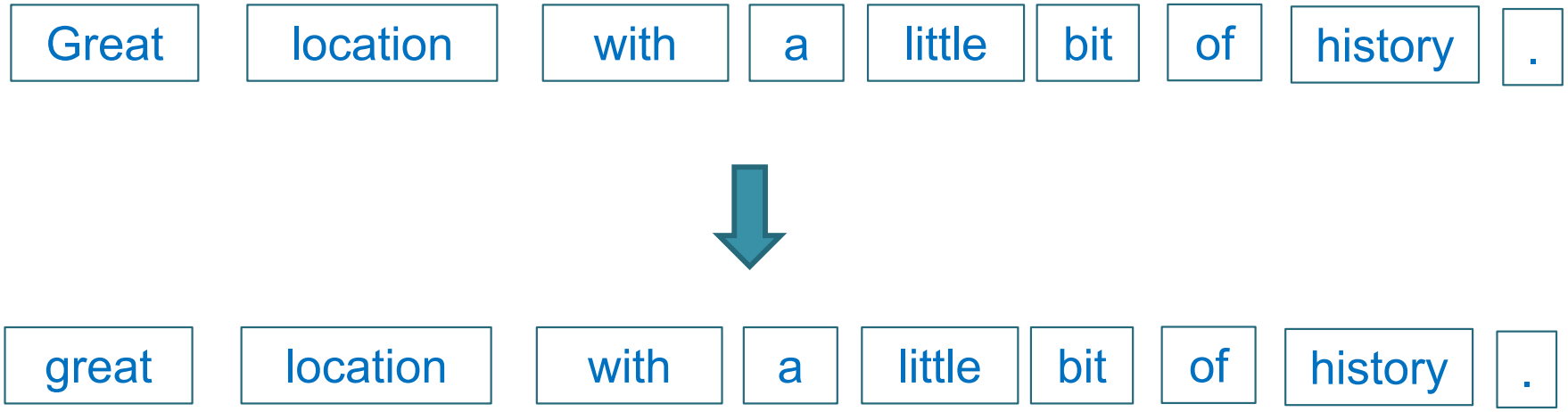
    U.S.A. | Dr.
  - ***’*** can be a closing internal quote, indicate a possessive, or be part of another token

    My friend’s | isn’t

© 2015-2023 NUS. All rights reserved.

Case Lowering

- Also known as **case normalization**, to convert all tokens to lower case to remove the variation of words due to case differences.

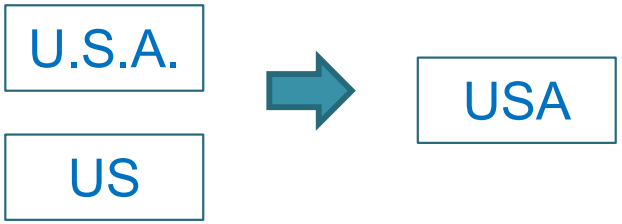


- May cause issue in some contexts
  - E.g., "Apple" -> "apple", "US" to "us"

© 2015-2023 NUS. All rights reserved.

Stemming/Lemmatization

- A word may come in varied forms and therefore need to be converted into a standard form
- Stemming can reduce the number of distinct features in a text corpus and increase the frequency of occurrence of some individual features.
  - **Inflectional** stemming (no change of POS) -- **Lemmatization**
  - **Derivational** Stemming (with change of POS)
  - Other normalisation (including case normalisation)



Refer to p5 in *Essential Linguistics*

© 2015-2023 NUS. All rights reserved.

Stemmers

- Some well-known stemming algorithms for English

Lovins Stemmer
by Julie Beth Lovins, 1968

- single pass, longest-match
- removing the longest suffix, ensuring the remaining stem is at least 3 characters long
- reforming the stem through recoding transformations

Porter Stemmer
by Martin Porter, 1980

- Widely used
- with implementations in various languages available online (C, java, Perl, python, C#, VB, Javascript, Tcl, Ruby, etc.)

Snowball Stemmer
by Porter

- a framework for writing Stemming algorithms
- Newer version of Porter Stemmer

© 2015-2023 NUS. All rights reserved.

How much stemming should be done?

- An inflectional stemmer needs to be partly rule-based and partly dictionary-based.
- Derivational stemming is more aggressive and therefore can reduce the number of features in a corpus drastically. However, meaning might be lost in the stemming process.

Too aggressive stemming can result in loss of meaning and non-legitimate words without the support of a dictionary.

```
[[6]]
battery life portability accessories style

[[7]]
   ability  store    music     ability  create playlists

[[8]]
portability capacity sound quality durability
```

```
[[6]]
batteri life portabl accessori style

[[7]]
   abil  store    music     abil  creat playlist

[[8]]
portabl capac sound qualiti durabl
```

© 2015-2023 NUS. All rights reserved.

Stopword Removal

- Some words are extremely common. They appear in almost all documents and carry little meaning. They are of limited use in text analytics applications.
  - Functional words (conjunctions, prepositions, determiners, or pronouns) like *the, of, to, and, it,* etc.
  - A stopword list can be constructed to exclude them from analysis.
  - **Depending on the domain**, other words may need to be included in the stopword list.

great | location | with | a | little | bit | of | history | .

© 2015-2023 NUS. All rights reserved.

Stopword List

- Also known as *filter dictionary* / *exclusion dictionary*
- To support the stopword removal step in preprocessing
- A list of very common words
  - usually functional words like *preposition*, *conjunction*, etc.
  - or words that are unimportant for the mining task
- Example stopword list (not complete):

a	*an*	*because*	*before*
about	*and*	*been*	*being*
above	*any*	*before*	*below*
after	*are*	*being*	*between*
again	*aren't*	*below*	*both*
against	*as*	*between*	*but*
all	*at*	*both*	*by*
am	*be*	*been*	*...*

From http://www.ranks.nl/resources/stopwords.html

© 2015-2023 NUS. All rights reserved.

Further clean-up

- Punctuation removal
- Number removal, etc.

great | location | history | .

⬇

great | location | history

© 2015-2023 NUS. All rights reserved.

TDM - Indexing

- Create the vector representation of documents (*term-document matrix* or *document-term matrix*) using "bag-of-words" approach

$$
\begin{pmatrix}
 & T_1 & T_2 & \dots & T_t \\
D_1 & w_{11} & w_{21} & \dots & w_{t1} \\
D_2 & w_{12} & w_{22} & \dots & w_{t2} \\
\vdots & \vdots & \vdots & & \vdots \\
\vdots & \vdots & \vdots & & \vdots \\
D_n & w_{1n} & w_{2n} & \dots & w_{tn}
\end{pmatrix}
$$

T: term
D: document
w: **weight** of the term

- Vector features: terms/keywords, usually only content words (adjectives, adverbs, nouns, and verbs).

© 2015-2023 NUS. All rights reserved.

Term Weighting

- Binary
  - 0 or 1, simply indicating whether a word has occurred in the document (but that's not very helpful).
- Frequency-based
  - *term frequency*, the frequency of words in the document, which provides additional information that can be used to contrast with other documents.

	amazing	service	lost	glamour	disappoint	brilliant	super	expensive	noisy	…
Doc1	1	1	0	0	0	1	0	0	0	
Doc2	0	0	1	1	1	0	0	1	0	
Doc3	0	0	0	1	0	0	1	0	0	
Doc4	0	0	0	0	2	0	0	1	1	
…										

© 2015-2023 NUS. All rights reserved.

Frequent Word List

- With <u>frequency-based</u> TDM, a list of words and their frequencies in the corpus can be generated
  - *Global frequency* – how many times a word appears in the corpus
  - *Document frequency* – how many unique documents contain the word

- This list, sorted by frequency, can give us a rough idea of what the corpus is about.

- Word Cloud is a nice visualization of such information.

research
work
project
students
graduate
undergraduate
class
astronomy
one
student
department
ability
skills
data
school
level
working
extremely
several
first
telescope
program
time
also
physics
galaxies
undergraduates
year
far
tremendous
much
problem
began
already
radio
However
since
courses
support
think
make
took
initiative
best
highly
images
exceptional
interest
now
remarkable
actually
scientist
well
many
great
gone
researcher
high
course
excellent
application
scientific
second
always
leadership
truly
within
summer
new
strong
faculty
met
known
take
career
field
without
possible
carried
intellectual
interesting
learned
large
beyond
combination
dark
letter
abilities
might
given
background
astronomical
maturity
bright
cluster
enough
seems
creative
senior
involved
successful
top
postdocs
necessary
matter
thought
highest
around
needed
clear
performance
three
go
years
amount
abviously
real
know
motivated
became
clarity
point
together
Like
past
independent
worked
intelligence
questions
pleased
taking
HST
stellar
although
drive
required
problems
two
creative
seems
enough
cluster
actually
maturity
bright
scientist
well
background
astronomical
yet
many
great
gone
researcher
might
leadership
abilities
combination
letter
dark
carried
possible
without
beyond
learned
interesting
intellectual
large
projects
abviously
amount
go
years
three
around
needed
clear
performance
thought
highest
postdocs
bit
top
successful
involved
senior
necessary
matter
asked
code
Spring
given
background
astronomical
maturity
bright
cluster
enough
seems
creative
actually
remarkable
scientist
well

Word Cloud: another example

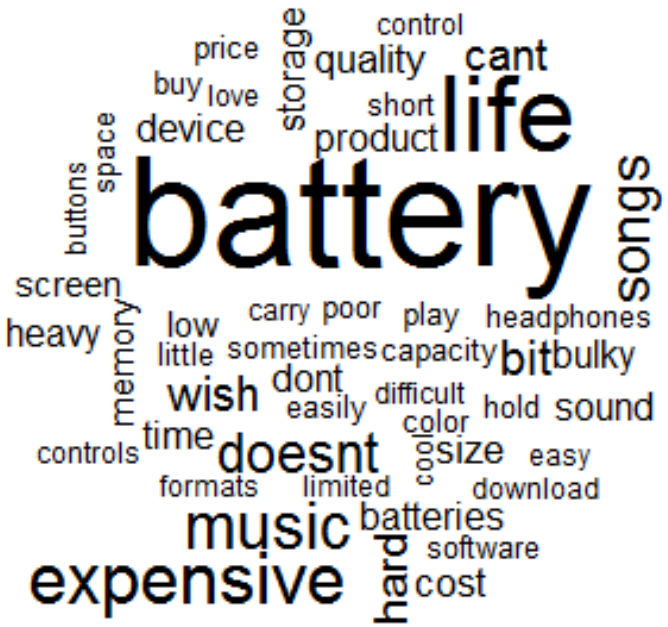
- Generated from http://worditout.com/word-cloud/make-a-new-one

© 2015-2023 NUS. All rights reserved.

Contrasting Two Groups

- "What do you like most…"

- "What do you like least…"



© 2015-2023 NUS. All rights reserved.

Other Weighting Methods

- Normalized frequency
  - To deal with varied document length, since a long document definitely has more occurrences of terms than a short document

$$normalized_frequency = \frac{frequency\ of\ a\ term\ in\ a\ document}{total\ number\ of\ terms\ in\ the\ document}$$

- ***tf-idf***
  - To modify the frequency of a word in a document by the perceived importance of the word(the *inverse document frequency)*, widely used in information retrieval
    - When a word appears in many documents, it's considered unimportant.
    - When the word is relatively unique and appears in few documents, it's important.

© 2015-2023 NUS. All rights reserved.

tf-idf Indexing

- *tf-idf* weighting :

$$tf\text{-}idf_{t,d} = tf_{t,d} * idf_t$$

  - $tf_{t,d}$: <u>*term frequency*</u> – number of occurrences of term t in document d
  - idf_t : <u>*inverted document frequency*</u> of term t

$$idf_t = \log \frac{N}{df_t}$$

N : the total number of documents in the corpus

df_t : the document frequency of term t, i.e., the number of documents that contain the term.

© 2015-2023 NUS. All rights reserved.

tf-idf Indexing – An Example

Note that in this example, stopwords and very common words are not removed, and terms are not reduced to root terms.

TERM VECTOR MODEL BASED ON $w_i = tf_i * IDF_i$

Query, Q: "gold silver truck"
D_1: "Shipment of gold damaged in a fire"
D_2: "Delivery of silver arrived in a silver truck"
D_3: "Shipment of gold arrived in a truck"
$D = 3$; $IDF = \log(D/df_i)$

		Counts, tf_i						Weights, $w_i = tf_i * IDF_i$			
Terms	**Q**	**D_1**	**D_2**	**D_3**	**df_i**	**D/df_i**	**IDF_i**	**Q**	**D_1**	**D_2**	**D_3**
a	0	1	1	1	3	3/3 = 1	0	0	0	0	0
arrived	0	0	1	1	2	3/2 = 1.5	0.1761	0	0	0.1761	0.1761
damaged	0	1	0	0	1	3/1 = 3	0.4771	0	0.4771	0	0
delivery	0	0	1	0	1	3/1 = 3	0.4771	0	0	0.4771	0
fire	0	1	0	0	1	3/1 = 3	0.4771	0	0.4771	0	0
gold	1	1	0	1	2	3/2 = 1.5	0.1761	0.1761	0.1761	0	0.1761
in	0	1	1	1	3	3/3 = 1	0	0	0	0	0
of	0	1	1	1	3	3/3 = 1	0	0	0	0	0
silver	1	0	2	0	1	3/1 = 3	0.4771	0.4771	0	0.9542	0
shipment	0	1	0	1	2	3/2 = 1.5	0.1761	0	0.1761	0	0.1761
truck	1	0	1	1	2	3/2 = 1.5	0.1761	0.1761	0	0.1761	0.1761

http://www.miislita.com/term-vector/term-vector-3.html

© 2015-2023 NUS. All rights reserved.

Alternative Representation of TDM

- The term document matrix is sparse, expected to have most of the values to be zero, since typically a document will only contain a small subset of the vocabulary in a corpus

```
<<DocumentTermMatrix (documents: 1000, terms: 17887)>>
Non-/sparse entries: 92858/17794142
Sparsity           : 99%
Maximal term length: 56
Weighting          : term frequency (tf)
```

- It saves memory to store the matrix as a set of sparse vectors, where a row is represented by a list of pairs, (ColumnIndex, Value)

Matrix

0	5	2	0
4	0	0	0
3	1	0	6

Sparse Vectors

(2, 5) (3, 2)
(1, 4)
(1, 3) (2, 1) (4, 6)

© 2015-2023 NUS. All rights reserved.

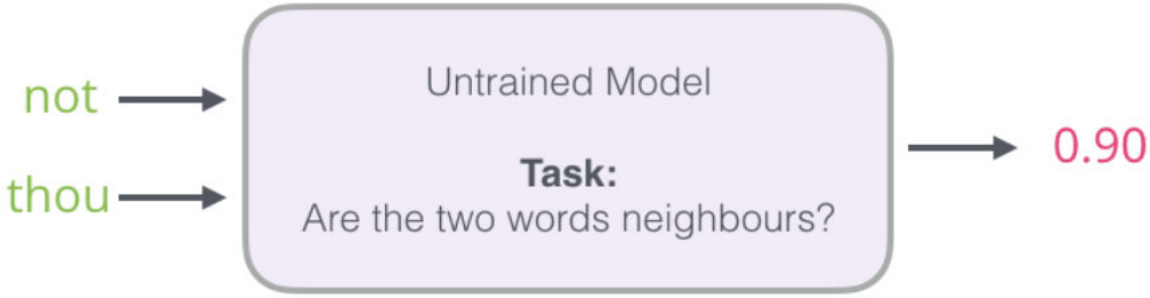
Vectorization – Embeddings

Other Vectorization Methods

- Word Embeddings:
  - A dense representation of words as vectors of real numbers in a continuous vector space with a much lower dimension
  - Learned by deep neural networks during a prediction task

    e.g. GloVe, Word2Vec, FastText, etc.

"Thou shall not make a machine in the likeliness of human mind…"



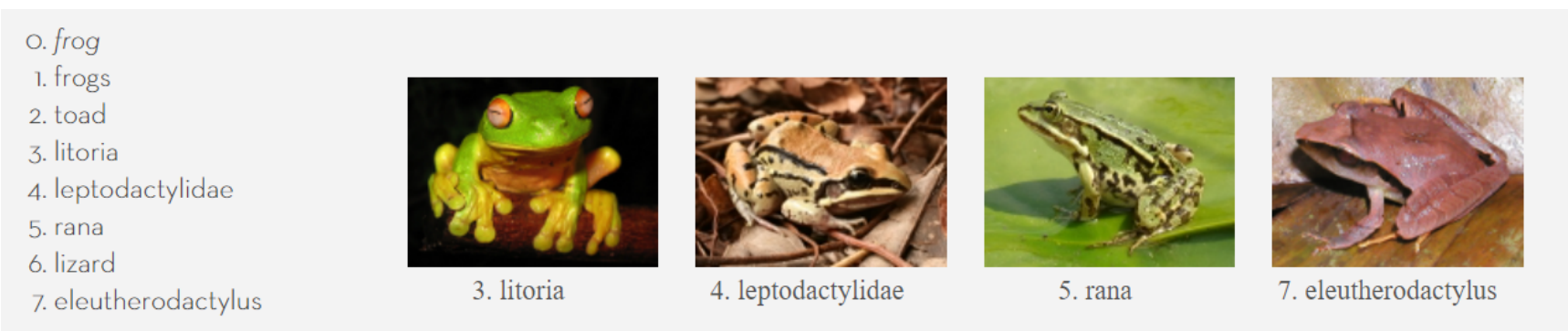
- Look up embeddings
- Calculate prediction

© 2015-2023 NUS. All rights reserved.

Word Embeddings Lookup

- Using pre-trained word vectors, e.g., GloVe (Global Vectors for Word Representation) by Stanford
- Lookup each token (word) in the vectors.

- Pros: the embeddings capture semantic or linguistic information of the words

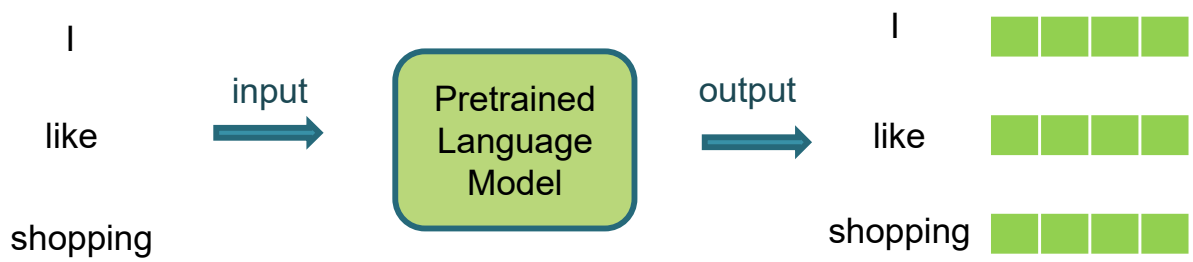


- Cons: one word one vector (what about words with multiple meanings?)

© 2015-2023 NUS. All rights reserved.

Contextualized Word Embeddings

- Using pre-trained large language models, e.g., BERT, GPT2, etc.
- Given text input, the model generates embedding for each word based on the context, better than lookup embeddings



- Can handle different meanings of the same word with different context

  e.g., "Move the table closer to the window"

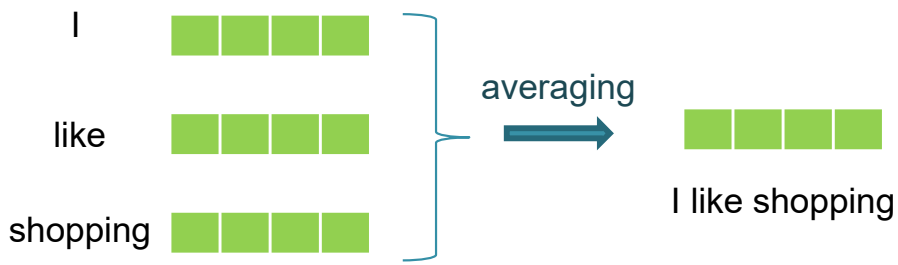
  vs.

  "Remove the first column in the table"

© 2015-2023 NUS. All rights reserved.

Document Embeddings

- Given each word in the document(any sequence of text) represented as an embedding vector, how to get the vector representation of the whole document?
- Common approach: summarizing all the word vectors by averaging (or summing) them

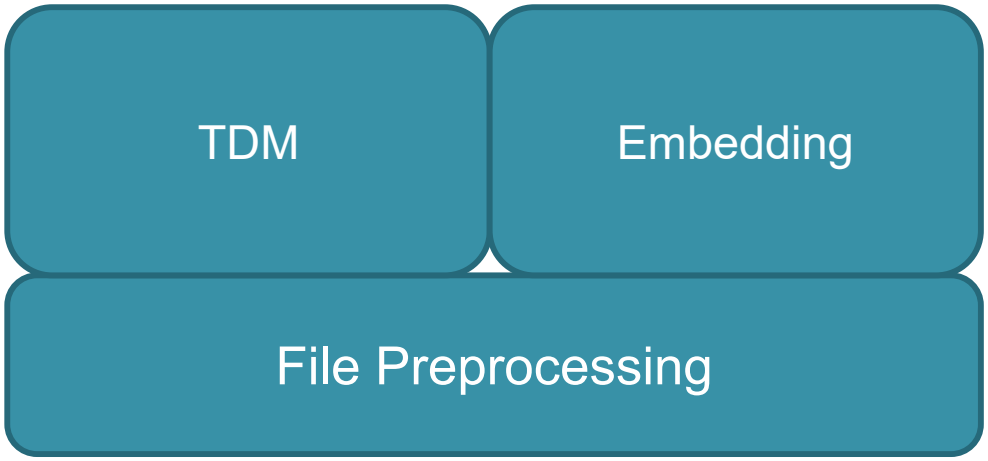


- However, there are also approaches to map documents to vectors, e.g., using DL models to learn to represent documents(*doc2vec*, *skip-thought, FastSent, Sentence-BERT, etc.*)

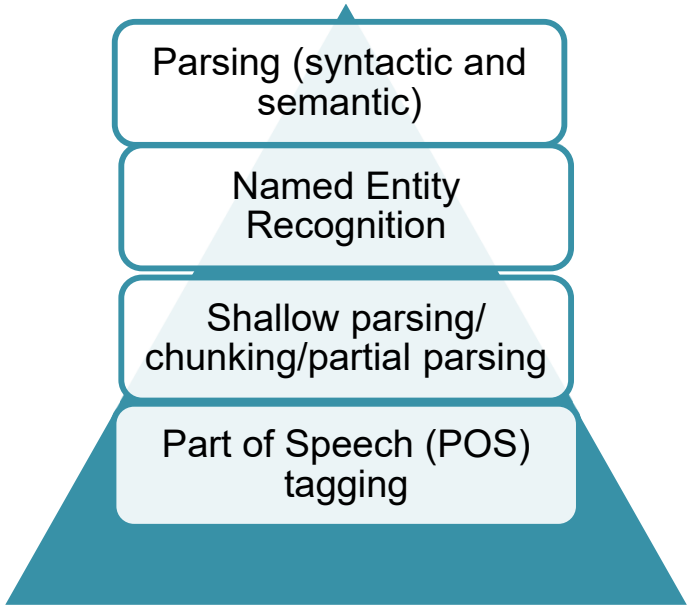
© 2015-2023 NUS. All rights reserved.

Summary

Vectorization
(for document-level tasks)



Later…



© 2015-2023 NUS. All rights reserved.

Reference & Resources

- Jurafsky, Dan. *Speech & language processing*. Pearson Education India, 2000. (continuously updated)
- Weiss, Indurkhya, & Zhang. Chapter 2 "From Textual Information to Numerical Vectors", *Fundamentals of Predictive Text Mining*, Springer, 2010.

- List of online word cloud generators
  - http://www.techlearning.com/default.aspx?tabid=67&entryid=364

© 2015-2023 NUS. All rights reserved.

Text Analytics

Module 5: Information Extraction

Dr. Fan Zhenzhen
Dr. Leong Mun Kew
NUS-ISS
National University of Singapore

Outline for this Module

- What is Information Extraction (IE)?
- How does an IE System work?
  - Main components of IE systems
  - Rule based methods for IE
  - Machine-learning based methods for IE
- Practical Information Extraction

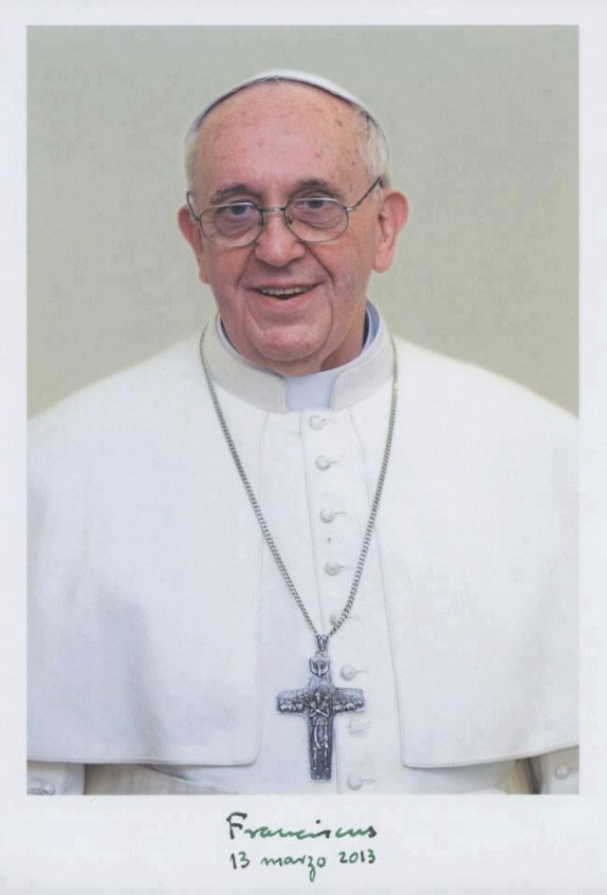
© 2013-2024 NUS. All rights reserved.

What is Information Extraction

From words to concepts

"Cardinal Jorge Mario Bergoglio of Buenos Aires was elected the 266th pope of the Catholic Church on 13 May 2013, taking the name Pope Francis"

- Cardinal: a concept = a prince of the Catholic Church
- Jorge Mario Bergoglio: a concept = Cardinal of Buenos Aires
- Buenos Aires: a concept = city in Argentina
- Pope: a concept = head of Catholic Church
- Catholic Church: a concept = religious organization
- 13 May 2013: a concept = a date
- Pope Francis: a concept = current Pope



© 2013-2024 NUS. All rights reserved.

Concept vs. Named Entity vs. Information

- Named Entity = lowest level of recognition by an IE system
  - Normally recognized by dictionaries or rules
- Concept = rule or heuristic to create an abstraction
  - Sometimes called a "natural class" = different people at different times and in different places would refer to the same referent with that concept
  - "president of the United States" vs. "president of the United Kingdom"
- Information = words, named entities, concepts which fulfill a need
  - So if you have a question, and a phrase answers that question, then that phrase is an example of information
  - Information is often regular, i.e., with a pattern
    - Eg, information about a person = name, age, sex, address, hp#, …
    - Information about a company = name, address, stock symbol, Chairman, …

© 2013-2024 NUS. All rights reserved.

Exercise

The National University of Singapore and The Hebrew University of Jerusalem (HUJ) are launching a Joint Doctor of Philosophy degree programme in biomedical science from August 2013. Professor Tan Eng Chye, NUS Deputy President and Provost, and Professor Menahem Ben-Sasson, President of HUJ signed the joint degree agreement at NUS, in the presence of Ambassador of Israel to Singapore Her Excellency Amira Arnon and about 30 invited guests.

#	Concept	Category	#	Concept	Category
1	National University of Singapore	Place	7		
2	Tan Eng Chye	Name	8		
3	Menahem Ben-Sasson		9		
4			10		
5			11		
6			12		

© 2013-2024 NUS. All rights reserved.

Information extraction vs information retrieval

How tall is the Eiffel Tower?

- A traditional search engine (IR) returns a list of documents which (hopefully) includes the answer
- The user needs to open/read the document to get the answer

WEB IMAGES NEWS MORE

bing Beta

how tall is the eiffel tower

4,850,000 RESULTS Narrow by language Narrow by region

How Tall is the Eiffel Tower? - Buzzle
www.buzzle.com/articles/**how-tall-is-the-eiffel-tower**.html
How Tall is the Eiffel Tower? Eiffel Tower is the tallest man-made marvel in France. This massive latticework structure is owned by the city of Paris and is a ...

Eiffel Tower - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Eiffel_Tower**
History · Design of the **tower** · Tourism · Attempted relocation · Economics
The **tower** stands 320 metres (1,050 ft) **tall**, about the same height as an 81-story building. During its construction, **the Eiffel Tower** surpassed the Washington Monument to ...

How tall is The Eiffel Tower? | Height of the Celebrities
howtallis.info/**eiffel-tower**.html
Want to know **How tall is The Eiffel Tower**? Ok, here we are. Find the answer of this question. **Eiffel Tower** height is 324 metres (1,063 ft)

How tall is the Eiffel tower? - Yahoo! Answers India
in.answers.yahoo.com/question/index?qid=20071125233320AAnuotX
Resolved · 7 total answers
26/11/2007 · Best Answer: According to The Oxford Dictionary of Phrase and Fable: "**Eiffel Tower** a wrought-iron structure erected in Paris for the World Exhibition of ...

Eiffel Tower Height, **How Tall Is The Eiffel Tower?**
my**eiffeltower**.com/**eiffel-tower**-height
Find out the exact **Eiffel Tower** height here! Discover **how tall is the Eiffel Tower** - the most visited monument in the world!

How tall is the Eiffel Tower? - Yahoo! Answers
answers.yahoo.com/question/index?qid=20070216082426AA7wNpQ
Resolved · 3 total answers
16/2/2007 · Best Answer: **Eiffel Tower** (French: La Tour **Eiffel**, /tuʁ ɛfɛl/) is an iron **tower** built on the Champ de Mars beside the River Seine in Paris, France. It ...

© 2013-2024 NUS. All rights reserved.

Information extraction version

Google

how tall is the eiffel tower

Mun Kew Leong 3 + Shar

Search

About 2,420,000 results (0.26 seconds)

Web
Images
Maps
Videos
News
Shopping
More

Show search tools

324 m
Eiffel Tower, Height
Hide details

How tall is the Eiffel Tower - answers.com
The Eiffel tower is located in Paris, France. It is **324 m** ...

Eiffel Tower Height, How ... - myeiffeltower.com
Eiffel Tower height - **324 m**. of iron parts! In 1889 ...

How tall is the eifel tower? - Yahoo! Answers - yahoo.com
... the structure is **324 m** (1058 ft) ...

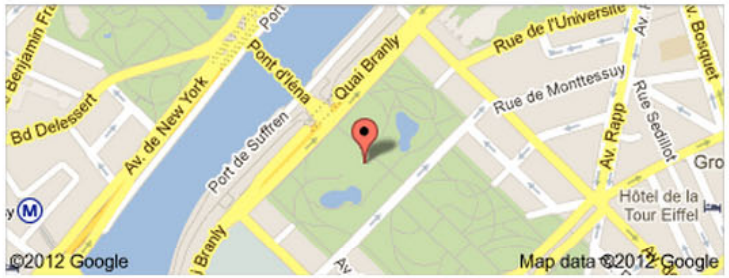
Is this accurate? Yes - No

Eiffel Tower - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Eiffel_Tower**
Jump to Lattice towers **taller** than the **Eiffel Tower**: The **Eiffel Tower** is a puddled iron lattice tower located on the Champ de Mars in Paris, named ...
History - Design of the tower - Tourism - Attempted relocation

How tall is the Eiffel Tower
wiki.answers.com › ... › France › Paris › Eiffel Tower
The **Eiffel Tower** is 100 meters **tall**. **How tall is the Eiffel tower** and which city is it in? The **Eiffel tower** is located in Paris, France. It is **324 m** (1,063 ft) **tall**.

Would Shard critics once have slammed the **Eiffel Tower**? | Daniel ...
www.guardian.co.uk/commentisfree/.../shard-critics-**eiffel-tow**... Share
5 Jul 2012 – But while the main body of the **Eiffel Tower** is only 300 metres **tall**, it stands at 324 metres once you add the base and television antenna.

Eiffel Tower Directions



en.wikipedia.org

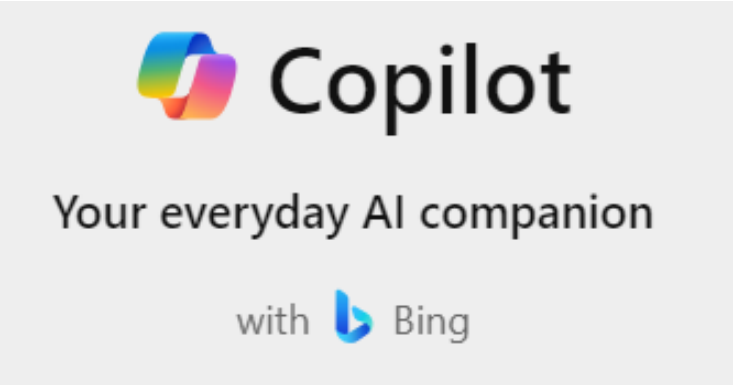
The Eiffel Tower is a puddled iron lattice tower located on the Champ de Mars in Paris, named after the engineer Gustave Eiffel, whose company designed and built the tower. Wikipedia

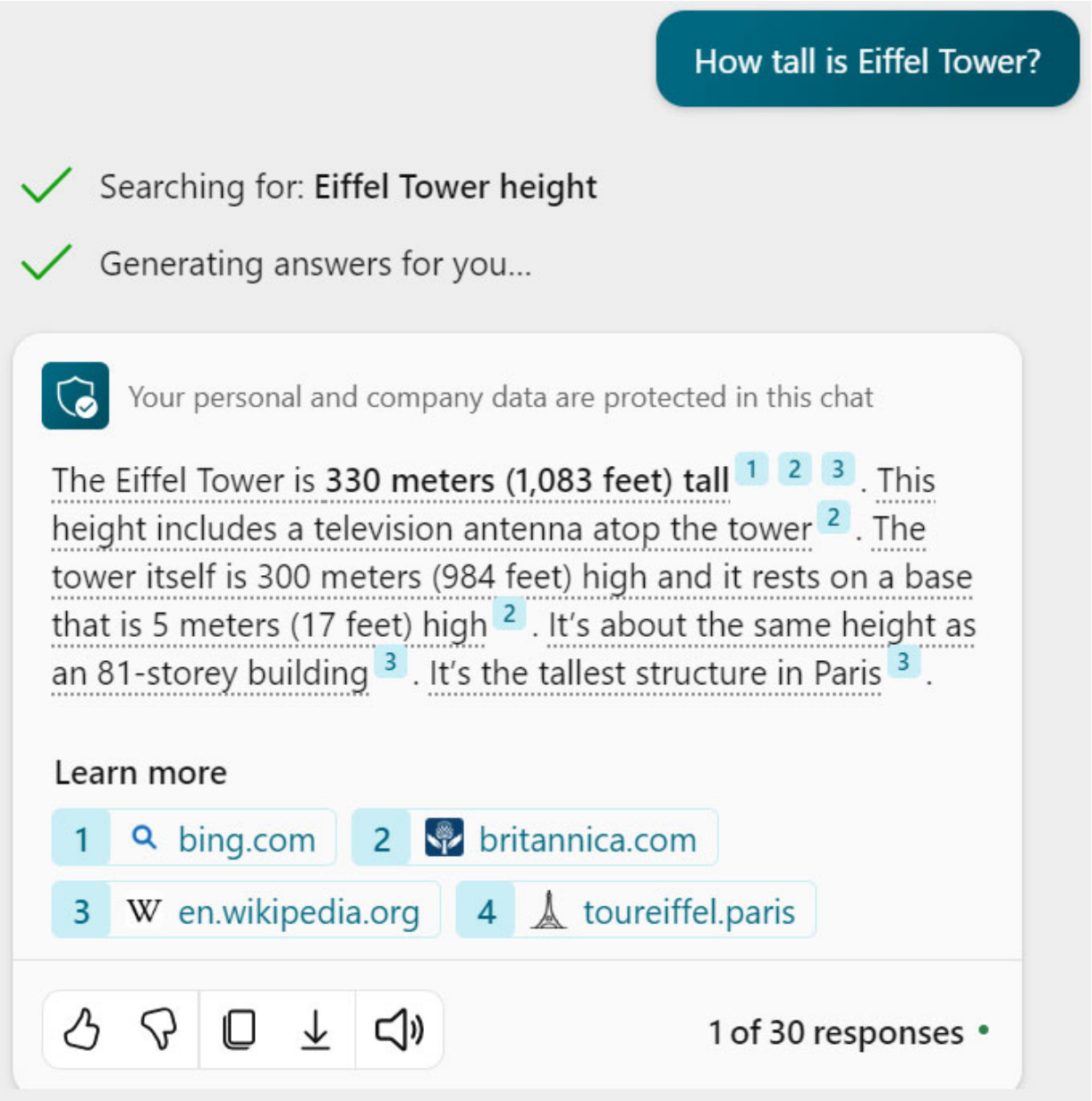
Height: 324 m
Construction started: January 28, 1887
Hours: Mon-Sun 9:30am–11:45pm
Address: Avenue Gustave Eiffel, 75007 Paris, France
Phone: 0892 70 12 39
Architect: Stephen Sauvestre

People also search for

© 2013-2024 NUS. All rights reserved.

The QA version





© 2013-2024 NUS. All rights reserved.

What is information extraction?

- The automatic extraction of (possibly pre-specified) information from natural language documents
  - Facts about types of entities, events, relations

- The automatic population of a structured information source (template) from natural language documents (i.e., create a table!)
  - Documents may be semi-structured (eg., patents), unstructured (e.g., websites) or free text (e.g., documents)

© 2013-2024 NUS. All rights reserved.

Drilling down on information…

- Entity/Concept:
  - an object of interest (date, person, building,…)
- Attribute:
  - property of an entity (name, height, type)
- Fact:
  - a predicate about an entity, e.g., tall(X)
  - a relation between two or more entities, e.g., married(X,Y)
- Event
  - an abstraction over a possibly sequential relationship of several entities e.g., terrorist attack, birthday party, etc.

© 2013-2024 NUS. All rights reserved.

What is named entity recognition?

- Recognition of particular types of proper noun phrases, specifically *persons*, *organizations*, *locations*, and sometimes *money*, *dates*, *times*, and *percentages*.
- Very useful in text analytics applications, by turning verbose text data into a more compact structural form

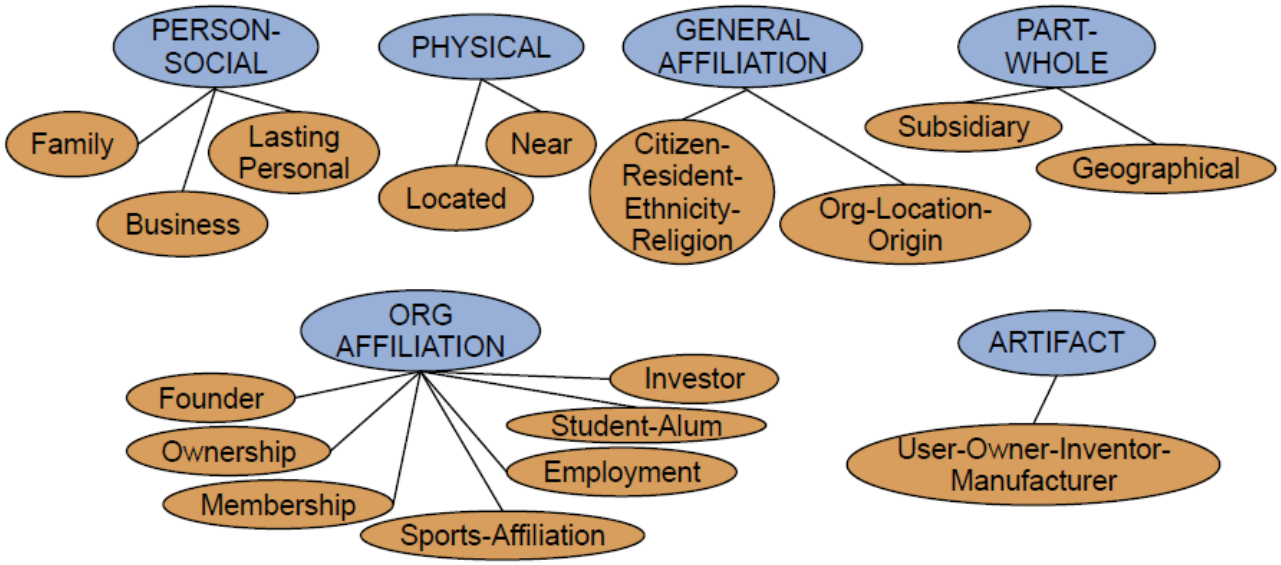
[**LOC Houston**] , Monday, July 21 -- Men have landed and walked on the moon. Two [**MISC Americans**] , astronauts of [**ORG Apollo**] 11, steered their fragile four-legged lunar module safely and smoothly to the historic landing yesterday at 4:17:40 P.M., Eastern daylight time. [**PER Neil A. Armstrong**] , the 38-year-old civilian commander, radioed to earth and the mission control room here: "[**LOC Houston**] , [**ORG Tranquility Base**] here; the Eagle has landed."

Generated by UIUC NER system

© 2013-2024 NUS. All rights reserved.

What is relation extraction?

- Extracting relations between entities, often binary relations, <*relation*, *entity1*, *entity2*>, like *child-of*, *employee-of*, *part-of*, *is-a*, etc.
- Often to populate a relational database, task -specific, domain-specific.
- Some relation categories:



© 2013-2024 NUS. All rights reserved.

Relations

- Relation names, types of entities involved
- Can be expressed as RDF (Resource Description Framework) triple
  - Subject(entity)-predicate(relation)-object(entity)

Relations	Types	Examples
Physical-Located	PER-GPE	**He** was in **Tennessee**
Part-Whole-Subsidiary	ORG-ORG	**XYZ**, the parent company of **ABC**
Person-Social-Family	PER-PER	**Yoko**'s husband **John**
Org-AFF-Founder	PER-ORG	**Steve Jobs**, co-founder of **Apple**...

© 2013-2024 NUS. All rights reserved.

What is event extraction?

- Finding events in which the entities participate, like *fare increases* by airlines, *price drop* of stocks, etc.
- Domain-specific, task-specific
- Usually verbs, sometimes NPs
- Event coreference

[EVENT Citing] high fuel prices, United Airlines [EVENT said] Friday it has [EVENT increased] fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR Corp., immediately [EVENT matched] [EVENT the move], spokesman Tim Wagner [EVENT said]. United, a unit of UAL Corp., [EVENT said] [EVENT the increase] took effect Thursday and [EVENT applies] to most routes where it [EVENT competes] against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

© 2013-2024 NUS. All rights reserved.

Related tasks

- Extracting temporal expressions (when the event happened), such as days, dates, times, sometimes normalized

Absolute	Relative	Durations
April 24, 1916	yesterday	four hours
The summer of '77	next semester	three weeks
10:15 AM	two weeks from yesterday	six days
The 3rd quarter of 2006	last quarter	the last three quarters

- Create event timeline (temporal ordering of events)

© 2013-2024 NUS. All rights reserved.

Example (Entities and Facts)

- Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.
- Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CFO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

Persons:	Organizations:	Locations:	Artifacts:	Dates:
Fletcher Maddox	UCSD Business School	La Jolla	Geninfo	June 1999
Dr. Maddox	La Jolla Genomatics	CA	Geninfo	
Oliver	La Jolla Genomatics			
Oliver	L.J.G.			
Ambrose				
Maddox				

PERSON	Employee_of	ORGANIZATION
Fletcher	Employee_of	UCSD Business
Maddox	Employee_of	School
Fletcher	Employee_of	La Jolla Genomatics
Maddox	Employee_of	La Jolla Genomatics
Oliver		La Jolla Genomatics
Ambrose		
ARTIFACT	Product_of	ORGANIZATION
Geninfo	Product_of	La Jolla Genomatics
LOCATION	Location_of	ORGANIZATION
La Jolla	Location_of	La Jolla Genomatics
CA	Location_of	La Jolla Genomatics

From: Ron Feldman, *Information Extraction: Theory and Practice*, http://cs.fit.edu/~pkc/icdm03/printing/tutorials/extraction/extraction.tutorial.pdf

© 2013-2024 NUS. All rights reserved.

Events and Attributes

COMPANY-FORMATION_EVENT:

COMPANY:	La Jolla Genomatics
PRINCIPALS:	Fletcher Maddox Oliver Ambrose
DATE:	
CAPITAL:	

RELEASE-EVENT:

COMPANY:	La Jolla Genomatics
PRODUCT:	Geninfo
DATE:	June 1999
COST:	

NAME:	Fletcher Maddox Maddox
DESCRIPTOR:	former Dean of the UCSD Business School his father the firm's CEO
CATEGORY:	PERSON
NAME:	Oliver
DESCRIPTOR:	His son Chief Scientist
CATEGORY:	PERSON
NAME:	Ambrose
DESCRIPTOR:	Oliver's brother the CFO of L.J.G.
CATEGORY:	PERSON
NAME:	UCSD Business School
DESCRIPTOR:	
CATEGORY:	ORGANIZATION
NAME:	La Jolla Genomatics L.J.G.
DESCRIPTOR:	
CATEGORY:	ORGANIZATION
NAME:	Geninfo
DESCRIPTOR:	its product
CATEGORY:	ARTIFACT
NAME:	La Jolla
DESCRIPTOR:	the Maddox family's hometown
CATEGORY:	LOCATION
NAME:	CA
DESCRIPTOR:	
CATEGORY:	LOCATION

From: Ron Feldman, *Information Extraction: Theory and Practice*, http://cs.fit.edu/~pkc/icdm03/printing/tutorials/extraction/extraction.tutorial.pdf

© 2013-2024 NUS. All rights reserved.

Unstructured Text

POLICE ARE INVESTIGATING A ROBBERY THAT (
ELEVEN STORE LOCATED AT 2545 LITTLE RIVE
LINCOLNIA AREA ABOUT 12:30 AM FRIDAY. A 2
ALEXANDRIA AREA EMPLOYEE WAS APPROAC
DEMANDED MONEY. SHE RELINQUISHED AN U
OF CASH AND THE MEN LEFT. NO ONE WAS IN
DESCRIBED AS BLACK, IN THEIR MID TWENTIE
FEET NINE INCHES TALL, WITH MEDIUM BUILD
CLEAN SHAVEN. THEY WERE BOTH WEARING
BLACK COATS. ANYONE WITH INFORMATION A
OR THE SUSPECTS INVOLVED IS ASKED TO CA
5555.

Structured (Desired) Information

Crime	Address	Town	Time	Day
ABDUCTION	8700 BLOCK OF LITTLE RIVER TURNPIKE,	ANNANDALE	11:30 PM	SUNDAY
...	...	...	...	...
ROBBERY	7- ELEVEN STORE LOCATED AT 2545 LITTLE RIVER TURNPIKE,	LINCOLNIA	12:45 AM	FRIDAY
ROBBERY	7- ELEVEN STORE LOCATED AT 5624 OX ROAD,	FAIRFAX	3:00 AM	FRIDAY

From: Ron Feldman, *Information Extraction: Theory and Practice*, http://cs.fit.edu/~pkc/icdm03/printing/tutorials/extraction/extraction.tutorial.pdf

© 2013-2024 NUS. All rights reserved.

Exercise

Read this text:

- During the Boston Marathon on April 15, 2013, two pressure cooker bombs exploded at 2:49 pm EDT killing 3 people and injuring 264 others. The bombs exploded about 13 seconds and 210 yards (190 m) apart, near the finish line on Boylston Street. The Federal Bureau of Investigation (FBI) took over the investigation, and on April 18, released photographs and surveillance video of two suspects. The suspects were identified later that day as the Chechen brothers Dzhokhar and Tamerlan Tsarnaev.

(from: https://en.wikipedia.org/wiki/Boston_Marathon_bombings)

Fill up this template:

Attribute	Value
Occurrence	Terrorist Attack
Event	Boston Marathon
Location	Boylston Street
Start Date	15 April 2013
Start Time	14:49
End Date	15 April 2013
End Time	14:49 (13 seconds later)
Method	Pressure cooker bombs
Incident_1	Bombs exploded about 13 sec & 210 yards
Incident_2	
Incident_3	
Perpetrators	2 suspects
Caught (Y/N)	N

© 2013-2024 NUS. All rights reserved.

How does an IE System work?

Greedy heuristics

- Start with nouns or names or simple <u>named entities (NEs)</u>
  - Identify NEs using dictionaries = lists = gazetteer = …
  - Eg, name lists: boy names, girl names, world names,…
  - Eg, days, months, cities, countries, … (can be multiwords)
  - Proper nouns = words that start with Capitals (in the middle of sentence)
- If nouns or names are <u>contiguous</u>, then aggregate them
  - Jorge + Mario + Bergoglio = "Jorge Mario Bergoglio"
  - Pope + Francis = "Pope Francis"
  - Catholic + Church = "Catholic Church"
- Some lexical and syntactic markup can help
  - Rule: if **X** $_{\textbf{[title]}}$ **+ Y** $_{\textbf{[name]}}$ **+ "of" + Z** $_{\textbf{[place]}}$ then **X Y of Z** is a concept
  - Given:

    Cardinal $_{[\text{title}]}$
    Jorge Mario Bergoglio$_{[\text{name}]}$
    Buenos Aires$_{[\text{place}]}$
  - with the rule above: "Cardinal Jorge Mario Bergoglio of Buenos Aires" is a concept

© 2013-2024 NUS. All rights reserved.

Exercise

Professor Tan Eng Chye, NUS Deputy President and Provost, and Professor Menahem Ben-Sasson, President of HUJ signed the joint degree agreement at NUS, in the presence of Ambassador of Israel to Singapore Her Excellency Amira Arnon and about 30 invited guests.

- *The following is a simple lexical rule to recognize names of universities in the paragraph above:*

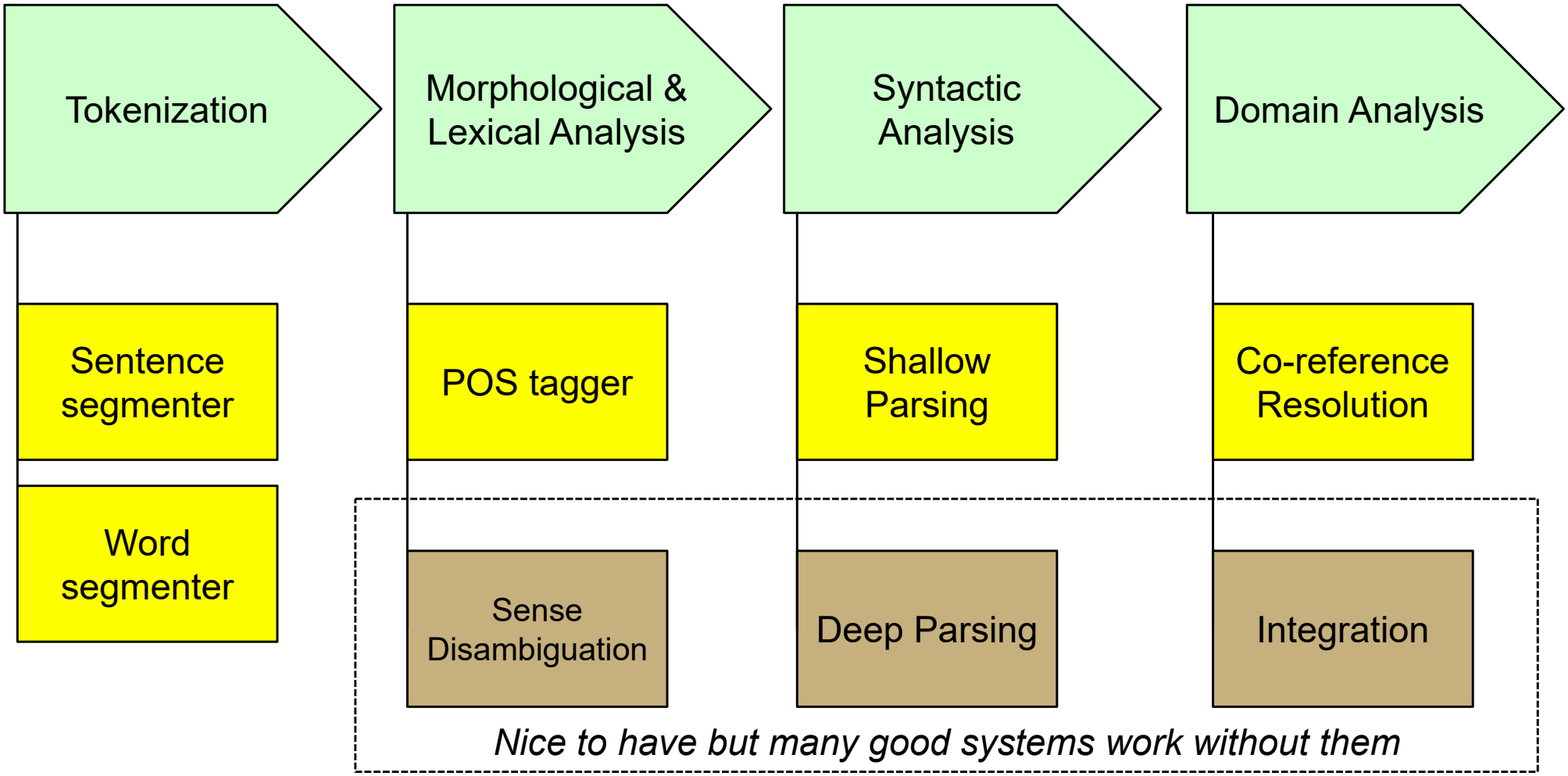
$$\textbf{If "The"} + X_{[capitalized]} + \textbf{"University"} + \textbf{"of"} + Y_{[place]}$$

$$\textbf{then "The X University of Y" is the name of a university}$$

- *Write a rule similar in structure to recognize the names of persons mentioned in the paragraph above. You may assume reasonable dictionaries (lists) exist and that suitable parts of speech and meaningful markup has been done:*

© 2013-2024 NUS. All rights reserved.

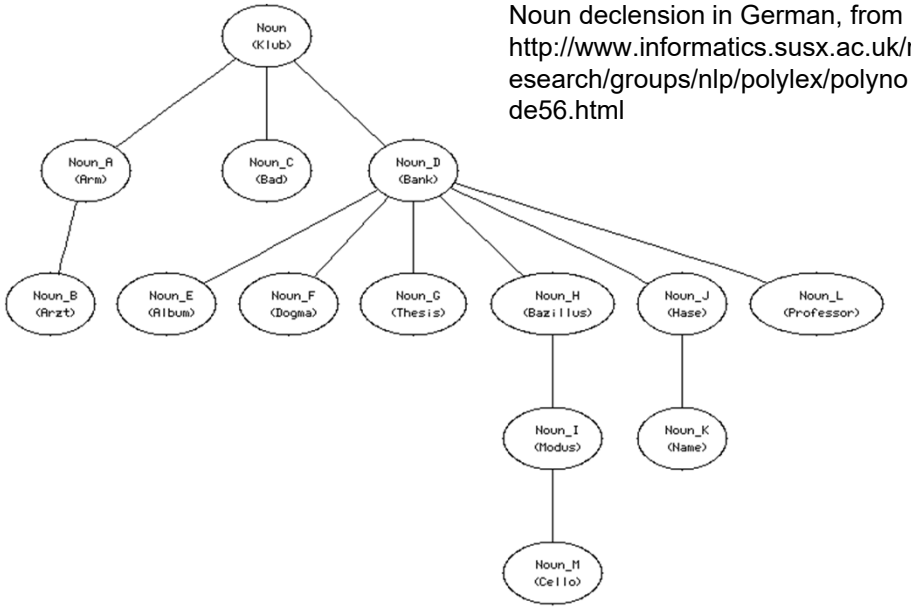
Main components of an IE system



© 2013-2024 NUS. All rights reserved.

Tokenization

- Language Considerations
  - Word boundary easy in roman script, ambiguous in character scripts
  - Morphology is easy in English, much harder in German, Hebrew
  - English is one-dimensional, other languages have more than one



Noun declension in German, from http://www.informatics.susx.ac.uk/research/groups/nlp/polylex/polynode56.html

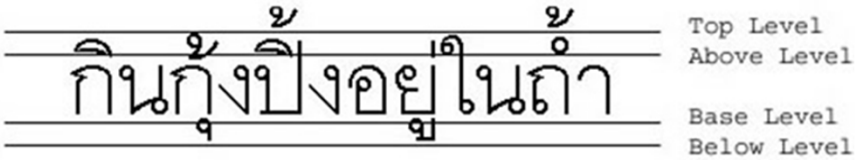
original, un-segmented text

再往远些看，随着汉字识别和语音识别技术的发展，
中文计算机用户将跨越语言差异的鸿沟，
在录入上走向中西文求同的道路。

separated word entities after segmentation

再 往 远 些 看 ， 随着 汉字 识别 和 语音 识别 技术 的 发展 ，
中文 计算机 用户 将 跨越 语言 差异 的 鸿沟 ，
在 录入 上 走 向 中 西文 求 同 的 道路 。

Word segmentation in Chinese, from http://blog.faroo.com/2009/06/09/lightweight-chinese-word-segmentation/



Word composition in Thai, from http://www-01.ibm.com/software/globalization/topics/thai/character.html

© 2013-2024 NUS. All rights reserved.

POS Tagging

- To determine POS or grammatical category of a term
  - Dictionary with word-POS correspondence is needed

```
IN/ About CD/ six CC/ and DT/ a JJ/ half NNS/ hours RB/ later ,/ , NNP/ Mr. NNP/
Armstrong VBD/ opened DT/ the NN/ landing NN/ craft POS/ 's NN/ hatch ,/ , VBD/
stepped RB/ slowly IN/ down DT/ the NN/ ladder CC/ and VBD/ declared IN/ as PRP/ he
VBD/ planted DT/ the JJ/ first NN/ human NN/ footprint IN/ on DT/ the NN/ lunar NN/
crust :/ : ``/ " DT/ That VBZ/ 's CD/ one JJ/ small NN/ step IN/ for NN/ man ,/ , CD/ one
JJ/ giant NN/ leap IN/ for NN/ mankind ./ . ''/ "
```

Generated by UIUC POS Tagger

- POS disambiguation
  - 14-15% of words in the vocabulary, mostly common words, are ambiguous, hence 55-67% of word tokens in running text are ambiguous.
  - Baseline: choose the tag which is the most frequent in the training corpus
  - Using rule-based or stochastic approach

earnings growth took a **back/JJ** seat
a small building in the **back/NN**
a clear majority of senators **back/VBP** the bill
Dave began to **back/VB** toward the door
enable the country to buy **back/RP** about debt
I was twenty-one **back/RB** then

© 2013-2024 NUS. All rights reserved.

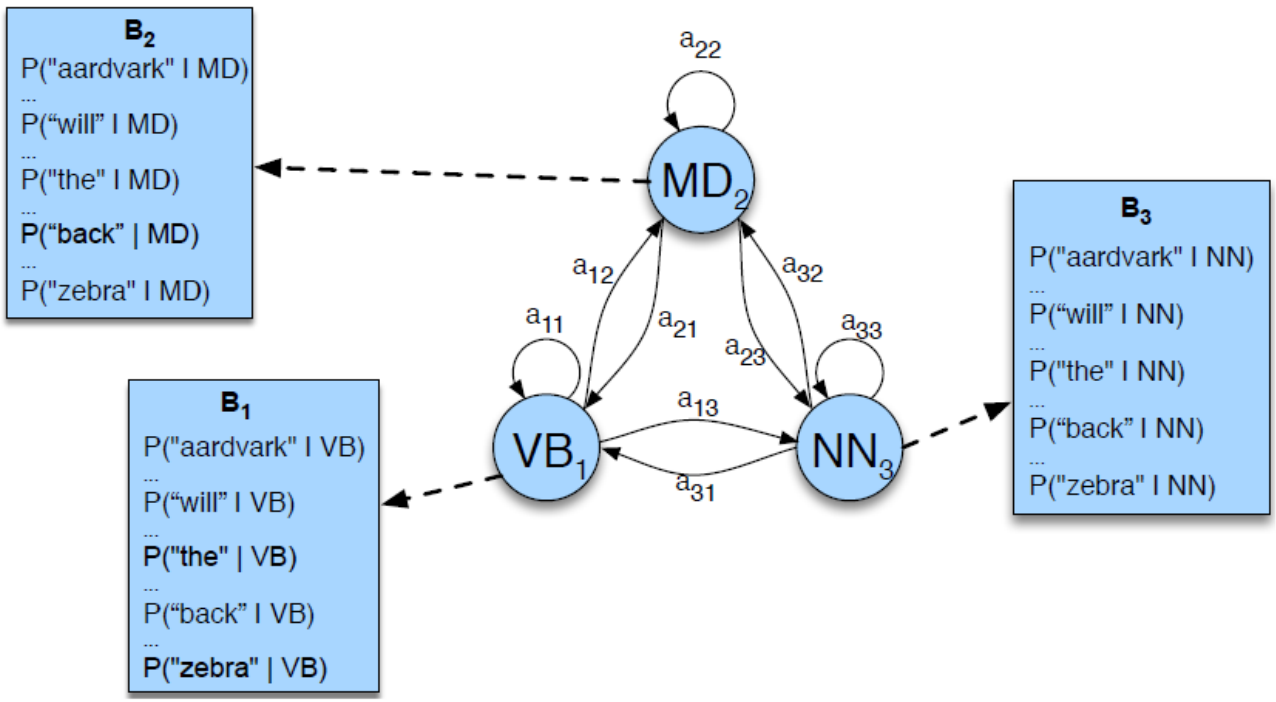
POS Taggers

- POS tagging is a sequence labelling task
- Rule-based - e.g. Brill's tagger by Eric Brill
  - Error-driven transformation-based tagger learned from annotated data
  - Initially assign the most frequent tag to each word, based on dictionary and morphological rules
  - Contextual rules are then applied repeatedly to correct any errors
- Stochastic taggers – e.g. CLAWS, Viterbi, Baum-Welch, etc.
  - based on Hidden Markov Models (HMMs) and n-gram probabilities
  - Manually tagged corpus is needed to estimate probabilities
- Many machine learning methods have also been applied. (Stanford's Statistical NLP website lists many free taggers.)

© 2013-2024 NUS. All rights reserved.

Sequence labelling with HMM

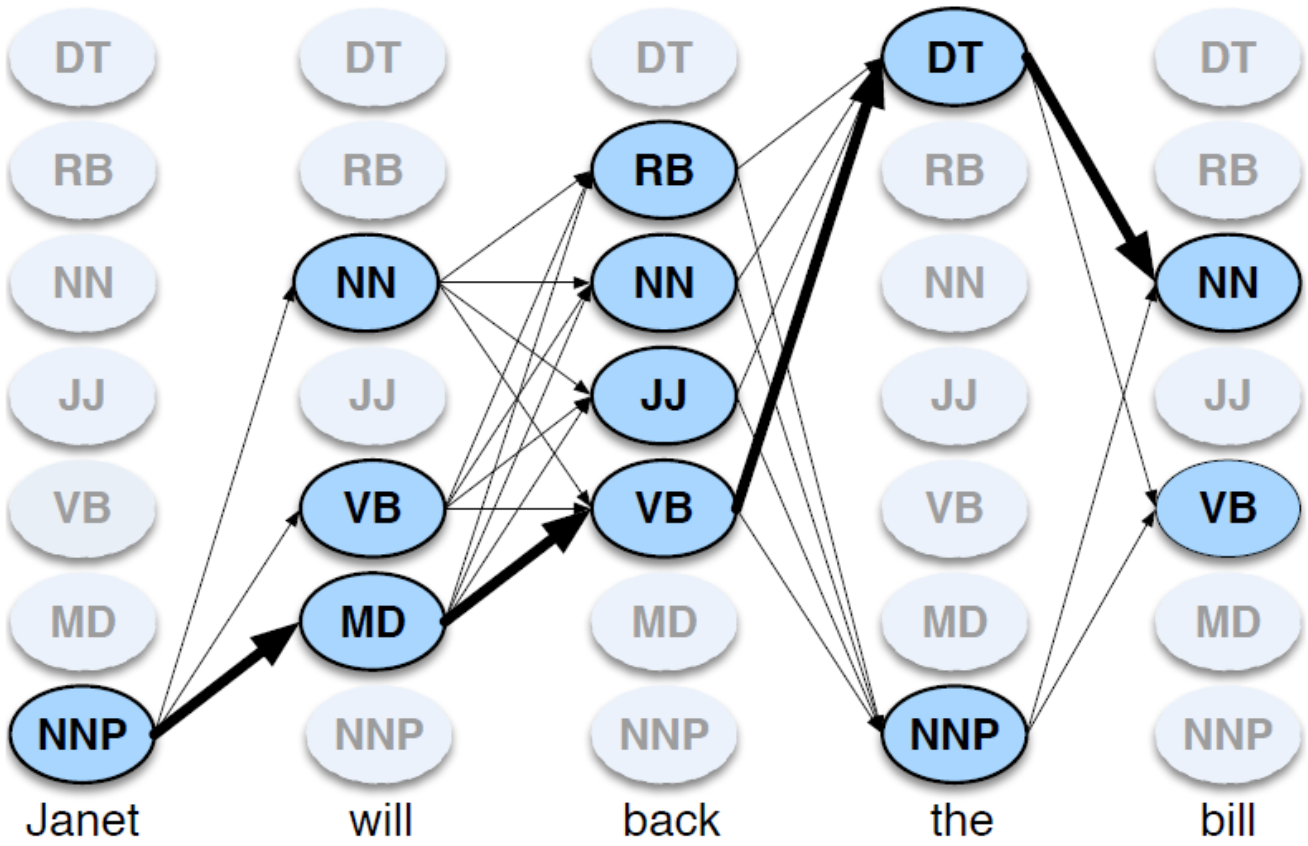
- HMM: a probabilistic sequence model/classifier, trained from tagged corpus.
  - A - *transition probabilities* – the probability of a tag occurring given the previous tag.
  - B - *observation likelihoods* – the probability that a given tag will be associated with a given word



© 2013-2024 NUS. All rights reserved.

HMM Decoding

- Given a sequence of words (observations), and an HMM, compute a probability distribution over possible sequences of labels (states) and chooses the best label sequence.



© 2013-2024 NUS. All rights reserved.

Shallow Parsing / Chunking

- To identify phrases in a text (noun phrases, verb phrases, and prepositional phrases, etc.)
- Largely stochastic techniques based on probabilities derived from an annotated corpus – segmenting and labeling
- Faster, more robust than full parsing
- Also a sequence labelling task

[NP About six and a half hours] [ADVP later] , [NP Mr. Armstrong] [VP opened] [NP the landing craft] [NP 's hatch] , [VP stepped] [ADVP slowly] [PP down] [NP the ladder] and [VP declared] [SBAR as] [NP he] [VP planted] [NP the first human footprint] [PP on] [NP the lunar crust] : "[NP That] [VP 's] [NP one small step] [PP for] [NP man] , [NP one giant leap] [PP for] [NP mankind] ."

Generated by UIUC chunker

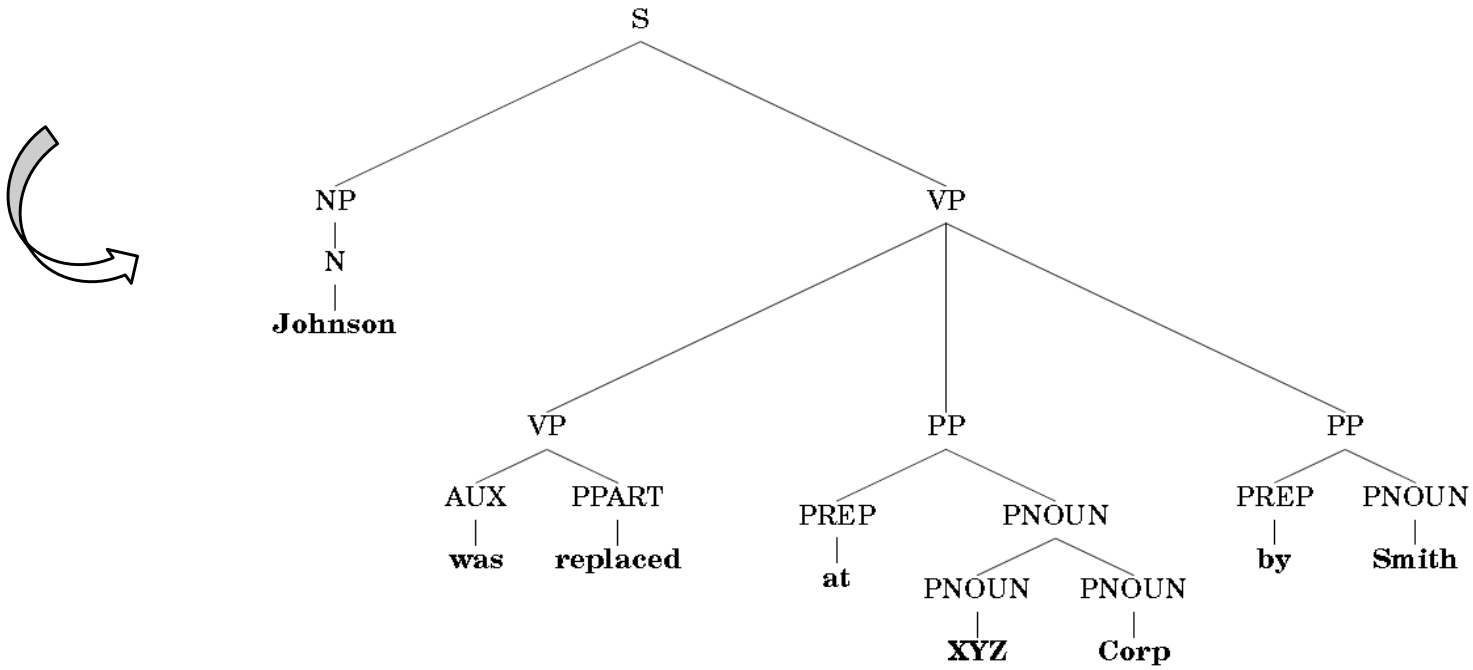
The	cat	sat	on	the	mat
B-NP	I-NP	B-VP	B-PP	B-NP	I-NP

© 2013-2024 NUS. All rights reserved.

Parsing

- Or *Syntactic Analysis*, the more sophisticated kind of text processing
- To produce a full parse of a sentence, typically as a tree, with syntactic functions of each word (e.g. subject, object, etc.)
- Comparatively expensive process, but it can provide information that shallow parsing can not provide.

Johnson was replaced at XYZ Corp by Smith



© 2013-2024 NUS. All rights reserved.

Trees

Trees can be represented in bracketed forms:

Tagging

```
John/NNP  was/VBD  replaced/VBN  at/IN  XYZ/NNP  Corp/NNP  by/IN  Smith/NNP  ./.
```

Parse

```
(ROOT
  (S
    (NP (NNP John))
    (VP (VBD was)
      (VP (VBN replaced)
        (PP (IN at)
          (NP (NNP XYZ) (NNP Corp)))
        (PP (IN by)
          (NP (NNP Smith)))))
    (. .)))
```

From Stanford Parser

© 2013-2024 NUS. All rights reserved.

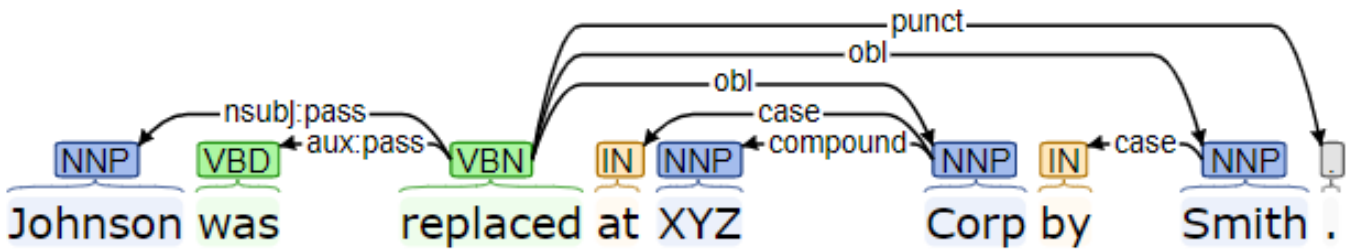
Dependencies

- Typed dependencies

Typed dependencies, collapsed

```
nsubjpass(replaced-3, John-1)
auxpass(replaced-3, was-2)
root(ROOT-0, replaced-3)
nn(Corp-6, XYZ-5)
prep_at(replaced-3, Corp-6)
agent(replaced-3, Smith-8)
```

- Usually visualized as:



From Stanford Parser

© 2013-2024 NUS. All rights reserved.

From Syntax to Semantics

- Semantic analysis can be applied on top of parsing result to help identify the right entities for the text mining task.

	SRL	Charniak
John	old thing [A1]	(S1 (S (NP (NNP John))
was		(VP (AUX was)
replaced	V: replace	(VP (VBN replaced)
at	location [AM-LOC]	(PP (IN at)
XYZ		(NP (NNP XYZ)
Corp		(NNP Corp)))
by	replacer [A0]	(PP (IN by)
Smith		(NP (NNP Smith)))))
.		(. .)))

Generated by UIUC Semantic Role Labeling system

© 2013-2024 NUS. All rights reserved.

Challenges in Parsing

- Robustness – graceful degradation
  - The input may not conform to what is normally expected
  - Ill-formed input or lack of coverage of grammars
  - To recover as much meaningful information as possible
- Disambiguation
  - Ambiguity accumulated from earlier steps can result in combinatorial increase of possible parses
  - Return the *n* best analyses, if not one, to the next level of processing
- Efficiency
  - Theoretical time complexity of most formalisms are polynomial

© 2013-2024 NUS. All rights reserved.

Word Sense Disambiguation

- Words are also ambiguous as to their meaning or reference (polysemous)
  - E.g. *table*: 1. a piece of furniture with a flat top supported by legs
    2. A list of numbers, facts, or information arranged in rows across and down a page
- Disambiguation of meanings in context has not been well solved, partly due to the lack of semantic concordances, corpora of disambiguated text to serve as training corpus for machine learning algorithms
  - E.g. You will find$^{9}_{v}$ that avocado$^{1}_{n}$ is$^{1}_{v}$ unlike$^{1}_{j}$ other$^{1}_{j}$ fruit$^{1}_{n}$ you have ever$^{1}_{r}$ tasted$^{2}_{v}$
- Usually not applied in a typical text analytics application

© 2013-2024 NUS. All rights reserved.

WordNet

- A large lexical database of English
- Created and maintained by the Cognitive Science Laboratory of Princeton University
- *Nouns*, *verbs*, *adjectives* and *adverbs* are grouped into sets of cognitive synonyms (*synsets*), each expressing a distinct concept

Number of words, synsets, and senses

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

Statistics from WordNet website
http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

© 2013-2024 NUS. All rights reserved.

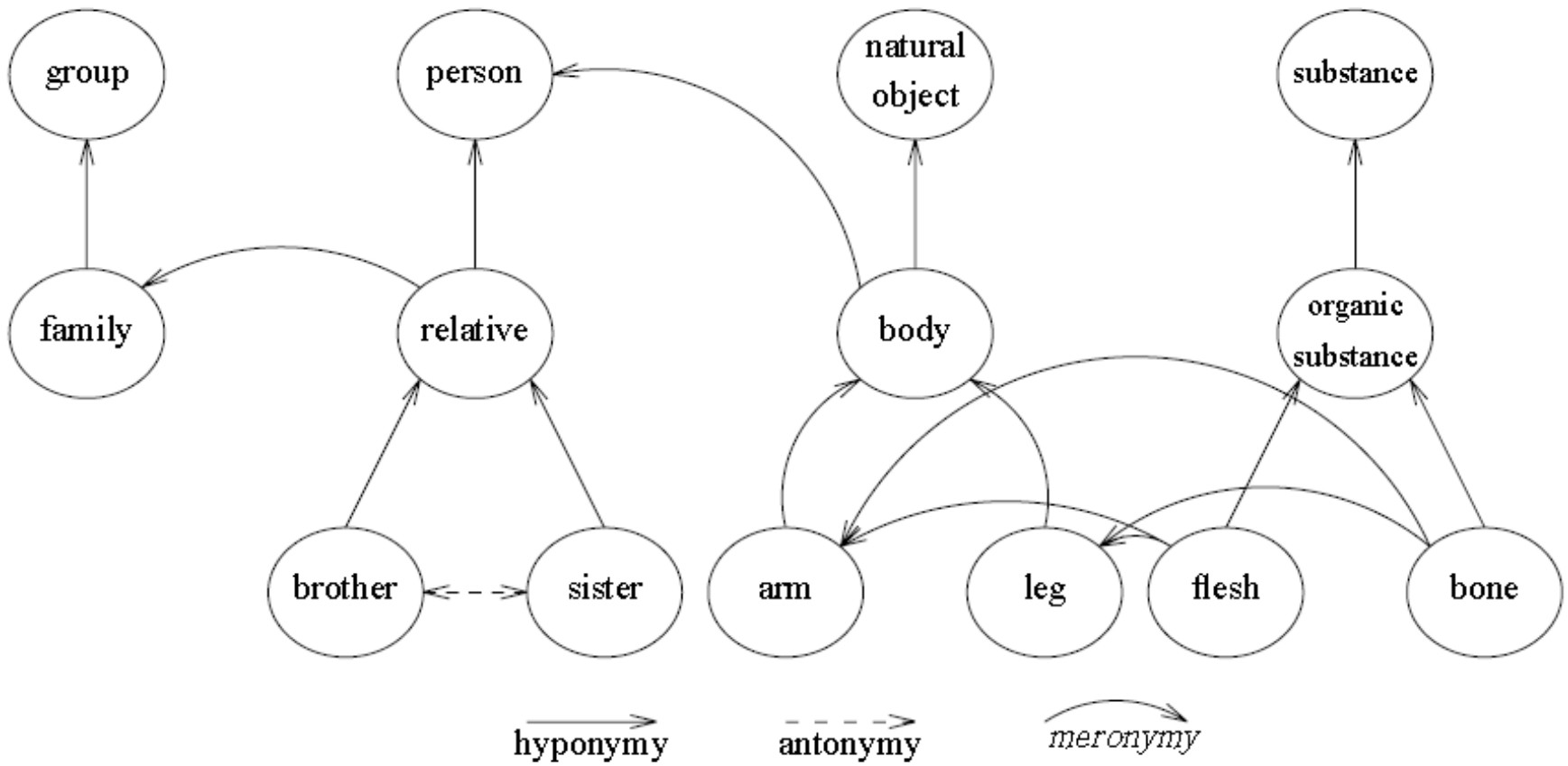
WordNet

- Synsets are linked by conceptual-semantic and lexical relations
  - Lexical relations
    - Synonymy – e.g. *shut* and *close*, *happy* and *joyful*
    - Antonymy – e.g. *wet* and *dry*, *young* and *old*, *happy* and *sad*
    - Morphological relations
  - Semantic relations
    - Hyponymy (or ISA relation, super-subordinate relation) – e.g. *apple* and *fruit*, *bed* and *furniture*, *communicate* and *talk* and *whisper*
    - Meronymy (part-whole relation) – e.g. *leg* and *chair*
  - And more…

© 2013-2024 NUS. All rights reserved.

WordNet

Figure 2. Network representation of three semantic relations among an illustrative variety of lexical concepts



From *Nouns in WordNet: A Lexical Inheritance System*

© 2013-2024 NUS. All rights reserved.

WordNet

- Example information in Wordnet for "happy":

Adjective

- (37)S: (adj) **happy#1** (enjoying or showing or marked by joy or pleasure)
- (2)S: (adj) felicitous#2, **happy#2** (marked by good fortune)
- S: (adj) glad#2, **happy#3** (eagerly disposed to act or to be of service)
- S: (adj) **happy#4**, well-chosen#1 (well expressed and to the point)

- Expanded view:

- (37)S: (adj) **happy#1** (enjoying or showing or marked by joy or pleasure)
  - *see also*
  - *similar to*
    - S: (adj) blessed#6 (characterized by happiness and good fortune)
    - S: (adj) blissful#1 (completely happy and contented)
    - S: (adj) bright#9 (characterized by happiness or gladness)
    - S: (adj) golden#2, halcyon#2, prosperous#3 (marked by peace and prosperity)
    - S: (adj) laughing#1, riant#1 (showing or feeling mirth or pleasure or happiness)
  - *attribute*
  - *antonym*
    - W: (adj) unhappy#1 [Opposed to: happy] (experiencing or marked by or causing sadness or sorrow or discontent)

© 2013-2024 NUS. All rights reserved.

WordNet

- Free and open source
- Proved useful for a wide range of Natural Language Processing applications
  - Word sense disambiguation
  - Word semantic distance measuring
  - Mono- and cross-lingual Information retrieval,
  - Question-answering systems
  - Machine translation
  - Document structuring and categorisation

© 2013-2024 NUS. All rights reserved.

Co-reference resolution

- What is co-reference
  - Determine relationship between entities which are related
    - Identity relation (morning star vs. evening star)
    - Whole-part relation
  - Simple version
    - Determine entities which have the same referent
      - Anaphora (Pronouns)
      - Proper names, proper nouns, noun phrases,…
    - Definite descriptions (may be time dependent)
      - Usain Bolt & “the fastest man in the world”

© 2013-2024 NUS. All rights reserved.

Recall our ealier example

Fletcher Maddox, former Dean of the UCSD Business School, announced the formation of La Jolla Genomatics together with his two sons. La Jolla Genomatics will release its product Geninfo in June 1999. Geninfo is a turnkey system to assist biotechnology researchers in keeping up with the voluminous literature in all aspects of their field.

Dr. Maddox will be the firm's CEO. His son, Oliver, is the Chief Scientist and holds patents on many of the algorithms used in Geninfo. Oliver's brother, Ambrose, follows more in his father's footsteps and will be the CFO of L.J.G. headquartered in the Maddox family's hometown of La Jolla, CA.

Type text here

© 2013-2024 NUS. All rights reserved.

Heuristics rule pronouns to reflect the nearest one.

Examples & Discussion

Coreference Examples

- Anaphora
  - The elephant stepped on the rabbit and it died.
  - The elephant stepped on the landmine and it died.
- Proper nouns
  - John Smith and Mary Brown were married this morning. The groom was dressed in a white tuxedo while the bride was…
- Definite descriptions
  - Usain Bolt has won the Olympic 100m gold medal. The fastest man in the world successfully defended his title last night.

What do you need to know to do accurate co-reference?

Died and subject. elephant or landmine which has more probabilty to die

Gender contraints -

Type text here

© 2013-2024 NUS. All rights reserved.

Approaches To Building IE systems

Main approaches

- Rule-based Systems
  - Hand-coded rules
    - Coded by linguists, with domain input
    - Iterative method based on document inspection
    - Slow but very good results
  - Induced (machine learning) rules
    - Fully machine learning
      - Given an annotated corpus, derive a basis set of rules that cover a pre-determined % of the annotated examples (and only the annotated examples)
      - Heuristic approach: one rule at a time!
    - Hybrid systems – machine learning to fine-tune the rules

© 2013-2024 NUS. All rights reserved.

Main approaches

- Machine learning-based systems
  - Start with a well-annotated corpus
  - Depending on the method (e.g., hidden Markov models), derive a model that generates the examples
  - Advantages compared to Rule based systems
    - Language independent (within representational limits)
    - No linguistic or domain knowledge needed in the team
    - Relatively small effort in creating the models
  - Issues
    - The complexity moves to the corpus – must be well annotated and must cover the full space of possibilities
    - Requires very large number of training examples to get good results

© 2013-2024 NUS. All rights reserved.

Rule-Based IE Methods

When to use rule-based systems?

- Rule-based systems can and do work well
  - Corpus is relatively static (in terms of vocabulary, language structure, etc.)
  - Can be fast especially in well-defined limited domains (compared to annotating training examples)
- A typical rule-based system comprises
  - Set of rules
  - Policies to control when and how (multiple) rules are applied, e.g., order, looping.

© 2013-2024 NUS. All rights reserved.

What does a rule look like?

- Form:
  - Match(pattern) then Do(action)

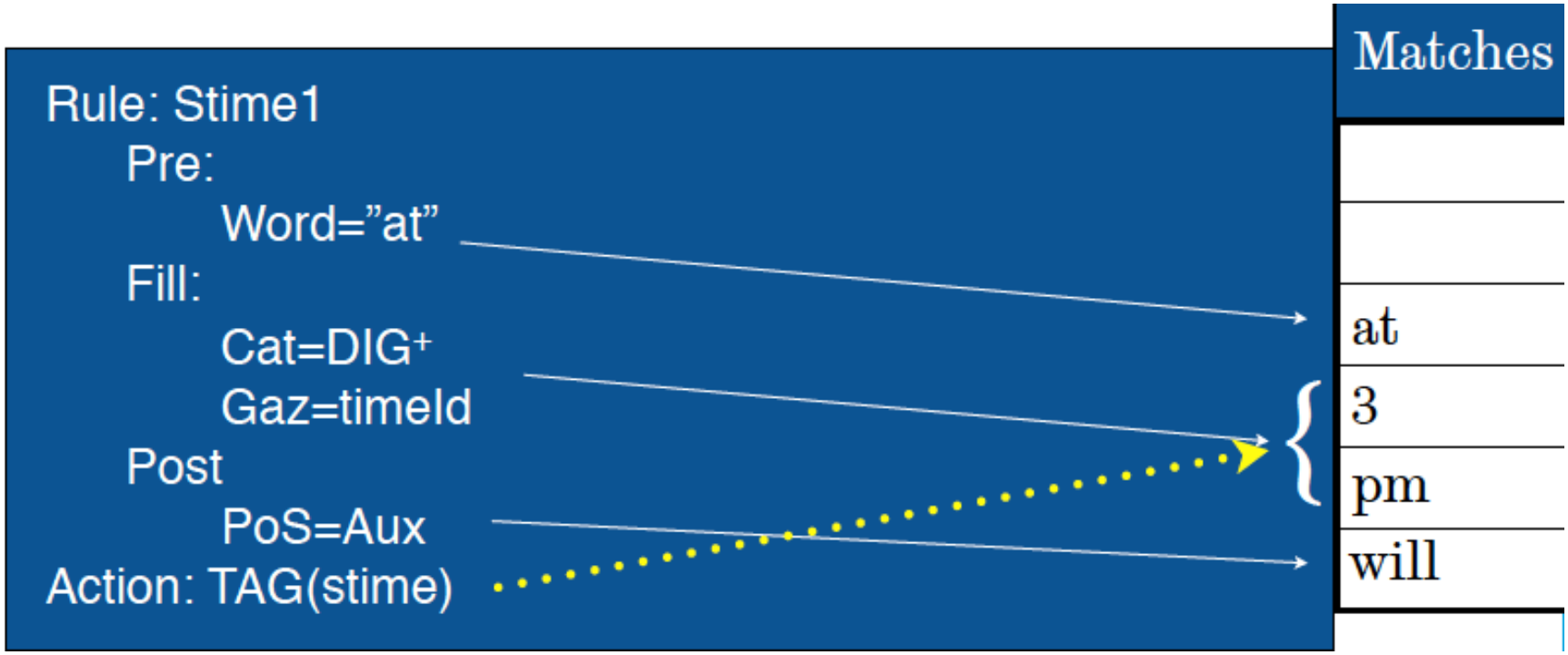
```
Rule: Company1                                          from gate.ac.uk
    ( ( {Token.orthography == upperInitial} )+
        {Lookup.kind == companyDesignator}
    ):match
  -->
     :match.NamedEntity = { kind=company, rule="Company1" }
```

Type text here

© 2013-2024 NUS. All rights reserved.

Types of IE recognition rules

- Lexical pattern matching – whole entity patterns
  - Classic rule: (left context + filler + right context) patterns
  - Simplistic: entities are considered independent
  - More natural for hand-coded rules



© 2013-2024 NUS. All rights reserved.

Types of IE recognition rules

- Structural pattern recognition
  - Typically used only for highly structured text
  - Can recognize more than one entity at a time

Example:

<p> Capitol Hill- *1* br twnhme. D/W W/D. Pkg incl
$*675*. *3* BR upper flr no gar. $*995*. (206)999-9999

Rule:

ID:7
Pattern: * ('Capitol Hill') * (*Digit*) * '$' (*Number*)
Output: Rental {Neighborhood $1} {Bedrooms $2} {Price $3}

Rule from: STEPHEN SODERLAND:
Learning Information Extraction Rules for Semi-structured and Free Text,
Machine Learning 1, 440

© 2013-2024 NUS. All rights reserved.

Example rule from CoreNLP

- TokensRegex
- E.g. for recognizing company names

```
$COMPANY_BEGINNING = "/[A-Z][A-Za-z]+/"
$COMPANY_ENDING = "/(Corp|Inc)\.?/“

{ ruleType: "tokens",
  pattern: ([{word:$COMPANY_BEGINNING} & {tag:"NNP"}]+
            [{word:$COMPANY_ENDING}]),
  action: Annotate($0, ner, "COMPANY"),
  result: "COMPANY_RESULT" }
```

© 2013-2024 NUS. All rights reserved.

Regex for patterns

- Regular expressions are very useful in extracting concepts expressed in a certain way, e.g. *currency*, *dates*, *e-mail addresses, phone numbers*, etc.
- Using operators to define strings containing certain characters, or not containing certain characters, or strings with pre-specified patterns of letters or numbers.
- Such patterns can be defined in a very compact way
  - E.g. regular expression for email addresses
    [a-zA-Z0-9._-]+@([a-zA-Z0-9.-]+\.)+[a-zA-Z]{2,4}
  - Strings matching this expression can then be extracted
    - E.g. *zhenzhen@nus.edu.sg*

© 2013-2024 NUS. All rights reserved.

Common Operators

- Special characters (operators) are used to define character patterns

Operator	Purpose
. (period)	Match any single character E.g. .in matches both ***Windows***, and ***Linux***
^	Match the empty string that occurs at the beginning of a line or string E.g. ^tre will not match ***stretch***
$	Match the empty string that occurs at the end of a line
\d	Match any single digit
\D	Match any single non-digit character
\w	Match any single alphanumeric character
\b	Match empty string, frequently used to indicate a word boundary E.g. \bhis\b matches *his* only, not *this* or *history*
|	Separate two alternative values E.g. gr(a|e)y matches both *gray* and *grey*
()	Group parts of search expression together

© 2013-2024 NUS. All rights reserved.

Common Operators

Operator	Purpose
?	Match the preceding character 0 or 1 time E.g. colou?r matches ***color*** *(0)* and ***colour*** *(1)*
*	Zero or more of the preceding character E.g. tre* matches ***tree*** *(2)*, ***tre**ad (1)*, and ***tr**ough (0)*
+	Match the preceding character 1 or more times E.g. tre+ matches ***tree***, and ***tre**ad*
[…]	Match anything inside the square brackets for one character position once E.g. [0-9] matches any character in the range 0-9 [abc] matches ***a***, ***b***, or ***c***
[^…]	Match any character excluding those in the square brackets E.g. [^A-M]in matches ***Win**dows*, but not ***Lin**ux*
{n}	Match the preceding character, or character range, n times E.g. [0-9]{3}-[0-9]{4} matches local phone number like *123-4567*
{n,m}	Match the preceding character at least n times but not more than m times E.g. [A-Z]{2,4} matches *com*, *sg*, but not *abcde*

© 2013-2024 NUS. All rights reserved.

Regular Expression

- Take a look at our email pattern regex again:

2 to 4 such characters

[a-zA-Z0-9._-]+@([a-zA-Z0-9.-]+\.)+[a-zA-Z]{2,4}

One or more such characters

One or more such characters

'xxx.' is repeated one or more times

Followed by '@'

Followed by '.'

© 2013-2024 NUS. All rights reserved.

Machine Learning based IE Methods

When to use ML based systems?

- Better performance
  - Machine learning (ML) on very large corpora is state-of-the-art
- Annotation based corpora for training
  - You have a well annotated corpora with many features
  - Various ML techniques from simple to sophisticated
  - You don't have domain or language resources in that area
  - Note that models don't transfer well across domains

© 2013-2024 NUS. All rights reserved.

NER as Sequence Labeling

- Or token classification
- Label each token in the sequence a tag, indicating if it's the boundary/part of an entity and its type
- IOB tagging: I-inside, O-outside, B-beginning
- Or IO tagging

Words	IOB Label	IO Label
American	B-ORG	I-ORG
Airlines	I-ORG	I-ORG
,	O	O
a	O	O
unit	O	O
of	O	O
AMR	B-ORG	I-ORG
Corp.	I-ORG	I-ORG
,	O	O
immediately	O	O
matched	O	O
the	O	O
move	O	O
,	O	O
spokesman	O	O
Tim	B-PER	I-PER
Wagner	I-PER	I-PER
said	O	O
.	O	O

[ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said.

© 2013-2024 NUS. All rights reserved.

Typical Features

- The word and its neighboring words
- POS of the word and its neighboring words
- Syntactic chunk label of the word and its neighboring words
- Their presence in a gazetteer
- Containing certain prefix/suffix
- Word shape, case information
- Presence of hyphen, etc.

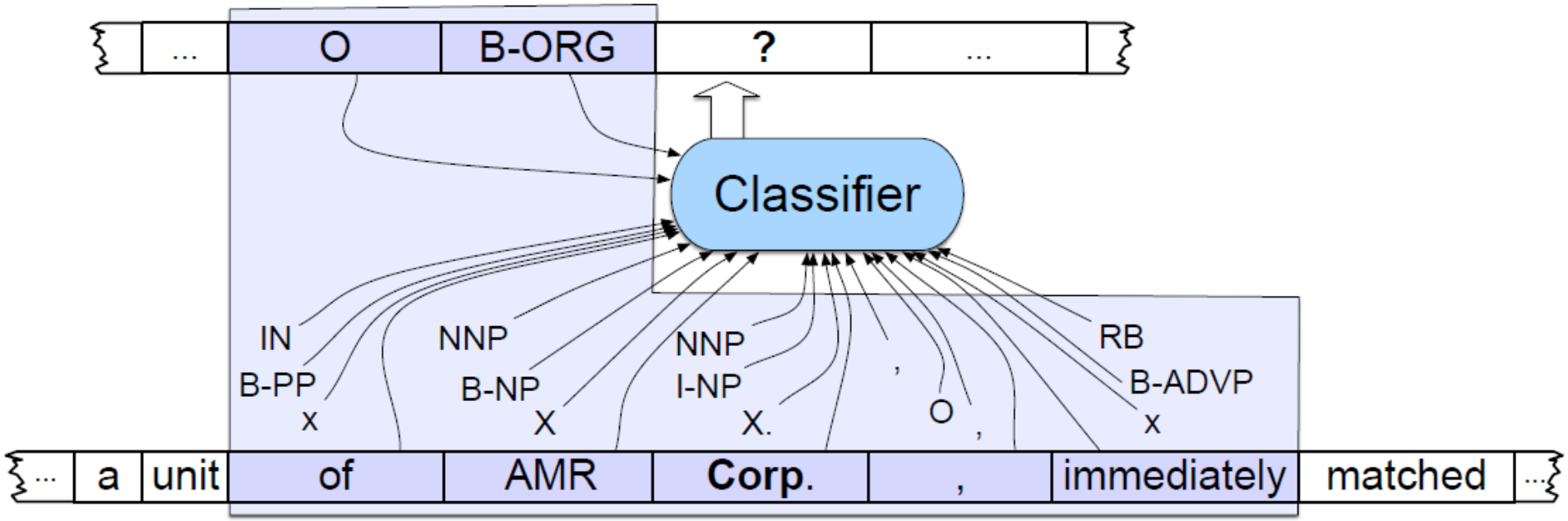
© 2013-2024 NUS. All rights reserved.

Example of word features

Word	POS	Chunk	Short shape	Label
American	NNP	B-NP	Xx	B-ORG
Airlines	NNPS	I-NP	Xx	I-ORG
,	,	O	,	O
a	DT	B-NP	x	O
unit	NN	I-NP	x	O
of	IN	B-PP	x	O
AMR	NNP	B-NP	X	B-ORG
Corp.	NNP	I-NP	Xx.	I-ORG
,	,	O	,	O
immediately	RB	B-ADVP	x	O
matched	VBD	B-VP	x	O
the	DT	B-NP	x	O
move	NN	I-NP	x	O
,	,	O	,	O
spokesman	NN	B-NP	x	O
Tim	NNP	I-NP	Xx	B-PER
Wagner	NNP	I-NP	Xx	I-PER
said	VBD	B-VP	x	O
.	,	O	.	O

© 2013-2024 NUS. All rights reserved.

Feature-based Sequence Labelling



© 2013-2024 NUS. All rights reserved.

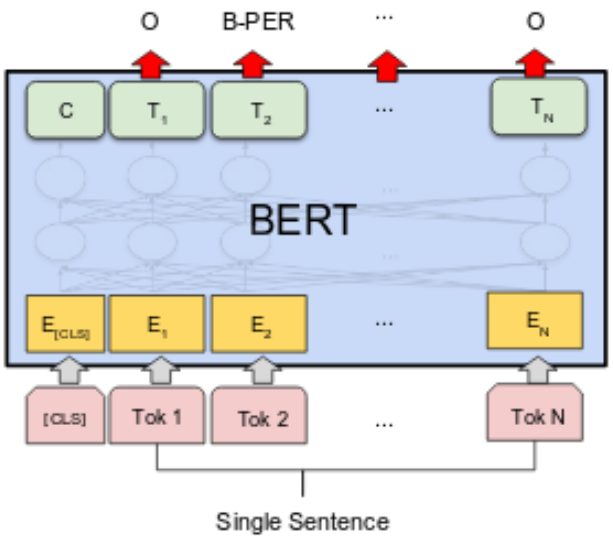
Popular models

- Hidden Markov Models (HMM)
  - Simple, joint probability
- Conditional Random Fields (CRF)
  - Conditional probability
  - Considers features of current token, and of preceding *n* tokens (window=*n*)
- Similarity algorithms
  - Measure distance of group of words to a dictionary list
  - Works especially well for jargon and other terminology
- Support Vector Machines (SVM)
  - Training method for standard perceptron
  - Optimize the points to determine the hyperplane dividing the positive training samples from the negative ones

© 2013-2024 NUS. All rights reserved.

Models based on DNN

- More recent models are based on Deep Neural Networks
  - Bi-LSTMs, Transformers
  - Sequence labeling task - classification of tokens in a sequence (BIO tags)
  - The last layer creates a probability distribution over all NER tags, followed by decoding
  - Good result with transfer learning and finetuning



© 2013-2024 NUS. All rights reserved.

Large Language Models

- Large Language Models (LLMs) like ChatGPT/GPT4, LLaMA, Bard, etc.
  - Instruction tuning enables LLMs to follow instructions, with zero-shot generation
  - *The art of asking* - prompt engineering/design required to induce the expected response from the models

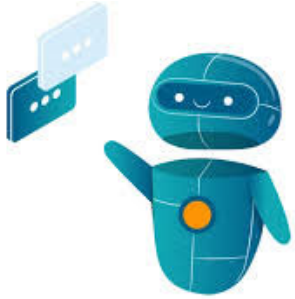
Please identify the named entities in the quoted sentence: "American Airlines, a unit of AMR Corp., immediately matched the move, spokesman Tim Wagner said."

Generating answers for you...

The named entities in the quoted sentence are:

- American Airlines (ORG)
- AMR Corp. (ORG)
- Tim Wagner (PER) [1]

I hope that helps!



© 2013-2024 NUS. All rights reserved.

IE with LLM

- Handles the usual named entities well.

Please identify the named entities in the quoted sentence and list the entities in json format: "The National University of Singapore and The Hebrew University of Jerusalem (HUJ) are launching a Joint Doctor of Philosophy degree programme in biomedical science from August 2013. Professor Tan Eng Chye, NUS Deputy President and Provost, and Professor Menahem Ben-Sasson, President of HUJ signed the joint degree agreement at NUS, in the presence of Ambassador of Israel to Singapore Her Excellency Amira Arnon and about 30 invited guests."

{ "named_entities": [{ "name": "National University of Singapore", "type": "Organization" }, { "name": "The Hebrew University of Jerusalem", "type": "Organization" }, { "name": "HUJ", "type": "Organization" }, { "name": "August 2013", "type": "Date" }, { "name": "Professor Tan Eng Chye", "type": "Person" }, { "name": "NUS", "type": "Organization" }, { "name": "Professor Menahem Ben-Sasson", "type": "Person" }, { "name": "Ambassador of Israel to Singapore Her Excellency Amira Arnon", "type": "Person" }, { "name": "Singapore", "type": "Location" }] }

© 2013-2024 NUS. All rights reserved.

IE with LLM

- Powerful enough to cover other types of information

Please identify the human names an their titles in the quoted sentence and list them in json format: "The National University of Singapore and The Hebrew University of Jerusalem (HUJ) are launching a Joint Doctor of Philosophy degree programme in biomedical science from August 2013. Professor Tan Eng Chye, NUS Deputy President and Provost, and Professor Menahem Ben-Sasson, President of HUJ signed the joint degree agreement at NUS, in the presence of Ambassador of Israel to Singapore Her Excellency Amira Arnon and about 30 invited guests."

{ "people": [{ "name": "Tan Eng Chye", "title": "Professor, NUS Deputy President and Provost" }, { "name": "Menahem Ben-Sasson", "title": "Professor, President of HUJ" }, { "name": "Amira Arnon", "title": "Ambassador of Israel to Singapore, Her Excellency" }] }

Please identify the degree name and its details in the quoted sentence and list them in json format: "The National University of Singapore and The Hebrew University of Jerusalem (HUJ) are launching a Joint Doctor of Philosophy degree programme in biomedical science from August 2013. Professor Tan Eng Chye, NUS Deputy President and Provost, and Professor Menahem Ben-Sasson, President of HUJ signed the joint degree agreement at NUS, in the presence of Ambassador of Israel to Singapore Her Excellency Amira Arnon and about 30 invited guests."

{ "degree": { "name": "Joint Doctor of Philosophy", "field": "Biomedical Science", "start_date": "August 2013", "institutions": [{ "name": "National University of Singapore", "abbreviation": "NUS" }, { "name": "The Hebrew University of Jerusalem", "abbreviation": "HUJ" }] } }

© 2013-2024 NUS. All rights reserved.

Reference & Resources

- Ron Feldman, *Information Extraction: Theory and Practice*, http://cs.fit.edu/~pkc/icdm03/printing/tutorials/extraction/extraction.tutorial.pdf
- Fabio Ciravegna, *Tutorial on Information Extraction from Text*, http://www.isweb.uni-koblenz.de/files/ssms09/SSMS_Slides/ciravegna-IE text.pdf
- Chris Manning & Hinrich Schutze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999
- Jurafsky, Dan, *Speech & language processing*. Pearson Education India, 2000. (continuously updated)
- GA Miller. *WordNet: A Lexical Database for English*, *Communications of the ACM*, 1995
- NLP resources: http://nlp.stanford.edu/links/statnlp.html
- Regular Expression Tutorial: http://www.zytrax.com/tech/web/regex.htm

Type text here

© 2013-2024 NUS. All rights reserved.

TEXT ANALYTICS

MODULE 6: TEXT CATEGORIZATION

Dr. Wang Aobo

Email: isswan@nus.edu.sg

Objectives of this module

At the end of this module, you can:

- **Describe what is text categorization and how text categorization systems work**
- **Evaluate a text categorization system with respect to a business scenario**
- **Understand how supervised and unsupervised text categorization works**
- **Understand what is topic modeling**

PLP Cert

Job&Roles

- Chatbot developer
- NLP specialist /scientist/Engineer
- AI solution engineer
- Machine learning engineer
- Data scientist

Skill Sets Required by Job Scope

- **Chatbot design and implementation**
- **Language modelling** **LLMs pre-training**
  sequence learning models **& fine-tuning**
- **deep-learning libraries** **Transformers**
  Tensorflow/ Keras or Pytorch **BERT / T5 / GPT**
- **Expertise in text-analytics**
  Sentiment Analysis, Entity Extraction,
- **Python-based**
  machine learning libraries
- **Clustering, predictive analysis**
  statistical and probability methods
  NN, XGBoost, Bayesian methods,
  clustering, decision trees, random forests
  Content Extraction and Classification

PLP Cert

- CNI
- TPML
- NMSM
- TA

© 2020 National University of Singapore. All Rights Reserved

Outline for this module

- **What is text categorization?**
- **How does supervised text categorization work?**
  - Document data set
  - Building a classifier
  - Evaluation (*quiz*)
  - Running the classifier (*workshop*)
- **Text categorization application examples**
- **Unsupervised text categorization**
  - Document clustering
- **Topic Modeling**
  - LDA (*workshop*)

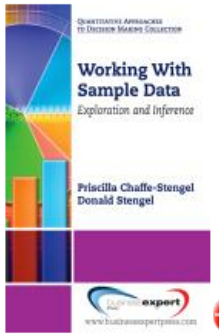
WHAT IS TEXT CATEGORIZATION?

Contrast with a library catalog

- **Example to right**
  - Subject: Statistics
- **Assigned by a cataloger**
- **Slow, tedious**
- **May be inconsistent**

Record 1 of 381 in National Library Board **Search type:** Search by Subjects
Search was: Statistics

Next



Title Working with sample data : exploration and inference / Priscilla Chaffe-Stengel, Donald N. Stengel.
Author Chaffe-Stengel, Priscilla M.
Publisher New York : Business Expert Press, c2012.
Physical Description 151 p. : ill. ; 23 cm.
Notes Includes index.
"The quantitative approaches to decision making collection"--Cover.
Originally published in 2011.
Other Contributors Stengel, Donald N.
Search by Subjects Commercial statistics.
Statistics.

MESH index of a single journal paper

Below is an example of a **complete reference** in Medline (OvidSP) showing the journal article details and the list of MeSH headings (some with subheadings) assigned to it by the NLM Indexers:

Unique Identifier	20980007
Record Owner	From MEDLINE, a database of the U.S. National Library of Medicine.
Status	MEDLINE
Authors	Asejczyk-Widlicka M. Srodka W. Schachar RA. Pierscionek BK.
Authors Full Name	Asejczyk-Widlicka, M. Srodka, W. Schachar, R A. Pierscionek, B K.
Institution	Institute of Physics, Wroclaw University of Technology, Wybrzeze Wyspianskiego 27, 50-370 Wrocl Poland.
Title	Material properties of the cornea and sclera: a modelling approach to test experimental analysis
Source	Journal of Biomechanics. 44(3):543-6, 2011 Feb 3.
Abbreviated Source	J Biomech. 44(3):543-6, 2011 Feb 3.
NLM Journal Name	Journal of biomechanics
Publishing Model	Journal available in: Print-Electronic Citation processed from: Internet
NLM Journal Code	0157375, hjf
Country of Publication	United States
MeSH Subject Headings	Computer Simulation *Cornea / ph [Physiology] Finite Element Analysis Humans *Intraocular Pressure / ph [Physiology] Muscle Rigidity *Sclera / ph [Physiology] Visual Acuity / ph [Physiology]
Abstract	Material prope es of cornea and sclera are important for maintaining the shape of the eye and the requisite ce curvatures for optics. They also need to withstand the forces of external and inte sculature and fluctuations in intraocular pressure (IOP). These properties are difficult tc iable results have been reported. A previously published experimental procedure, material properties of the eyeball coats were obtained, has been modelled in this te Element Analysis, in order to test the accuracy of the experiment. Material e calculated from the model and the resulting relationships between stress and rnea and sclera compared to their experimentally obtained counterparts. The ween model and experiment was close for the sclera but more varied for the cornea The pressure vessel model can be applied for measuring the material properties of the sclera but is less accurate for the cornea. Copyright Copyright 2010 Elsevier Ltd. All rights reserved.

Journal article details

MeSH descriptors assigned by indexers- taken from the list of preferred terms used to describe topics

The indexers have assigned the subheading Physiology to this MeSH descriptor

From: http://library.ulster.ac.uk/sci/mesh/

Automatic text categorization (also known as "classification")

- **Hard Classification**

  The process of assigning text documents uniquely into two or more categories (a document cannot be in more than one category)

  E.g., spam filtering – binary decision: "spam" or "not spam"

- **Soft Classification**

  The process of assigning one or more category labels to a text document (a document may have more than one category)

  E.g., news filtering – which category to assign to news articles:

  - Sports, Olympics, Football (natural class)
  - Political, Business, Home,… (news sections)
  - Asian, Europe, Middle-East, …(geographical)

Some Examples of Text Classification

- Assigning subject categories to documents
- Email spam detection
- Medical diagnosis
- Identifying a language (before further processing)
- Identifying fraud (anomaly detection)
- Sentiment analysis (e.g., positive/negative reviews)
- Monitoring a news feed (e.g., news about 737-MAX 8)
- Etc.

Is this document produced by humans or by chatgpt (generative Ai)

HOW DOES AUTOMATIC TEXT CATEGORIZATION WORK?

Text Categorization Phases

Two Phases for supervised method

1. **Training – creating the text "classifier" (automatic categorization engine)**
   - You need a set of documents, already categorized
   - Divide the set into training (typically 70%) and testing (30%)
   - Train your classifier such that it's able to accurately classify the training set of documents to your level of comfort
     - "level of comfort" depends on how hard is the task! ☺
   - Evaluate your classifier on the test set; ensure sufficient accuracy
2. **Running – using your classifier on new sets of documents**
   - You will not know how well it performs
   - Need to "audit" the results occasionally (use an assessor)
     - Assess random sample of the documents against the predicted categories

DOCUMENT DATA SET

Movie reviews classified as "Human" and "Robot" generated

- a mesmerizing cinematic poem from the first frame to the last ."
- a well-put-together piece of urban satire .
- one can't deny its seriousness and quality .
- hard to resist .
- a naturally funny film , home movie makes you crave chris smith's next movie .
- a true-blue delight .
- a fun ride .
- a surprisingly funny movie .
- the script is smart and dark - hallelujah for small favors .
- a flick about our infantilized culture that isn't entirely infantile .

- unfortunately the story and the actors are served with a hack script .
- too slow for a younger crowd , too shallow for an older one .
- terminally brain dead production .
- one lousy movie .
- this movie . . . doesn't deserve the energy it takes to describe how bad it is .
- a cleverly crafted but ultimately hollow mockumentary .
- it's an 88-minute highlight reel that's 86 minutes too long .
- the whole affair is as predictable as can be .

From: http://karpathy.ca/mlsite/lecture2.php

Movie reviews classified as "Human" and "Robot" generated

-a mesmerizing cinematic poem from the first frame to the last ."
-a well-put-together piece of urban satire .
-one can't deny its seriousness and quality .
-hard to resist .
-a naturally funny film , home movie makes you crave chris smith's next movie .
-a true-blue delight .
-a fun ride .
-a surprisingly funny movie .
-the script is smart and dark - hallelujah for small favors .
-a flick about our infantilized culture that isn't entirely infantile .

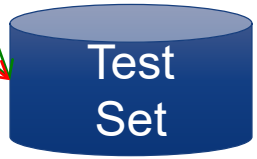
5000 reviews

-unfortunately the story and the actors are served with a hack script .
-too slow for a younger crowd , too shallow for an older one .
-terminally brain dead production .
-one lousy movie .
-this movie . . . doesn't deserve the energy it takes to describe how bad it is .
-a cleverly crafted but ultimately hollow mockumentary .
-it's an 88-minute highlight reel that's 86 minutes too long .
-the whole affair is as predictable as can be .

5000 reviews

70%

30%



From: http://karpathy.ca/mlsite/lecture2.php

BUILDING A CLASSIFIER

JUST SOME EXAMPLES (NOT EXHAUSTIVE)

Creating classifiers

- **Hand-coded classifiers (the "good old days!")**
  - If <conditions> then <category> else NOT<category>, where conditions are normally in disjunctive normal form

If	((wheat & farm)	**or**	
	(wheat & commodity)	**or**	
	(bushels & export)	**or**	
	(wheat & tonnes)	**or**	
	(wheat & winter & ¬soft))	**then**	WHEAT else ¬ WHEAT

From: F. Aiolli, *Text Categorization*, http://www.math.unipd.it/~aiolli/corsi/SI-0607/Lez09.251006.pdf

Generative Classifiers

- **Naïve Bayes Model**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as observations.

- Q: What is A and B in documents Classification context?

Naïve Bayes Model

Probabilistic Classifiers

Represent the probability that a document d_i belongs to category c_j by

$$P(c_j|d_i) = P(c_j)P(d_i|c_j) \,/\, \cancel{P(d_i)}$$

$$\begin{aligned} class_{MAP} &\approx \boldsymbol{argmax}_{class \in C}\, P(doc|class) * \mathrm{P}(class) \\ &\approx \boldsymbol{argmax}_{class \in C}\, P(w_1, w_2, \ldots w_n|class) * \mathrm{P}(class) \\ &\approx \boldsymbol{argmax}_{class \in C}\, P(w_1|class) * P(w_2|class) * \cdots * P(w_n|class) * \mathrm{P}(class) \end{aligned}$$

Naïve Bayes Model

$$class_{MAP} \approx \boldsymbol{argmax}_{class \in C}\, P(doc|class) * \mathrm{P}(class)$$
$$\approx \boldsymbol{argmax}_{class \in C}\, P(w_1, w_2, \dots w_n|class) * \mathrm{P}(class)$$
$$\approx \boldsymbol{argmax}_{class \in C}\, P(w_1|class) * P(w_2|class) * \cdots * P(w_n|class) * \mathrm{P}(class)$$

- **Example**
  - For sentiment detection problem: $\boldsymbol{class} \in \{+, -\}$
  - **Training**: calculate and keep all the likelihood of vocabulary words *wrt.* classes: $P(w|-)$, $P(w|+)$, $\mathrm{P}(+)$, $\mathrm{P}(-)$

    $P(w|-)$ = *count (neg docs having w) / count (neg docs)*

    $P(-)$ = ***count (neg doc) / count(total docs)***

  - **Testing**: calculate, compare and select the class offering bigger result

$$P(w_1|-) * P(w_2|-) * \cdots * P(w_n|-) * \mathrm{P}(-)$$
$$P(w_1|+) * P(w_2|+) * \cdots * P(w_n|+) * \mathrm{P}(+)$$

Naïve Bayes Model

$$class_{MAP} \approx \boldsymbol{argmax}_{class \in C}\, P(doc|class) * \mathrm{P}(class)$$
$$\approx \boldsymbol{argmax}_{class \in C}\, P(w_1, w_2, \dots w_n|class) * \mathrm{P}(class)$$
$$\approx \boldsymbol{argmax}_{class \in C}\, P(w_1|class) * P(w_2|class) * \cdots * P(w_n|class) * \mathrm{P}(class)$$

- **Example**
  - For sentiment detection problem: $\boldsymbol{class} \in \{+, -\}$
  - **Training**: calculate and keep all the likelihood of vocabulary words *wrt.* classes: $P(w|-)$, $P(w|+)$, $\mathrm{P}(+)$, $\mathrm{P}(-)$
  - **Testing**: calculate, compare and select the class offering bigger result

$$P(w_1|-) * P(w_2|-) * \cdots * P(w_n|-) * \mathrm{P}(-)$$
$$P(w_1|+) * P(w_2|+) * \cdots * P(w_n|+) * \mathrm{P}(+)$$

$$Dnew = "I\ hated\ the\ poor\ acting"$$

$$P(+|D_{new}) = P(I|+) * P(hated|+) * \cancel{P(the|+)} * P(poor|+) * P(acting|+) * P(+)$$
$$= \mathrm{a}$$

$$P(-|D_{new}) = P(I|-) * P(hated|-) * \cancel{P(the|-)} * P(poor|-) * P(acting|-) * P(-)$$
$$= \mathrm{b}$$

Discriminative Classifiers

- **Decision Tree Classifiers**

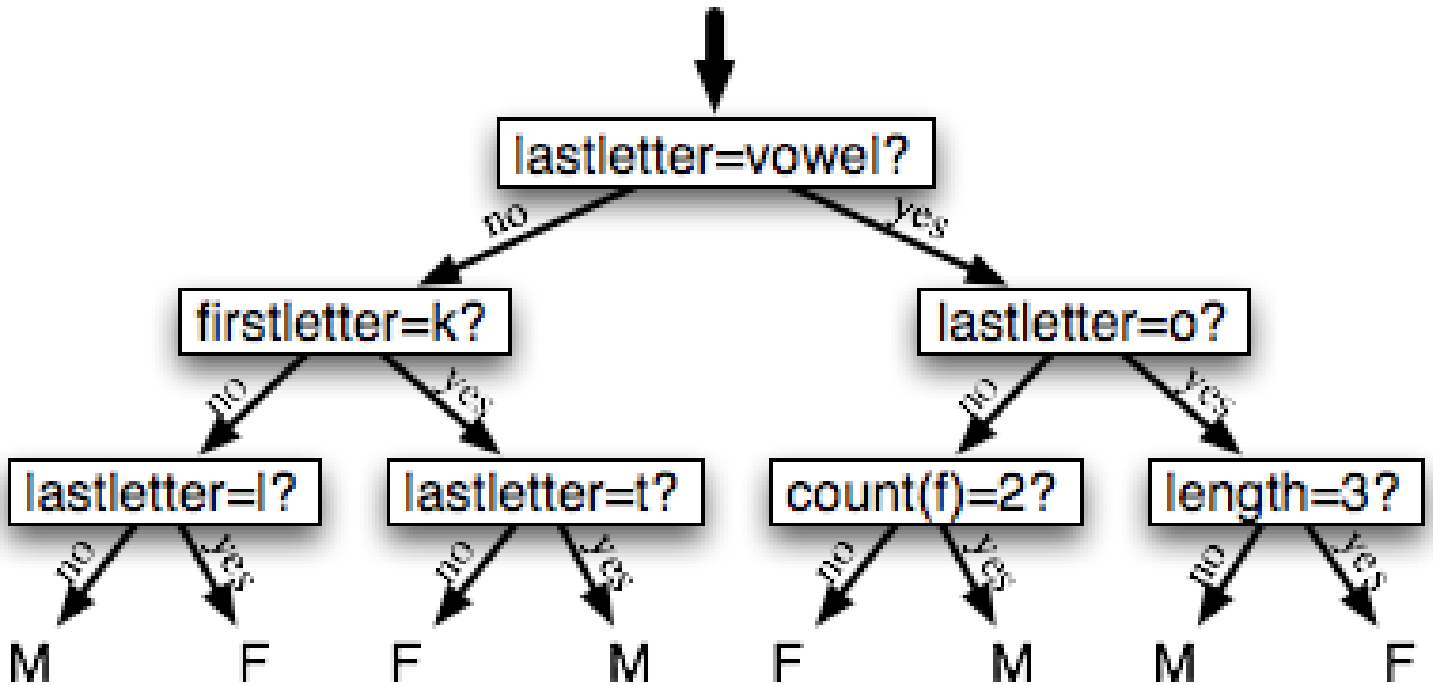
- List of boys names
  - Alan
  - Barry
  - Colin
  - Dexter
  - Edward
  - Frederick
  - Howard
  - ….

- List of girls names
  - Anna
  - Betty
  - Chelsea
  - Doris
  - Elizabeth
  - Fanny
  - Hortense
  - ….

Q: What is inside the leaf node?

Q: What is inside the internal node?

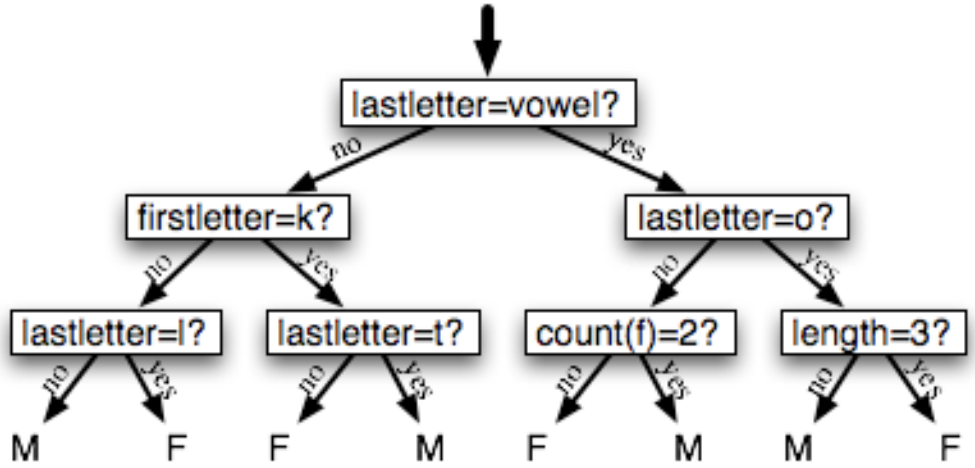
Example of a decision tree to decide if a name is male or female



From: http://nltk.googlecode.com/svn/trunk/doc/book/ch06.html

Example of a decision tree to decide if a name is male or female

	Lastletter ="vowel"	Firstletter= "k"	Lastletter ="t"	Lastletter ="o"	Count(f)	length
Fanny	0	0	0	0	1	5
Kate	0	0	0	0	0	4
Howard	0	0	0	0	0	6

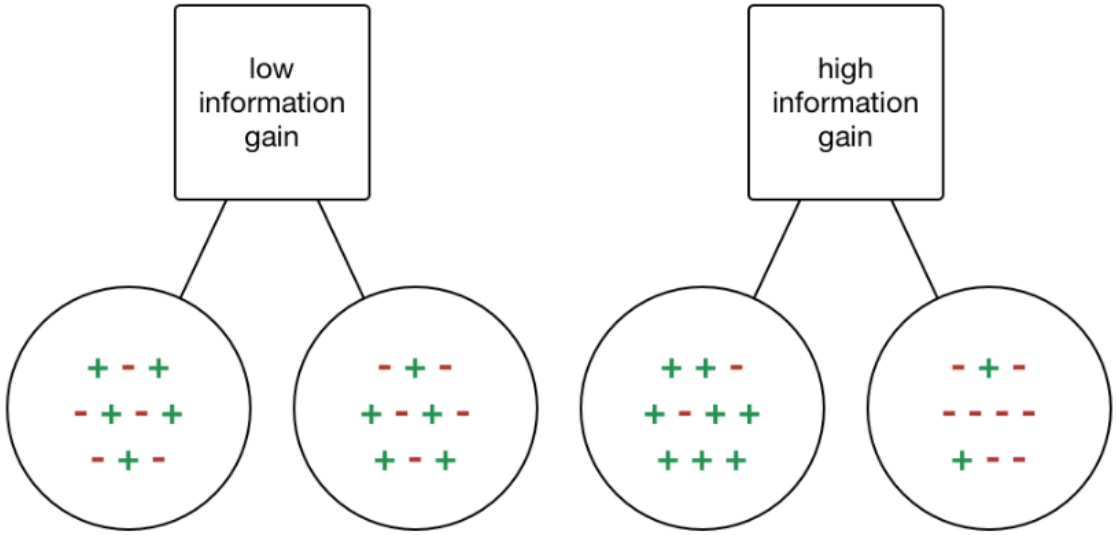


Information Gain

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N} I(D_{left}) - \frac{N_{right}}{N} I(D_{right})$$

f: feature split on
D_p: dataset of the parent node
D_{left}: dataset of the left child node
D_{right}: dataset of the right child node
I: impurity criterion (Gini Index or Entropy)
N: total number of samples
N_{left}: number of samples at left child node
N_{right}: number of samples at right child node

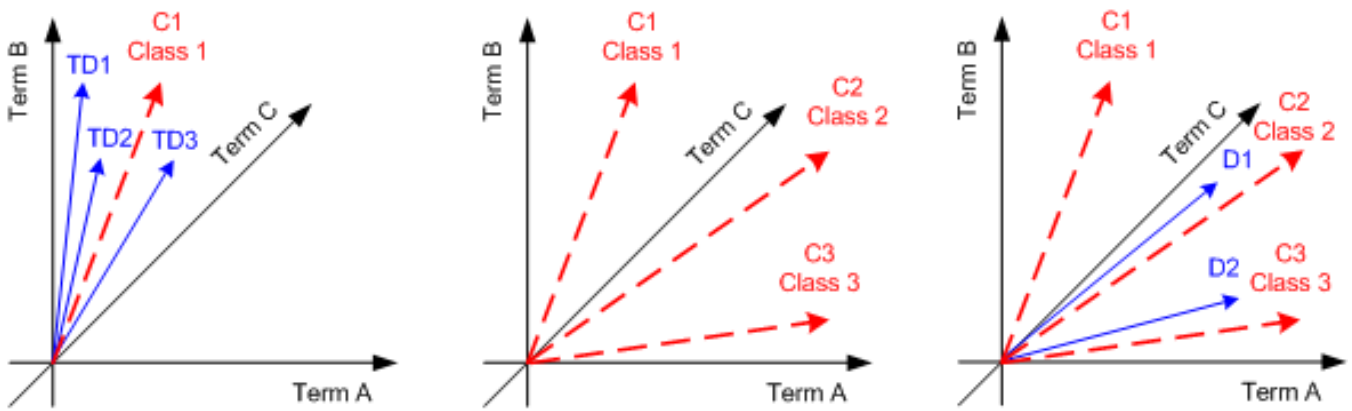
$$\mathrm{H}(T) = \mathrm{I}_E(p_1, p_2, \ldots, p_J) = -\sum_{i=1}^{J} p_i \log_2 p_i$$



© 2019 National University of Singapore. All Rights Reserved

Discriminative Classifiers

- **The Rocchio Classifiers**
  - Each category is represented by a prototypical document, i.e., profile vector
  - Documents are classified by similarity to the profile vector



A) Training Document Representations in Vector Space define Class Representation (Centroid)

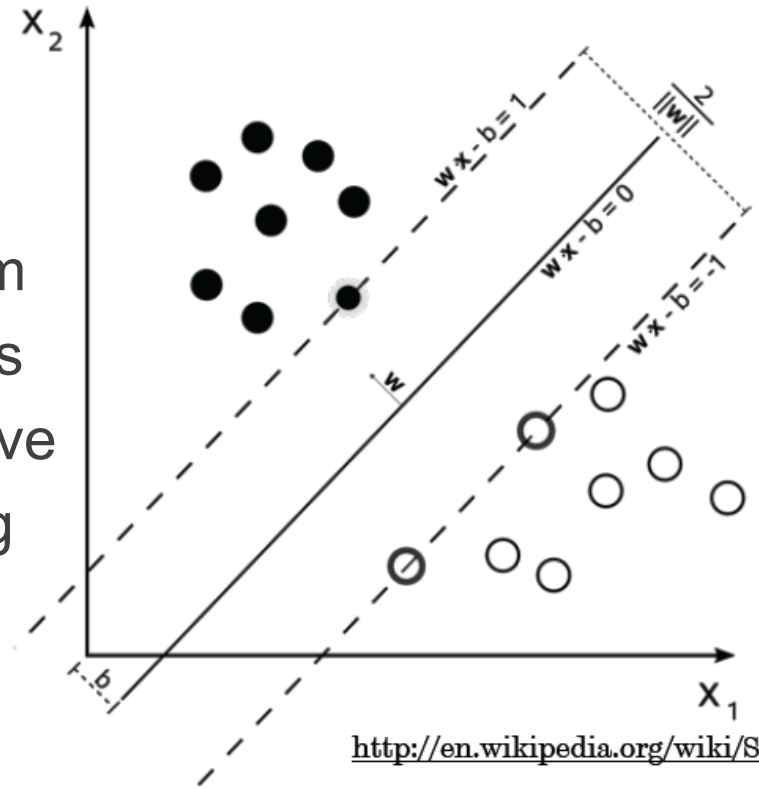
B) Class Representation by Class Vectors (CentroidS)

C) Classification of Documents by similarity between Document Vector and Class Vector

From: http://www.iicm.tugraz.at/about/Homepages/cguetl/courses/isr/opt/classification/Vector_Space_Model.html

Discriminative Classifiers

- **Support Vector Machines (SVMs)**
  - SVMs divide the term space in hyperplanes separating the positive and negative training samples.
  - The surface that provides the widest separation between the support surfaces is selected



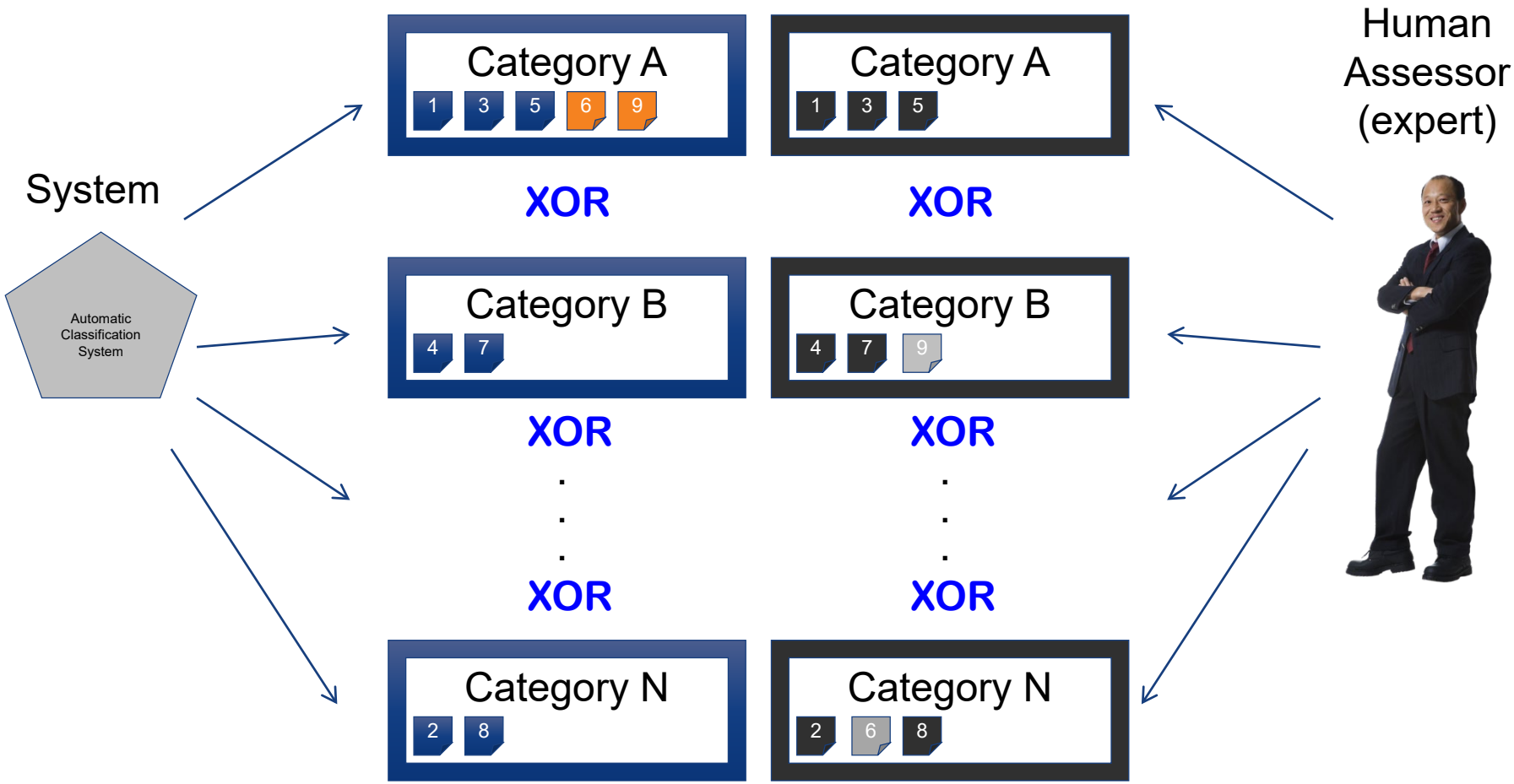
http://en.wikipedia.org/wiki/Support_vector_machines

EVALUATION

ACCURACY

EVALUATION

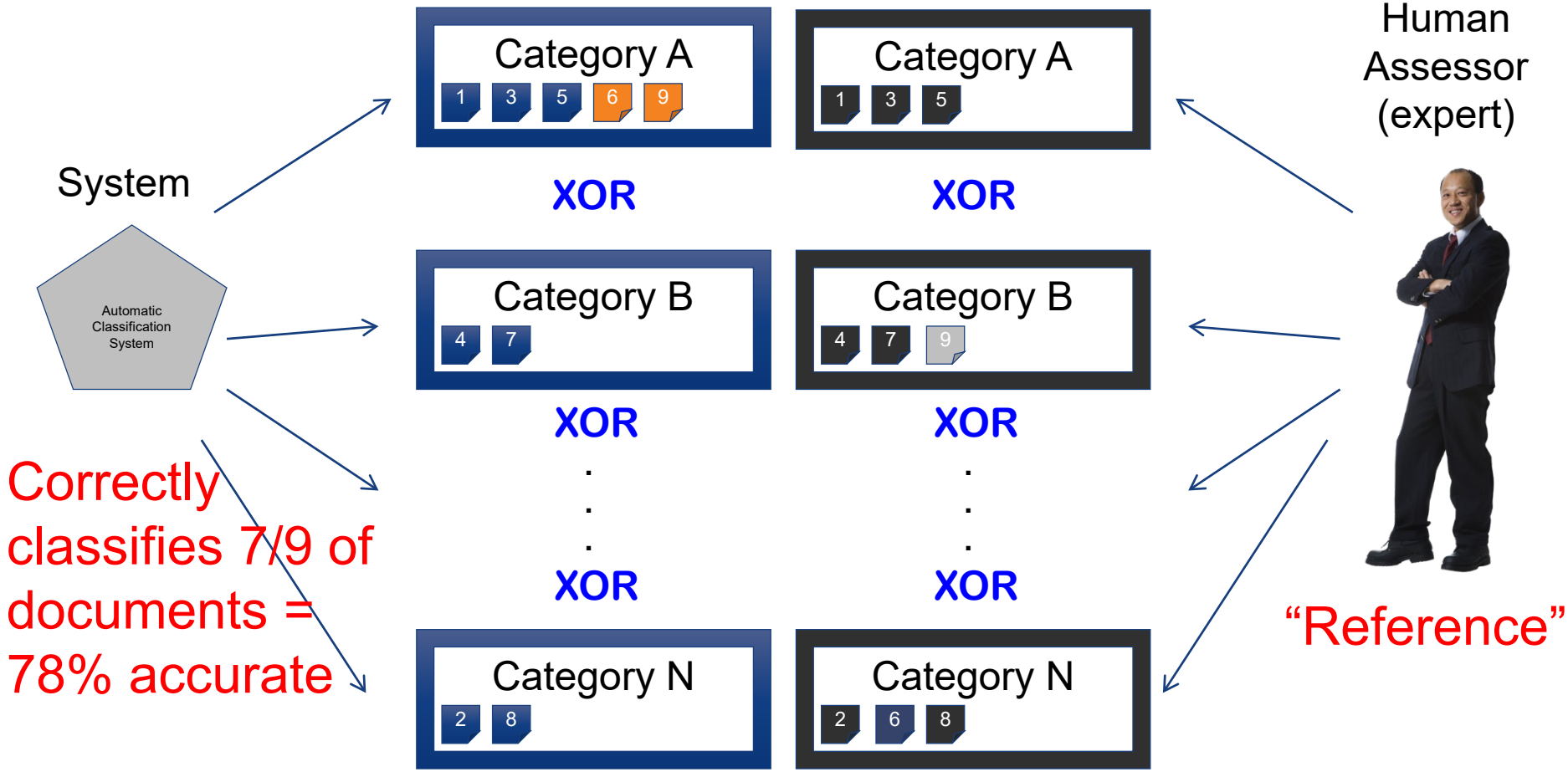
What is “accuracy”?



What do we mean by "Accuracy"

- **You measure an automatic categorization system by:**
  - How well it classifies a set of documents against a "reference"
  - This "reference" is normally a human expert
- **Reference**
  - Gold standard – accepted as being the best available
    - May not be perfect, e.g., tumour board for oncology
  - Good enough
    - Human expert(s), typically 80% agreement, good methodology
  - Better than nothing
    - "your boss tells you to do this, so you recruit your friends, family,…"
- **Most of the time, no such thing as "absolute truth"**

One measure of accuracy



Discussion

- **Weather prediction (system predicts one week in advance)**

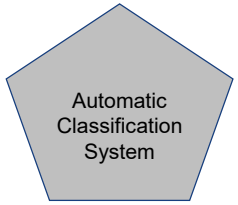
	Mon	Tue	Wed	Thu	Fri	Sat	Sun
System	Sunny	Drizzle	Rain	Sunny	Cloudy	Thunderstorms	Sunny
Actual	Sunny	Rain	Cloudy	Sunny	Drizzle	Thunderstorms	Cloudy

- **Questions:**
  - How “accurate” is the weather prediction system?
  - Do you have a tolerance for error? +/- margin of error?
  - How about outcomes – will I get wet if I use the system to decide whether or not to carry an umbrella? Will I get angry?

Confusion Matrix

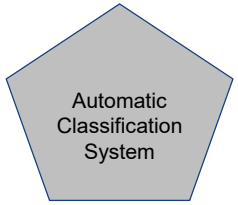
		Predicted Categories					
		A	B	C		. . .	N
Actual Categories	A						
	B						
	C						
	. . .						
	N						

Example (using #)



		Predicted Categories						
Actual Categories		A	B	C		. . .	N	
	A	143	34	17		. . .	2	= Tot(A) docs
	B	67	1289	44		. . .	239	= Tot(B) docs
	C	980	234	3454		. . .	88	= Tot(C) docs
						. . .		
	⋮					. . .		
	N	87	24	63		. . .	650	= Tot(N) docs

Example (using %)



		Predicted Categories						
Actual Categories		A	B	C		. . .	N	
	A	87%	2%	5%		. . .	1%	= 100%
	B	6%	90%	0%		. . .	2%	= 100%
	C	12%	2%	77%		. . .	4%	= 100%
						. . .		
	⋮					. . .		
	N	21%	0%	4%		. . .	65%	= 100%

Consider the simple 2x2 matrix (2000 documents were classified)

Desired positive prediction

False negative

		Predicted	
		Yes	No
Actual	Yes	1350 90%	150 10%
	No	100 20%	400 80%

False positive

Desired negative prediction

EVALUATING MULTIPLE CLASSIFIERS

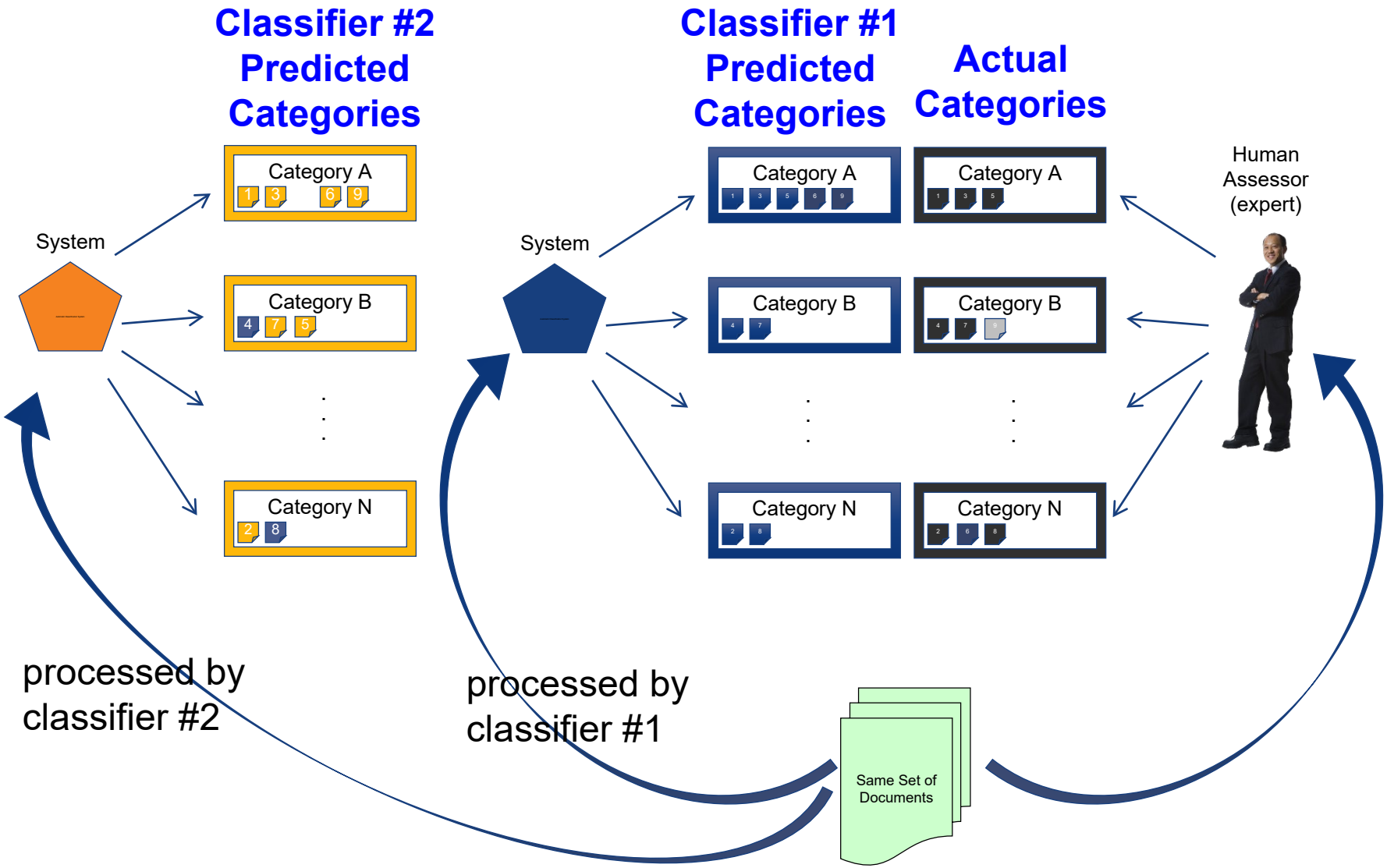
Discussion

- **Weather prediction. You ask two people to predict whether it will rain or not in the coming week:**

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
ZZ	Yes	No	Yes	No	Yes	No	No
MK	No	No	No	Yes	No	No	Yes
Actual	Yes	Yes	Yes	No	Yes	Yes	No

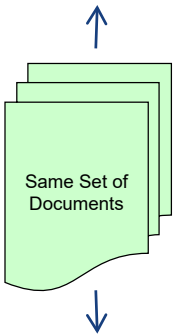
- **Questions:**
  - Who is more “accurate”? ZZ is right 5/7 times. MK is right 0/7 times.
  - Who should you ask in future if you don’t want to get wet?
  - Who is the better “classifier”?

What happens with 2 classifiers?



Comparing models: e.g. 2x2 matrix (actual numbers of documents)

Classifier #1



Classifier #2

		Predicted	
		Y	N
Actual	Y	900	100
	N	40	410

Seems quite good for both predictions

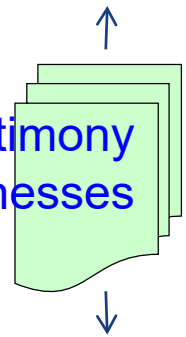
		Predicted	
		Y	N
Actual	Y	700	300
	N	2	448

Reduced the false positives but false negatives increased

? Which classifier is better?

Adding the semantics – the courtroom

Classifier #1



Classifier #2

		Predicted	
Actual		Guilty	Innocent
	Guilty	900	100
	Innocent	40	410

Let 100 guilty go free
Convict 40 innocent persons

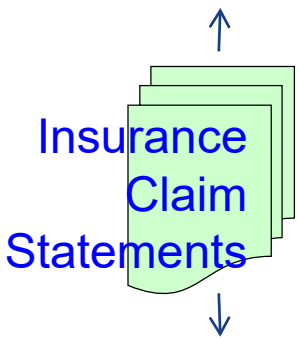
		Predicted	
Actual		Guilty	Innocent
	Guilty	700	300
	Innocent	2	448

Let 300 guilty go free
Convict 2 innocent persons

? Which classifier is better?

Adding a cost function – fraud investigation

Classifier #1



Classifier #2

		Predicted	
Actual		Honest	Fraud
	Honest	900	100
	Fraud	40	410

		Predicted	
Actual		Honest	Fraud
	Honest	700	300
	Fraud	2	448

? Which classifier is better?

The average fraud costs the company $2000
It costs the company $500 to investigate each suspected fraud

Adding a cost function – fraud investigation

- The average fraud costs the company $2000
- It costs the company $500 to investigate each suspected fraud

Company loses $80k in fraud
Company pays $255k in costs

		Predicted	
Actual		Honest	Fraud
	Honest	900	100
	Fraud	40	410

Company loses $4k in fraud
Company pays $374k in costs

		Predicted	
Actual		Honest	Fraud
	Honest	700	300
	Fraud	2	448

- ***Consider Doing nothing (don't act to identify fraud):***
  - *Predicted fraud = 0 cases @$500 per case costs $0k for investigation.*
  - *Undetected fraud is 450 cases @$2k/fraud loses $900k.*
  - *Overall -$0k -$900k = -$900k*
- ***Analysis for classifier #1:***
  - *Predicted fraud = 510 cases @$500 per case costs $255k for investigation.*
  - *Undetected fraud is 40 cases @$2k/fraud loses $80k.*
  - *Overall -$255k -$80k = -$335k*
- ***Analysis for classifier #2:***
  - *Predicted fraud = 748 cases @$500 per case costs $374k for investigation.*
  - *Undetected fraud is 2 cases @$2k/fraud loses $4k.*
  - *Overall -$374k -$4k = -$378k*

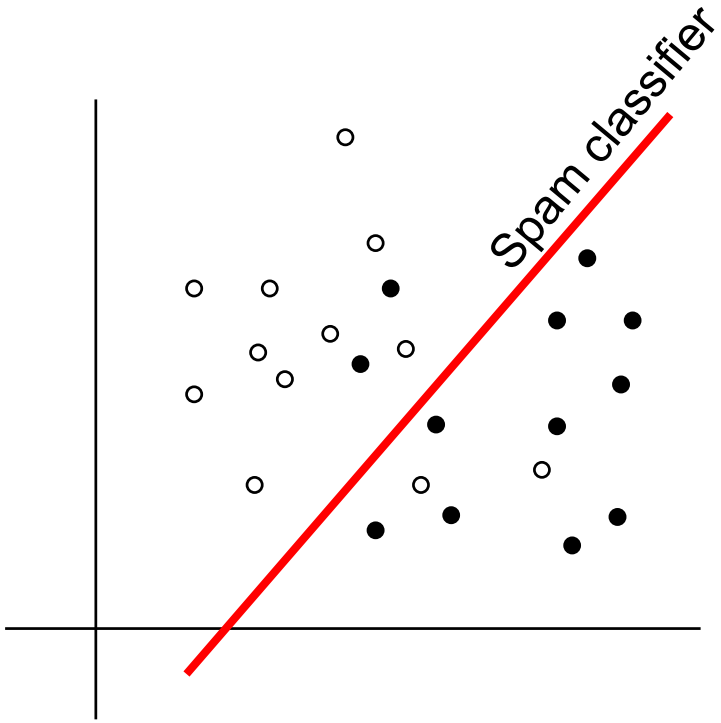
Classifier evaluation

- **Evaluation of classifiers is done with respect to a *business context***
- **Evaluation of classifiers is normally done *empirically***
- **Experimental evaluation focuses on *effectiveness*, i.e., the ability of the classifier to make the right classification decision**
- **Precision & Recall concepts as applied to (multi-class) categorization**
  - Precision is the probability that if a random document d_i is categorized under category c_j, that decision is correct
  - Recall wrt c_j is the probability that if a random document d_i should be categorized under c_j , then the decision is taken

RUNNING THE CLASSIFIER

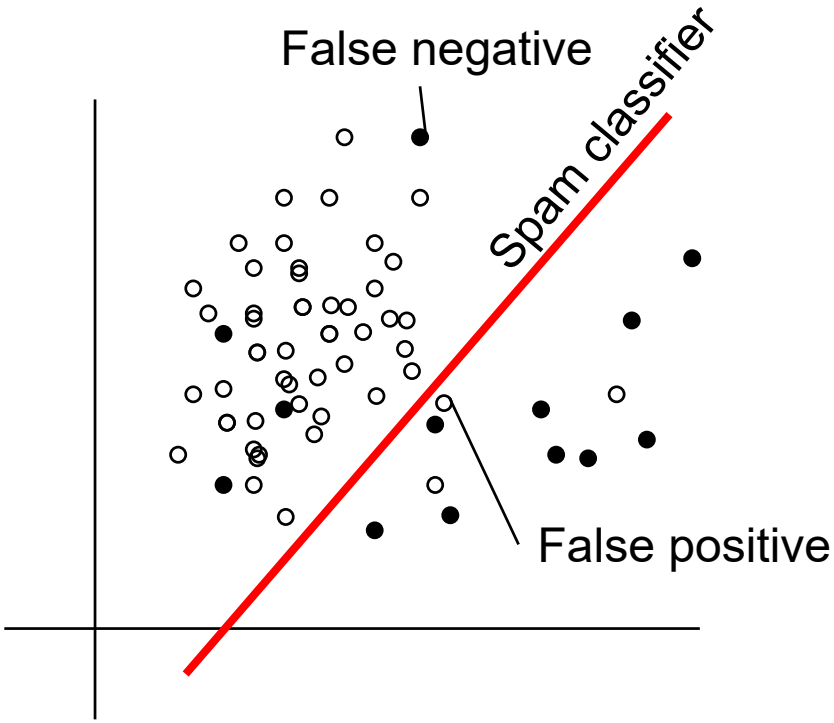
© 2015 National University of Singapore. All Rights Reserved

Expect False Results
eg: spam filtering



- ○ Email data – non-spam
- ● Email data – spam

Training Set



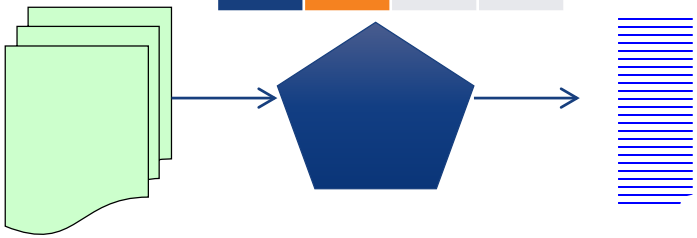
- ○ Email non-spam
- ● Email spam

Real email stream

Overfitting the Training Set

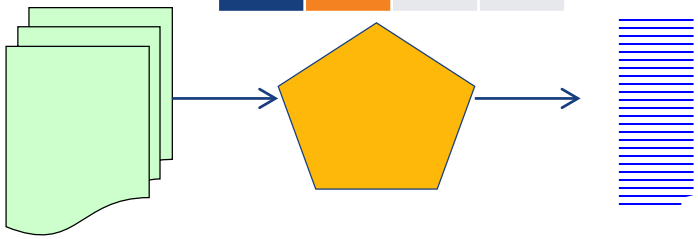
Training Set

		Predicted	
		Y	N
Actual	Y	700	0
	N	0	300



Training Set

		Predicted	
		Y	N
Actual	Y	600	100
	N	50	250

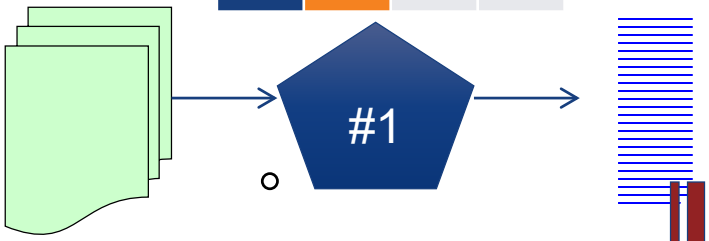


? Which classifier is better?

Overfitting the Training Set

Training Set

Actual	Predicted: Y	Predicted: N
Y	700	0
N	0	300



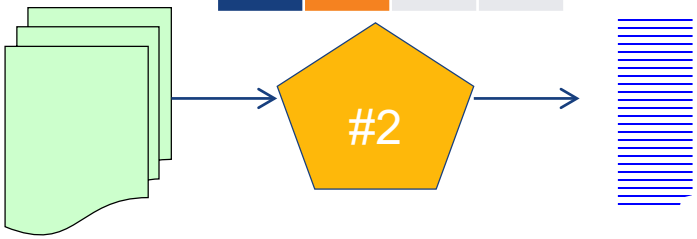
Real Data

Actual	Predicted: Y	Predicted: N
Y	45200	24800
N	13800	16200

REAL WORLD

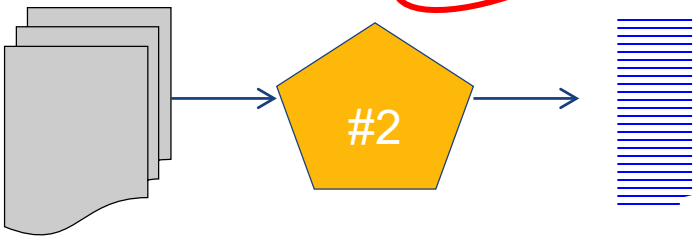
Training Set

Actual	Predicted: Y	Predicted: N
Y	600	100
N	50	250

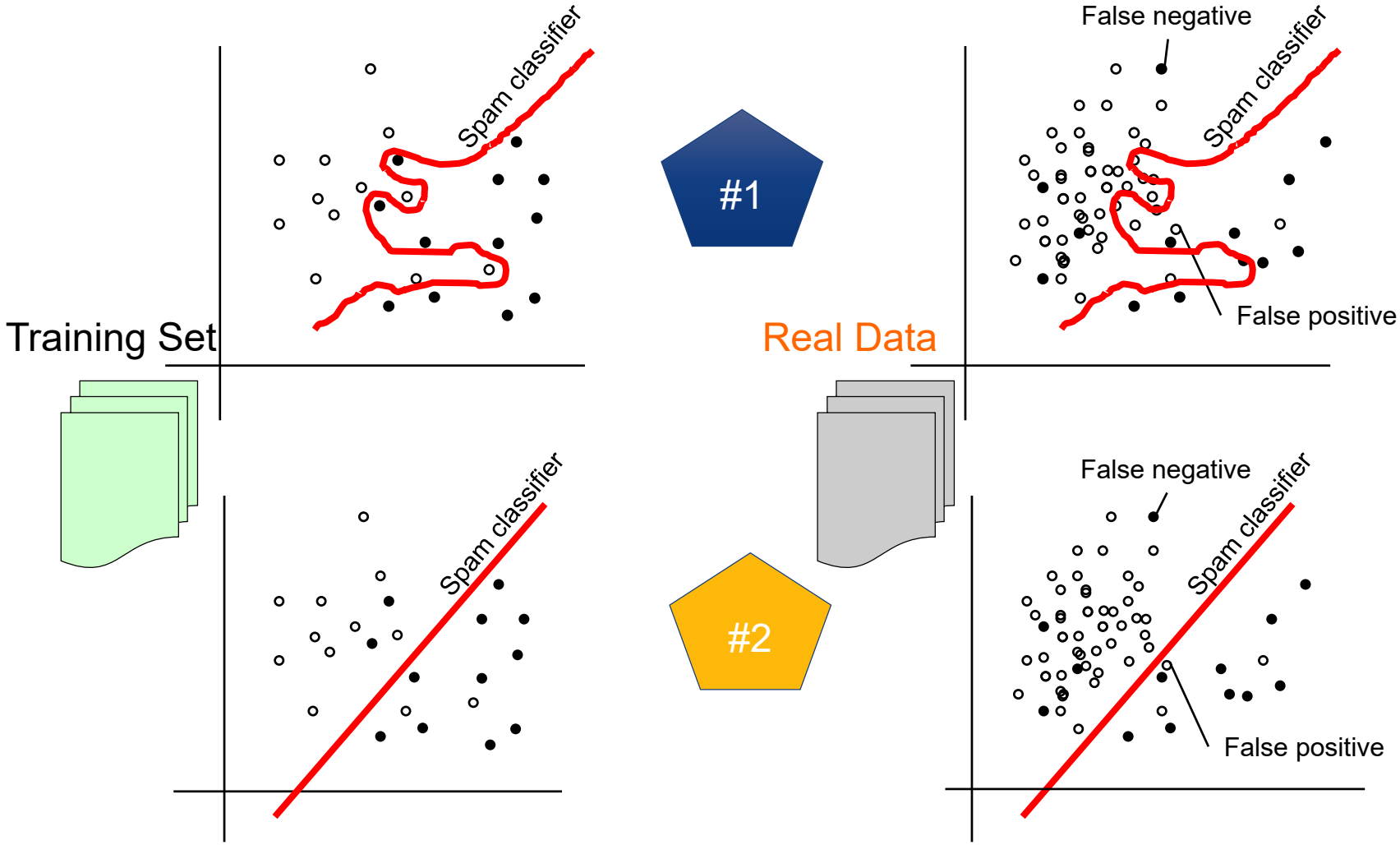


Real Data

Actual	Predicted: Y	Predicted: N
Y	56900	13100
N	1500	28500

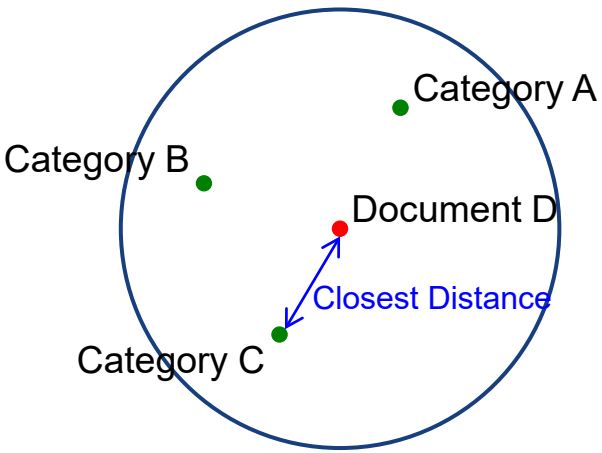


Overfitting the Training Set – what happened?



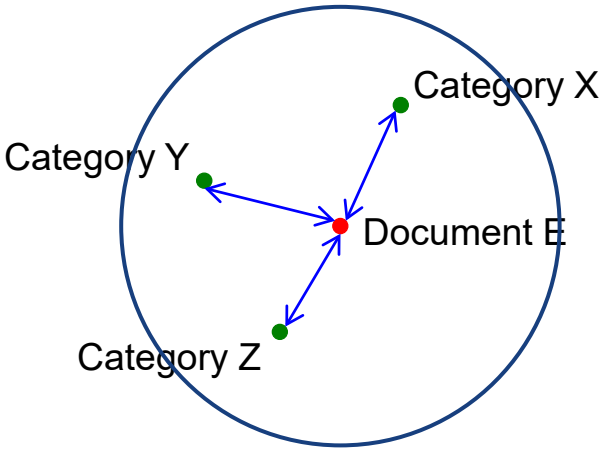
Hard and Soft Categorization

- **Fully automated classifiers make "hard" binary decisions**
  - In example to right, the document, D, is assigned to category C only.



- **Semi-automated (interactive) classifiers instead are created by allowing "soft" real-value decisions**
  - Rank the categories according to their measure of appropriateness for the document
  - In example to right, the document, E, is assigned 3 possible categories:

Rank	Category	Probability
1	Z	0.76
2	X	0.72
3	Y	0.54



  - Used for computer assisted human decision making
    - For example, in critical applications such as medical diagnosis

Aggregating multiple classifiers

Running with more than one classifier

Actual \ Predicted	Y	N
Y	900	100
N	40	410

Classifier #1

Same Set of Documents

Classifier #2

Actual \ Predicted	Y	N
Y	700	300
N	2	448

Lists of predicted documents

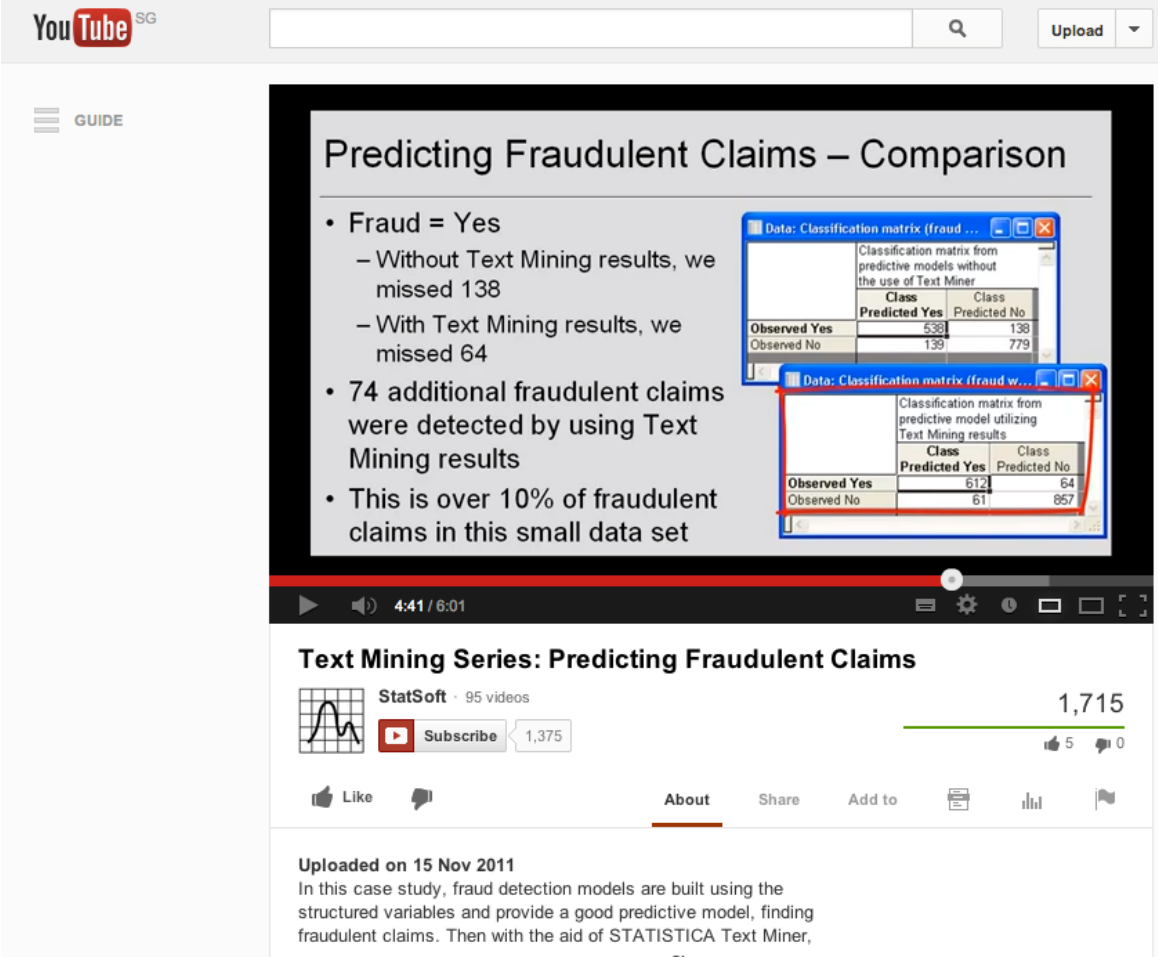
Aggregate the results

Various methods:
- Union
- Intersection
- Algorithmic
  - Voting
  - By confidence level
  - Etc.

Actual \ Predicted	Y	N
Y	880	120
N	20	430

TEXT CATEGORIZATION APPLICATION EXAMPLES

Boosting Identification of Fraudulent Claims



From: http://www.youtube.com/watch?v=OlQpm8qTog4

UNSUPERVISED TEXT CATEGORIZATION

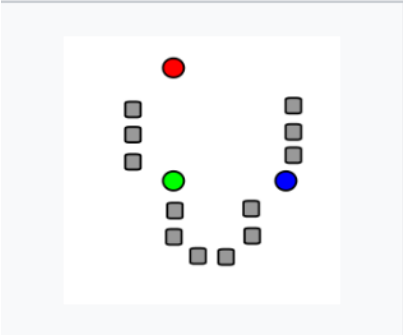
DOCUMENT CLUSTERING

What is text clustering?

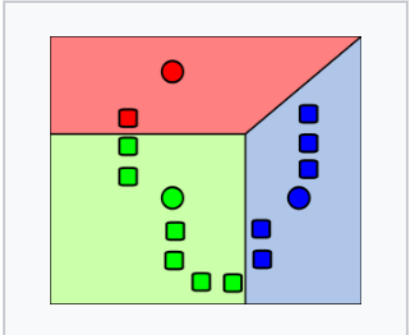
- **Clustering is the task of grouping a set of documents in such a way that the documents in each group are more "similar" to each other than to documents in other groups.**
- **Clustering lets you explore your data**
  - Many tools are interactive
- **You can understand your data better, e.g.:**
  - What groupings exist in your data?
  - How many are there? How big is each group?
  - What are the common terms?
  - Are there anomalies?

Clustering Example

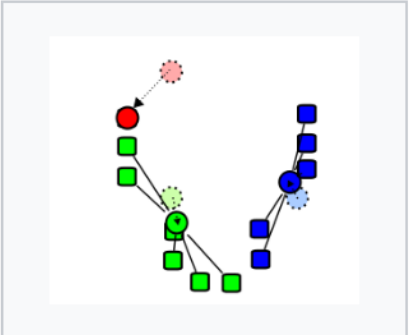
Demonstration of the standard algorithm



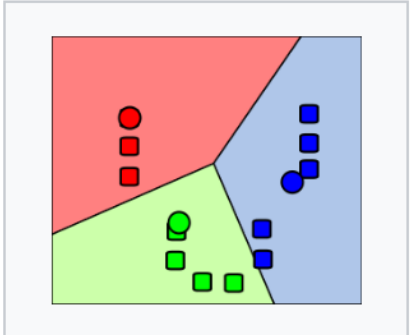
1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).



2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



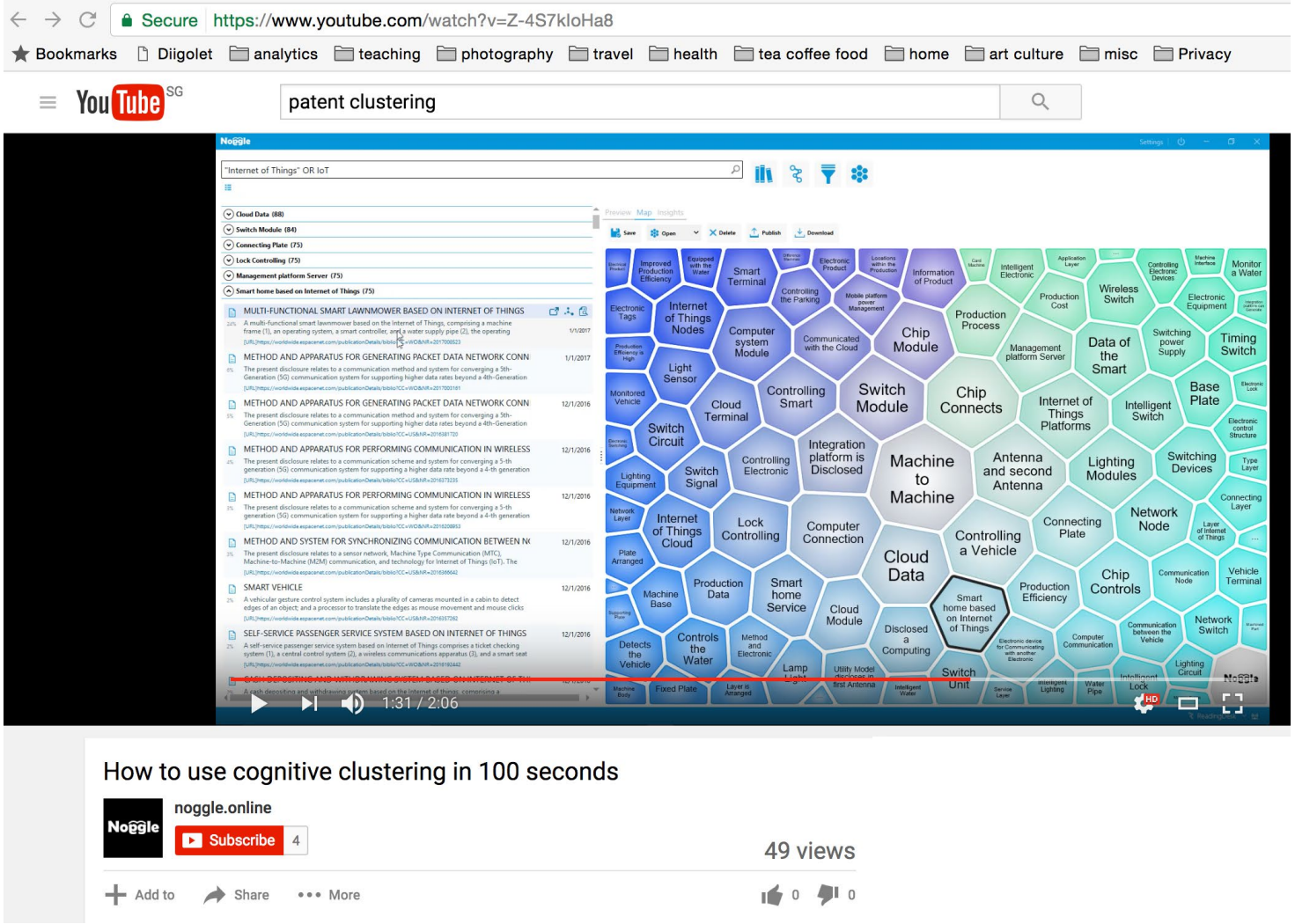
3. The centroid of each of the *k* clusters becomes the new mean.



4. Steps 2 and 3 are repeated until convergence has been reached.

From: https://www.youtube.com/watch?v=CHlrx4gsoJI

Patent Clustering



From: https://www.youtube.com/watch?v=Z-4S7kIoHa8

Notes about Clustering

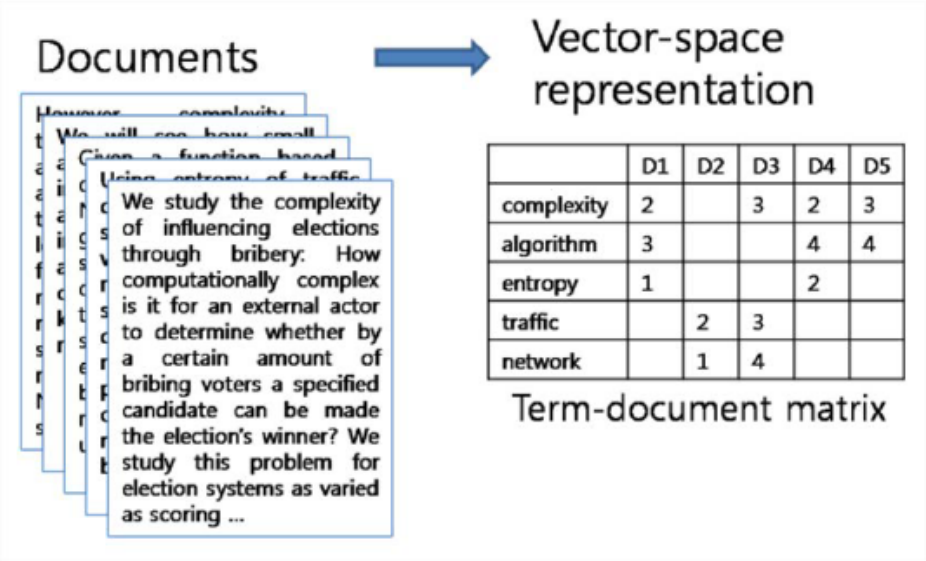
- **Your clusters may surprise you**
  - Documents tend to fall into natural classes (clusters)
  - There will be some surprising ones (worth drilling down!)
- **You can control the number of clusters (depends on the algorithm)**
  - You don't want too many clusters (overfit!)
  - You don't want too few clusters (meaningless)
  - **Clusters should lead to fulfilling business outcomes**
- **You don't need training phase to create clusters**
  - Clustering can be language independent (but monolingual)

DIMENSIONAL REDUCTION

Dimensional Reduction

- Sparsity
- High dimension
- SVD
- Low dimension

Document Clustering

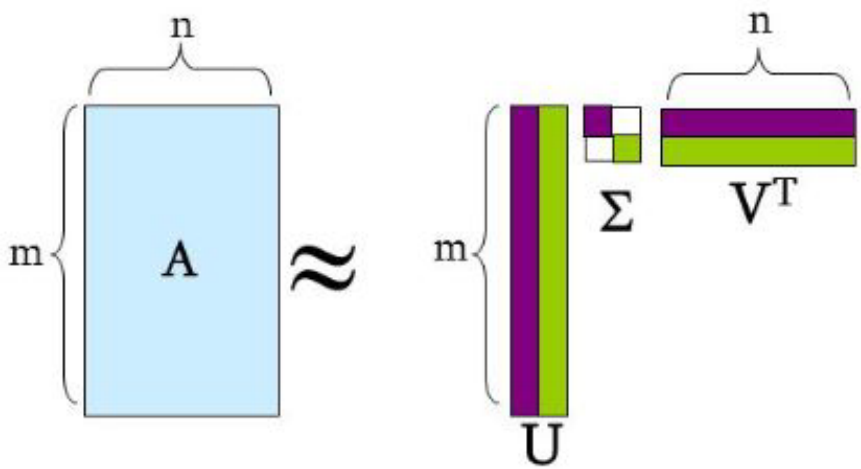


	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

	T1	T2	T3	T4	T5	T6	T7	T8
Doc1	2	0	4	3	0	1	0	2
Doc2	0	2	4	0	2	3	0	0
Doc3	4	0	1	3	0	1	0	1
Doc4	0	1	0	2	0	0	1	0
Doc5	0	0	2	0	0	4	0	0
Doc6	1	1	0	2	0	1	1	3
Doc7	2	1	3	4	0	2	0	2

Singular Value Decomposition

$$\mathbf{A} \approx \mathbf{U}\Sigma\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i^T$$



- **U,V**
  - Columns are orthogonal and unit vectors
- **Σ**
  - Entries (singular values) are positive and sorted in decreasing order of importance

Singular Value Decomposition

Original Matrix

	document	error	invalid	message	file	format	unable	to	open	using	path	variable
1	d1	1	1	1	1	1	0	0	0	0	0	0
2	d2	1	0	2	1	0	1	1	1	1	1	0
3	d3	1	0	0	0	1	1	1	0	0	0	1

(3,11)

≈

	document	SVD1	SVD2
1	d1	1.63	.49
2	d2	3.14	-.96
3	d3	1.35	1.64

(3,N)

×

Sorted Singular Values

12.29		
	6.2	
		...

(N,N)

×

T

Weights

	U_2	
error	43	.30
invalid	.11	.13
message	.55	-.37
file	.33	-.12
format	.21	.55
unable	.31	.18
to	.31	.18
open	.22	-.25
using	.22	-.25
path	.22	-.25
variable	.09	.42

(11,N)

T

- **Top N=2 dimensions**

Singular Value Decomposition

Original Matrix

	document	error	invalid	message	file	format	unable	to	open	using	path	variable
1	d1	1	1	1	1	1	0	0	0	0	0	0
2	d2	1	0	2	1	0	1	1	1	1	1	0
3	d3	1	0	0	0	1	1	1	0	0	0	1

- **New Matrix**

	document	SVD1	SVD2
1	d1	1.63	.49
2	d2	3.14	-.96
3	d3	1.35	1.64

[3,N]

x

Sorted Singular Values

12.29		
	6.2	
		...

[N,N]

[3,N]

- **Dimensions reduced from 11 to N=2**

Dimension Reduction

Original Matrix

	document	error	invalid	message	file	format	unable	to	open	using	path	variable
1	d1	1	1	1	1	1	0	0	0	0	0	0
2	d2	1	0	2	1	0	1	1	1	1	1	0
3	d3	1	0	0	0	1	1	1	0	0	0	1

Step1. Apply SVD/PCA

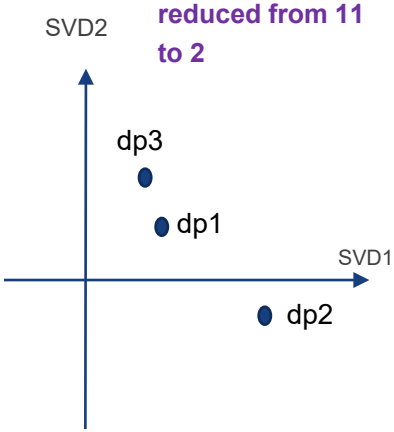
You get SVDs/Concepts.

	document	SVD1	SVD2
1	d1	1.63	.49
2	d2	3.14	-.96
3	d3	1.35	1.64

✖

Sorted Singular Values	
12.29	
	6.2

Dimensions reduced from 11 to 2



DataPoint1 = [1,1,1,1,1,0,0,0,0,0,0]

DataPoint2 = [1,1,2,1,0,1,1,1,1,1,0]

DataPoint3 = [1,0,0,0,1,1,1,0,0,0,1]

≈

Datapoint1 = [20.1, 3.0]

Datapoint2 = [38.6,-5.95]

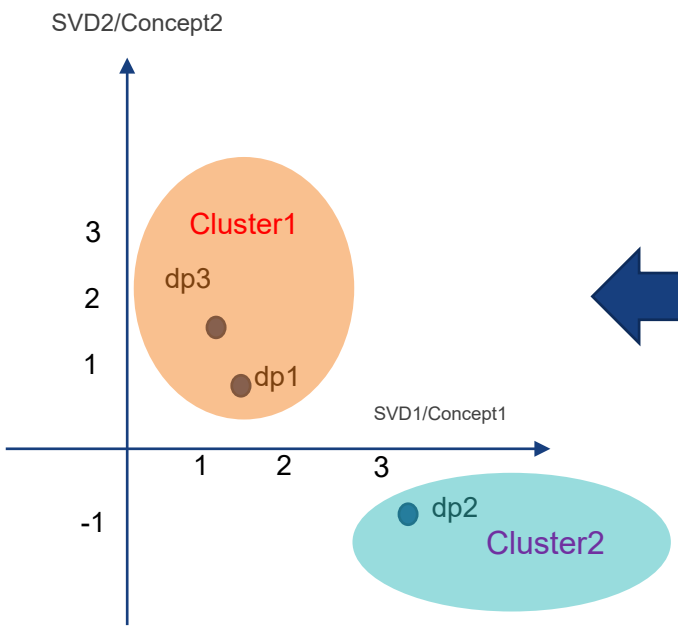
Datapoint3 = [16.6, 10.2]

Concept#

/SVD#

≠

Cluster#



Step 2. Apply KM
Or other classifiers

Singular Value Decomposition

SVD – How Many Dimensions?

- Usually no more than 5 to 20 dimensions extract most of the information from the TDM.
- More dimensions (up to a few hundred) can be retained if the processed data is for subsequent predictive modeling or clustering

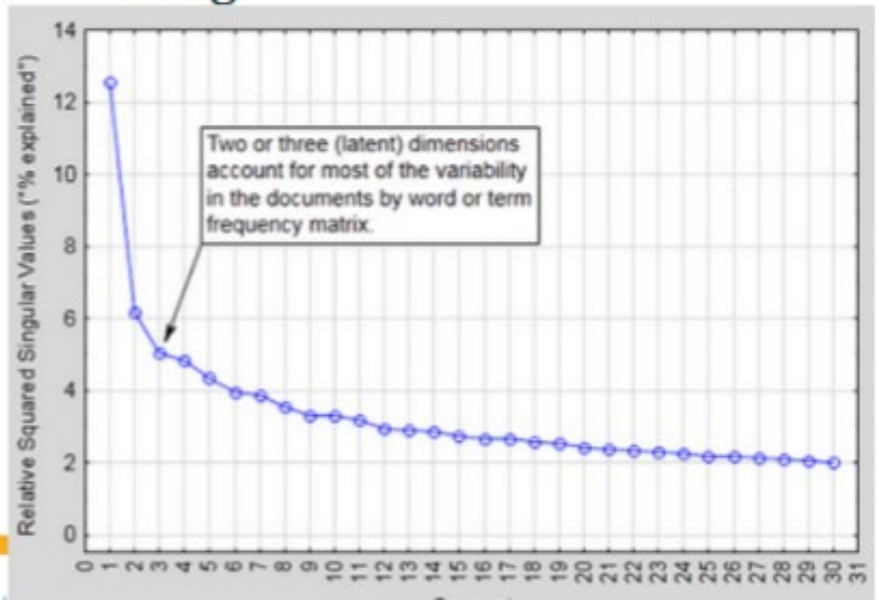
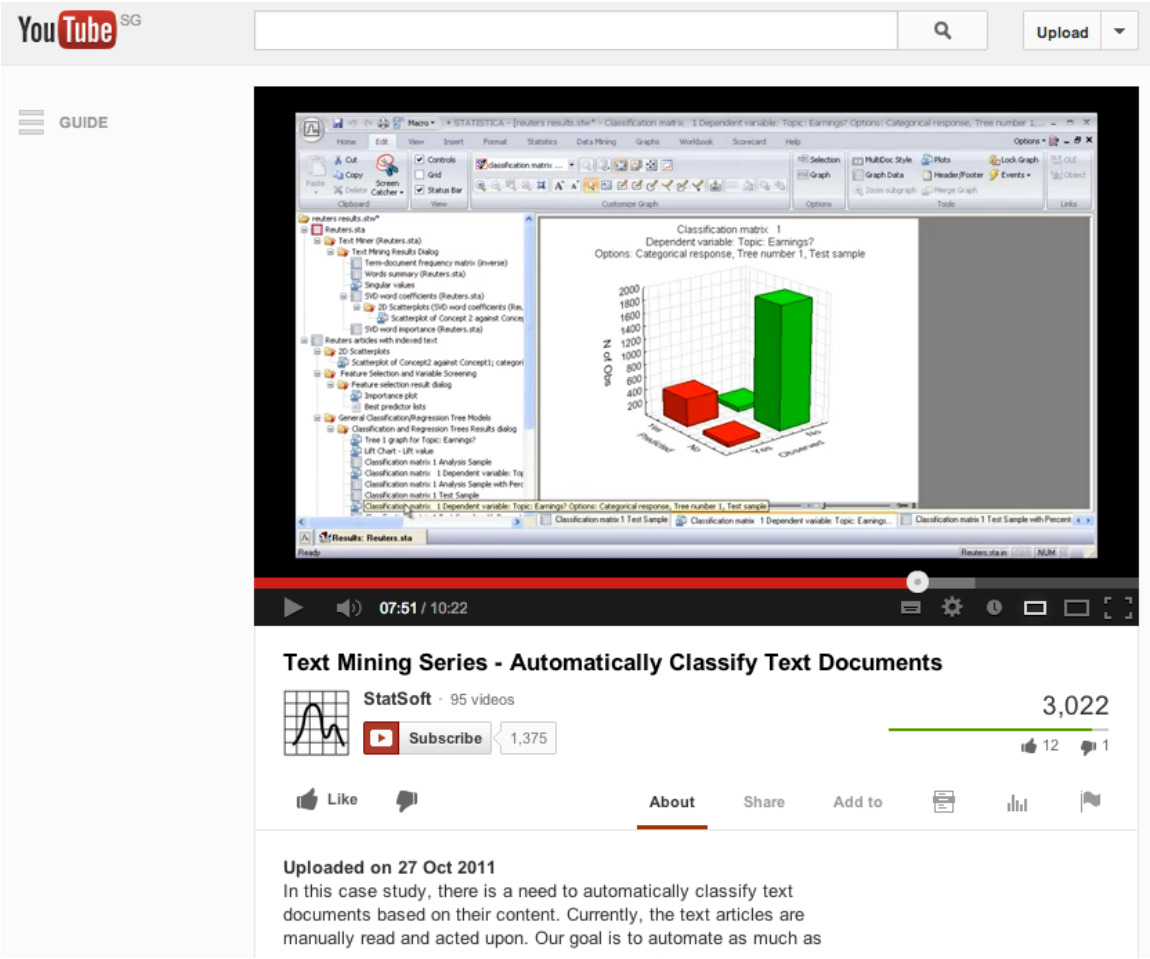


Figure 11.3 Plot of relative squared singular values by number of latent semantic dimensions
From *Practical Text Mining and Statistical Analysis for Non-structured Text data*

Automatic Categorization of Documents

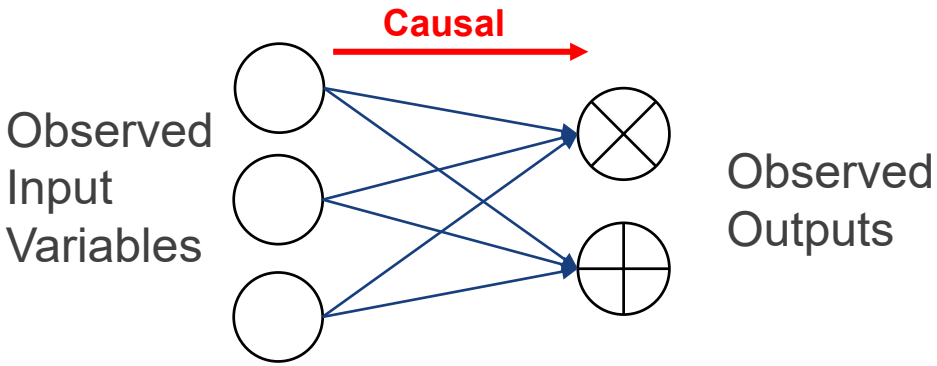


From: http://www.youtube.com/watch?v=Q5K3gyQJkC0

TOPIC MODELING

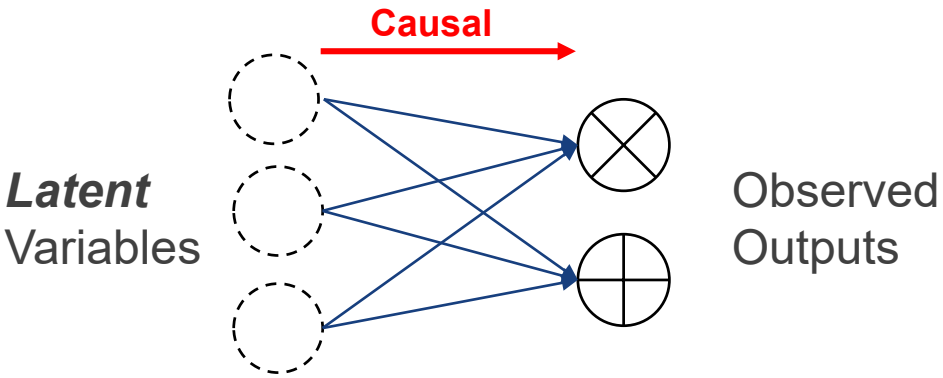
Supervised vs Unsupervised

- **Supervised categorization**



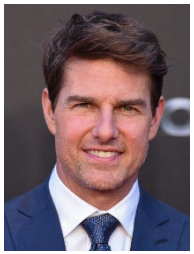
Essentially similar models
- **Inputs causally effects the outputs**
- **Input variables may not be observable**

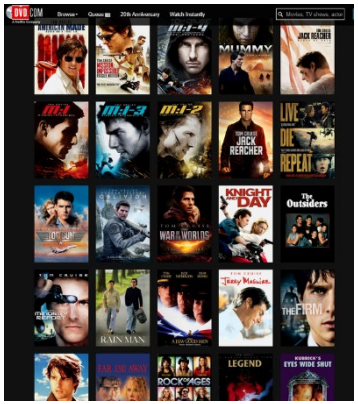
- **Unsupervised categorization**



Example

- Actors and their movies are observed inputs





- The potential tags of “Action”, “War” are latent variables

Text Example

- **What's unusual?**
  - Anomaly detection

Date	Amount	Location
2 Mar	$40	Penang
5 Mar	$20	KL
17 Mar	$30	KL
4 Apr	$80	Ipoh
9 Apr	$30	KL
14 Apr	$70	KL
20 May	$100	Johor
25 May	$20	KL
31 May	$3	Kiev
4 Jun	$40	KL
23 Jun	$50	KL
30 Jun	$30	KL
16 Jul	$70	Ipoh
16 Jul	$50	Ipoh

What is topic modeling?

- **"Topic" modeling**
  - Can we figure out what discourses (==latent variables) would generate the collection of documents?
  - These discourses are just bunches of words
    - If done well, the bunches of words would seem naturally to be together, e.g.,
      - "wag", "bark", "bone", "bite", "dog"
      - "pilot", "plane", "wing", "flight"
  - These bunches of words constitute **topics**

© 2015 National University of Singapore. All Rights Reserved

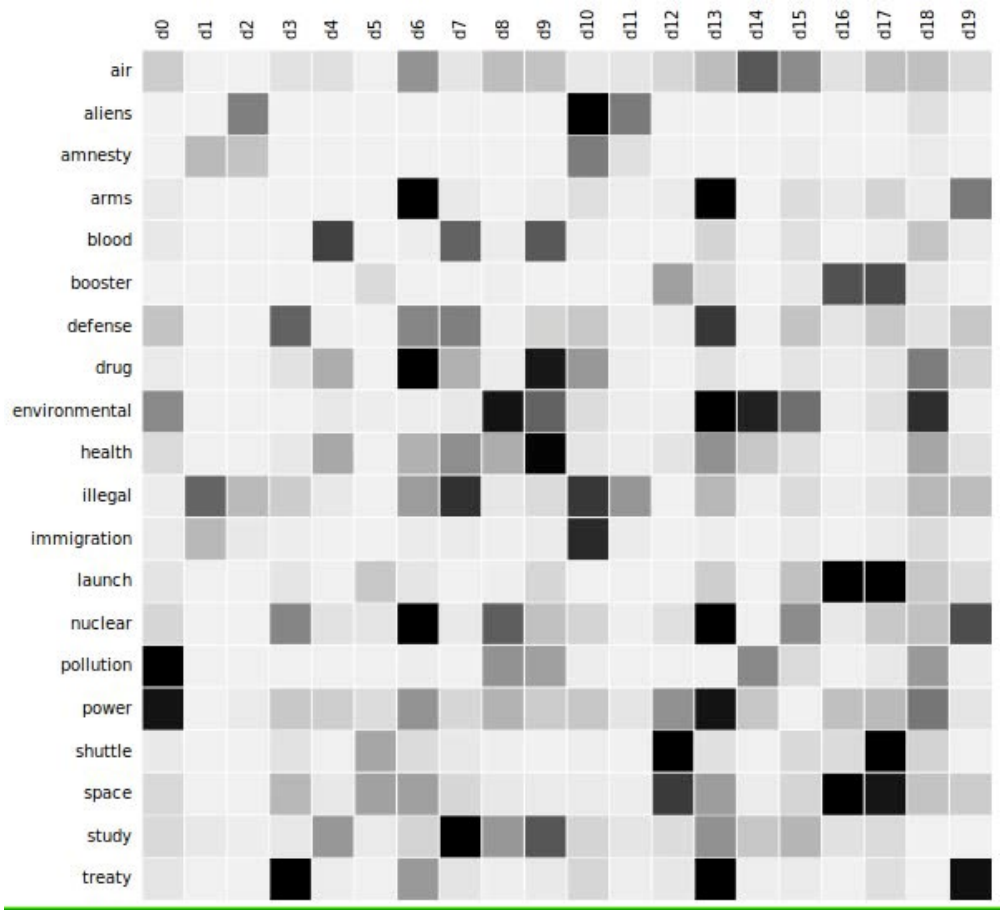
Animation of topic modeling

- **Columns = documents**
- **Rows = words**
- **Squares = frequency**
- **Darker = higher frequency**

- **Group:**
  - Documents using similar words
  - Words which occur in similar documents

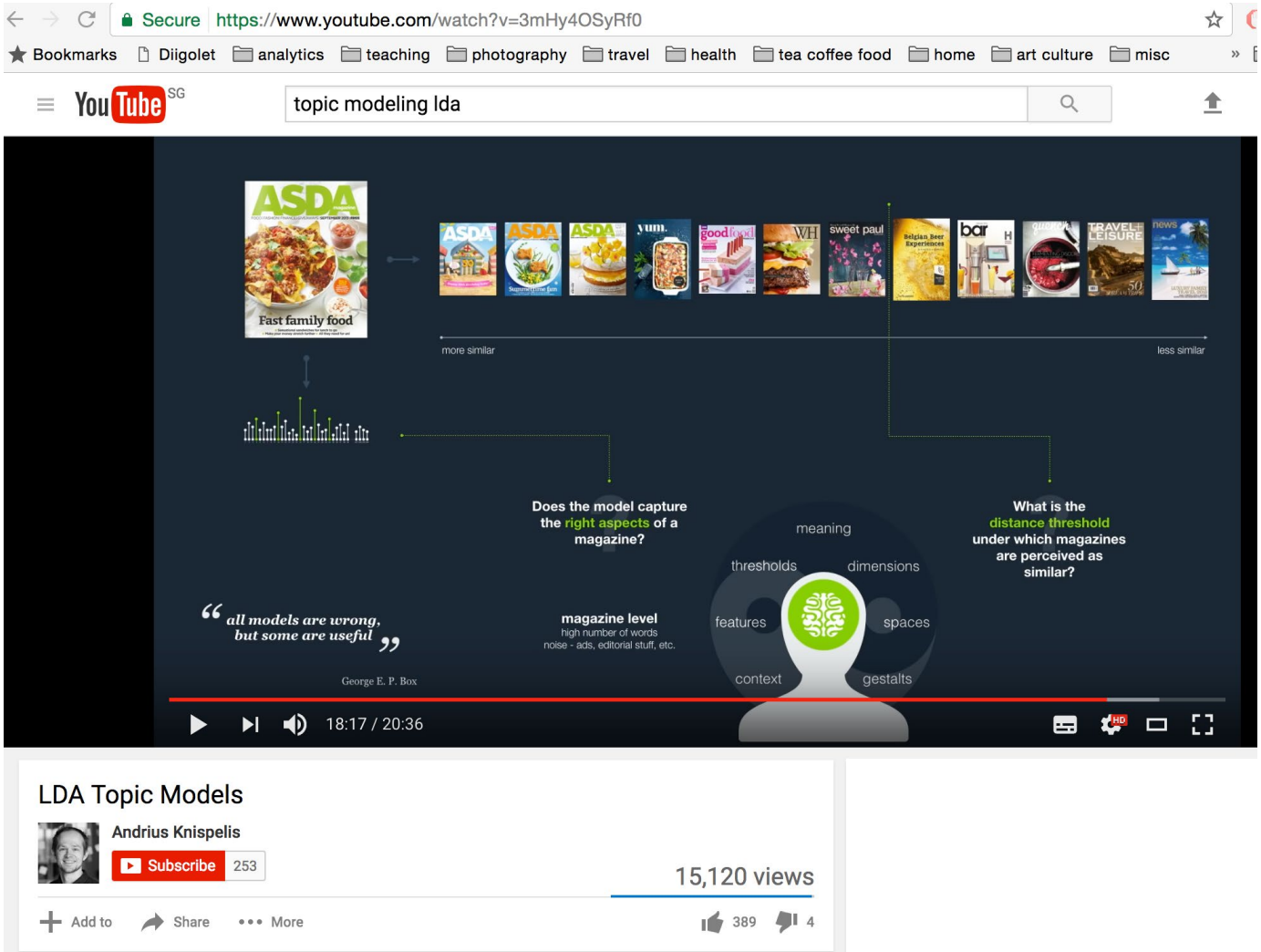
Resulting set of words are "topics"

Number of Groups are pre-defined



From: http://topicmodels.west.uni-koblenz.de/ckling/tmt/svd_ap.html

LDA Topic Model Explanation



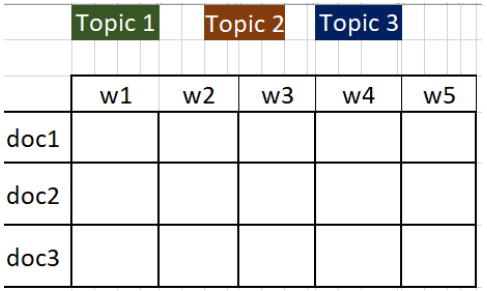
From: https://www.youtube.com/watch?v=3mHy4OSyRf0

More examples of applications

- **Analysis of text, e.g.,**
  - Diachronic analysis:
    - Speeches during election campaign
      - Economy, abortion, build wall, reduce taxes,…
    - Speeches after taking office
      - Reduce taxes, create jobs, immigration, China,…
  - Contrast analysis:
    - Different candidates positions and issues
    - Characteristics of various media publications

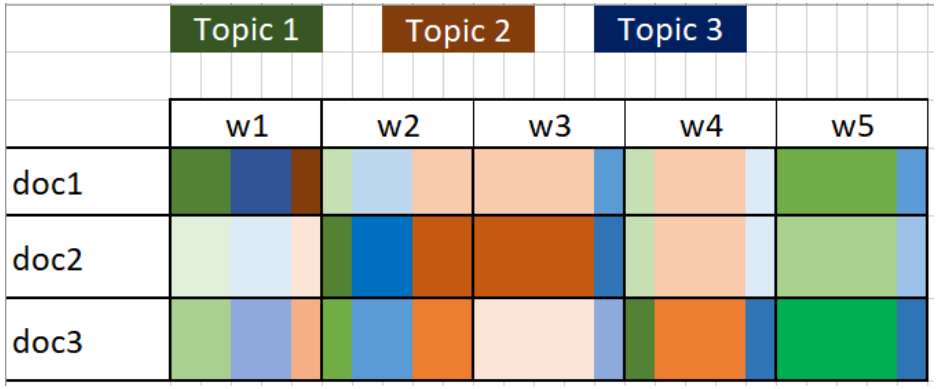
Animation of topic modeling

Input



- Predefine number of topics
- TDM

output



- Word-Topic distribution
- Doc-Topic distribution
- Topics are indexed with numbers without tags
- Topics are represented by a list of (important) words

Reference & Resources

- Fabrizio Sebastiani, *A Tutorial on Automated Text Categorization*, web.iiit.ac.in/~jawahar/PRA-03/textCat.pdf
- F. Aiolli, *Text Categorization*, downloaded from http://www.math.unipd.it/~aiolli/corsi/SI-0607/Lez09.251006.pdf
- John Elder, Gary Miner, Bob Nisbet. *Practical Text Mining and Statistical Analysis for non-Structured Text Data Applications*, Academic Press, 2012
- Chris Manning & Hinrich Schutze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999
- Scott Weingart, *Topic Modeling for Humanists: A Guided Tour*, downloaded from http://www.scottbot.net/HIAL/index.html@p=19113.html
- Ted Underwood, *Topic Modeling made just simple enough*, downloaded from https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/
- NLP resources: http://nlp.stanford.edu/links/statnlp.html

Document and Topic Models: pLSA and LDA

Andrew Levandoski and Jonathan Lobo

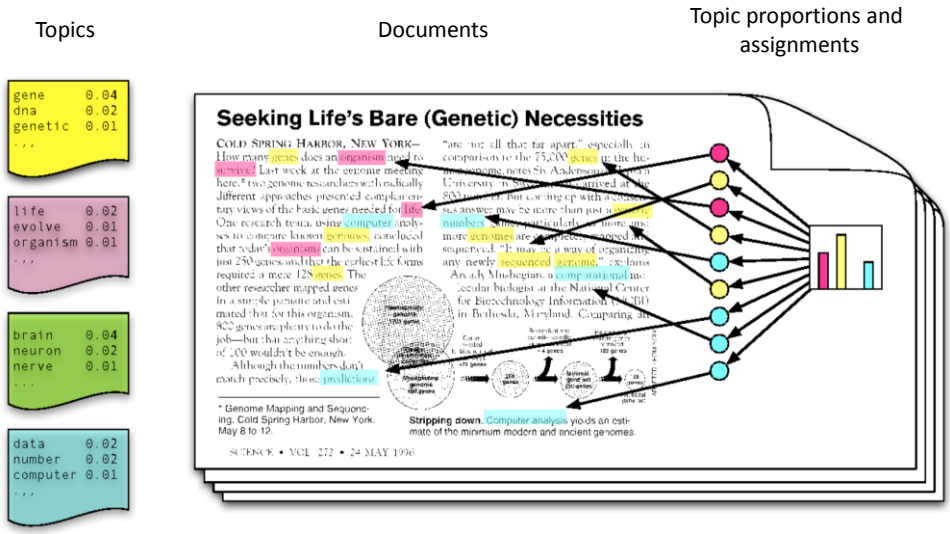
CS 3750 Advanced Topics in Machine Learning

2 October 2018

Outline

- Topic Models
- pLSA
  - LSA
  - Model
  - Fitting via EM
  - pHITS: link analysis
- LDA
  - Dirichlet distribution
  - Generative process
  - Model
  - Geometric Interpretation
  - Inference

Topic Models: Visual Representation



3

Topic Models: Importance

- For a given corpus, we learn two things:
  1. **Topic:** from full vocabulary set, we learn important subsets
  2. **Topic proportion:** we learn what each document is about
- This can be viewed as a form of dimensionality reduction
  - From large vocabulary set, extract basis vectors (topics)
  - Represent document in topic space (topic proportions)
  - Dimensionality is reduced from $\{w_i\} \in \mathbb{Z}_V^N$ to $\theta \in \mathbb{R}^K$
- Topic proportion is useful for several applications including document classification, discovery of semantic structures, sentiment analysis, object localization in images, etc.

4

Topic Models: Terminology

- **Document Model**
  - **Word:** element in a vocabulary set
  - **Document:** collection of words
  - **Corpus:** collection of documents
- **Topic Model**
  - **Topic:** collection of words (subset of vocabulary)
  - Document is represented by (latent) mixture of topics
    - $p(w|d) = p(w|z)p(z|d)$ (z : topic)
- **Note:** document is a *collection* of words (not a *sequence*)
  - *'Bag of words'* assumption
  - In probability, we call this the *exchangeability* assumption
    - $p(w_1, \dots, w_N) = p(w_{\sigma(1)}, \dots, w_{\sigma(N)})$ (σ: permutation)

5

Topic Models: Terminology (cont'd)

- Represent each document as a vector space
- A word is an item from a vocabulary indexed by $\{1, \dots, V\}$. We represent words using unit-basis vectors. The vth word is represented by a V vector w such that $w^v = 1$ and $w^u = 0$ for $v \neq u$.
- A document is a sequence of n words denoted by $\mathrm{w} = (w_1, w_2, \dots w_n)$ where w_n is the nth word in the sequence.
- A corpus is a collection of M documents denoted by
$$D = \{w_1, w_2, \dots w_m\}.$$

6

Probabilistic Latent Semantic Analysis (pLSA)

Motivation

- Learning from text and natural language
- Learning meaning and usage of words without prior linguistic knowledge
- Modeling semantics
  - Account for polysems and similar words
  - Difference between what is said and what is meant

Vector Space Model

- Want to represent documents and terms as vectors in a lower-dimensional space
- N × M word-document co-occurrence matrix $\boldsymbol{N}$

$$D = \{d_1, \ldots, d_N\}$$

$$\mathrm{W} = \{w_1, \ldots, w_M\}$$

$$\boldsymbol{N} = \left(n(d_i, w_j)\right)_{ij}$$

  - limitations: high dimensionality, noisy, sparse
  - solution: map to lower-dimensional latent semantic space using SVD

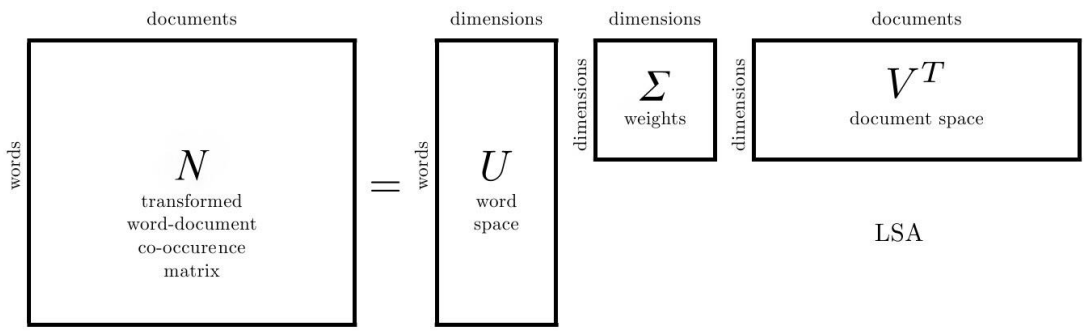
9

Latent Semantic Analysis (LSA)

- Goal
  - Map high dimensional vector space representation to lower dimensional representation in latent semantic space
  - Reveal semantic relations between documents (count vectors)
- SVD
  - $N = U\Sigma V^T$
  - U: orthogonal matrix with left singular vectors (eigenvectors of NN^T)
  - V: orthogonal matrix with right singular vectors (eigenvectors of N^TN)
  - Σ: diagonal matrix with singular values of N
- Select *k* largest singular values from Σ to get approximation $\widetilde{N}$ with minimal error
  - Can compute similarity values between document vectors and term vectors

10

LSA



LSA Strengths

- Outperforms naïve vector space model
- Unsupervised, simple
- Noise removal and robustness due to dimensionality reduction
- Can capture synonymy
- Language independent
- Can easily perform queries, clustering, and comparisons

LSA Limitations

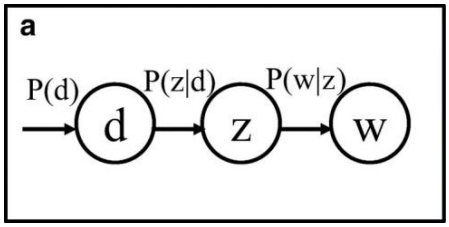
- No probabilistic model of term occurrences
- Results are difficult to interpret
- Assumes that words and documents form a joint Gaussian model
- Arbitrary selection of the number of dimensions k
- Cannot account for polysemy
- No generative model

Probabilistic Latent Semantic Analysis (pLSA)

- Difference between topics and words?
  - Words are observable
  - Topics are not, they are latent
- Aspect Model
  - Associates an unobserved latent class variable $z \in \mathbb{Z} = \{z_1, \ldots, z_K\}$ with each observation
  - Defines a joint probability model over documents and words
  - Assumes *w* is independent of *d* conditioned on *z*
  - Cardinality of *z* should be much less than than *d* and *w*

pLSA Model Formulation

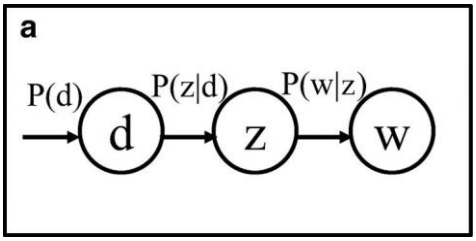
- Basic Generative Model
  - Select document *d* with probability *P(d)*
  - Select a latent class *z* with probability *P(z|d)*
  - Generate a word *w* with probability *P(w|z)*
- Joint Probability Model

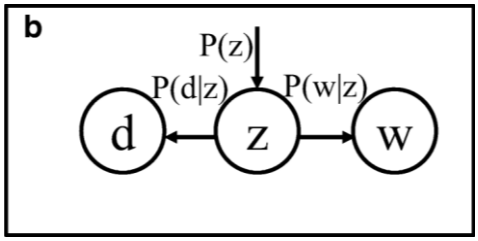


$$P(d,w) = P(d)P(w|d) \qquad P(w|d) = \sum_{z \in \mathbb{Z}} P(w|z)P(z|d)$$

15

pLSA Graphical Model Representation





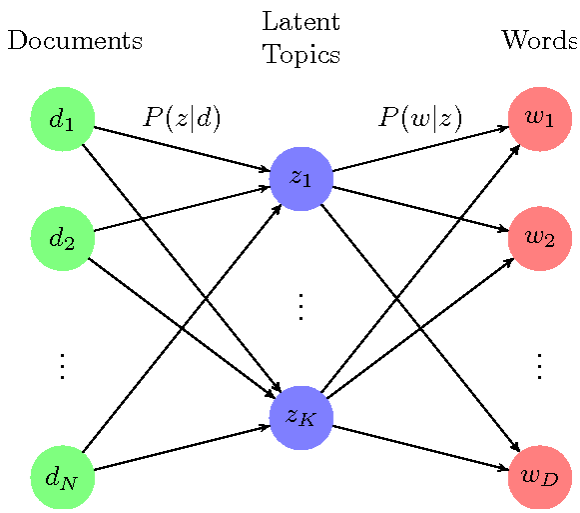
$$P(d,w) = P(d)P(w|d)$$

$$P(w|d) = \sum_{z \in \mathbb{Z}} P(w|z)P(z|d)$$

$$P(d,w) = \sum_{z \in \mathbb{Z}} P(z)P(d|z)P(w|z)$$

16

pLSA Joint Probability Model



$$P(d,w) = P(d)P(w|d)$$

$$P(w|d) = \sum_{z \in \mathbb{Z}} P(w|z)P(z|d)$$

Maximize:

$$\mathcal{L} = \sum_{d \epsilon D} \sum_{w \epsilon W} n(d,w) \log P(d,w)$$

Corresponds to a minimization of KL divergence (cross-entropy) between the empirical distribution of words and the model distribution P(w|d)

17

Probabilistic Latent Semantic Space

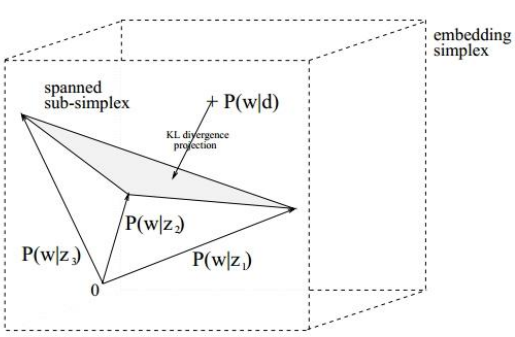


Figure 1: Sketch of the probability sub-simplex spanned by the aspect model.

- P(w|d) for all documents is approximated by a multinomial combination of all factors P(w|z)
- Weights P(z|d) uniquely define a point in the latent semantic space, represent how topics are mixed in a document

18

Probabilistic Latent Semantic Space

- Topic represented by probability distribution over words

$$z_i = (w_1, \ldots, w_m) \qquad z_1 = (0.3, 0.1, 0.2, 0.3, 0.1)$$

- Document represented by probability distribution over topics

$$d_j = (z_1, \ldots, z_n) \qquad d_1 = (0.5, 0.3, 0.2)$$

19

Model Fitting via Expectation Maximization

- E-step

$$P(z|d,w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z'} P(z')P(d|z')P(w|z')}$$

Compute posterior probabilities for latent variables z using current parameters

- M-step

$$P(w|z) = \frac{\sum_d n(d,w)P(z|d,w)}{\sum_{d,w'} n(d,w')P(z|d,w')}$$

$$P(d|z) = \frac{\sum_w n(d,w)P(z|d,w)}{\sum_{d',w} n(d',w)P(z|d',w)}$$

Update parameters using given posterior probabilities

$$P(z) = \frac{1}{R}\sum_{d,w} n(d,w)P(z|d,w), \qquad R \equiv \sum_{d,w} n(d,w)$$

20

pLSA Strengths

- Models word-document co-occurrences as a mixture of conditionally independent multinomial distributions
- A mixture model, not a clustering model
- Results have a clear probabilistic interpretation
- Allows for model combination
- Problem of polysemy is better addressed

pLSA Strengths

- Problem of polysemy is better addressed

tie					spring				
trousers	season	scoreline	wires	operatic	beginning	dampers	flower	creek	humid
blouse	teams	goalless	cables	soprano	until	brakes	flowers	brook	winters
waistcoat	winning	equaliser	wiring	mezzo	months	suspension	flowering	river	summers
skirt	league	clinching	electrical	contralto	earlier	absorbers	fragrant	fork	ppen
sleeved	finished	scoreless	wire	baritone	year	wheels	lilies	piney	warm
pants	championship	replay	cable	coloratura	last	damper	flowered	elk	temperatures

pLSA Limitations

- Potentially higher computational complexity
- EM algorithm gives local maximum
- Prone to overfitting
  - Solution: Tempered EM
- Not a well defined generative model for new documents
  - Solution: Latent Dirichlet Allocation

pLSA Model Fitting Revisited

- Tempered EM
  - Goals: maximize performance on unseen data, accelerate fitting process
  - Define control parameter β that is continuously modified

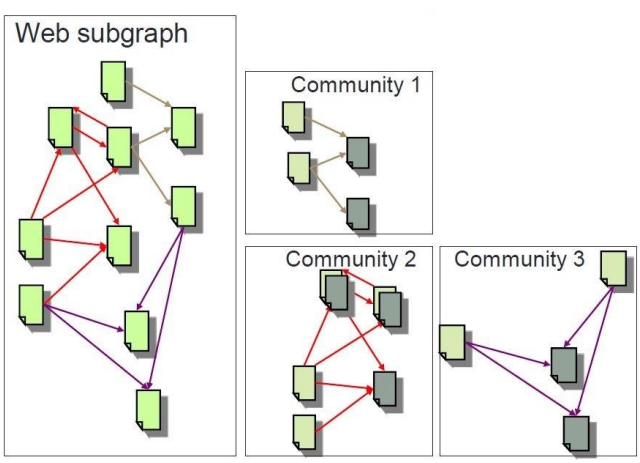
- Modified E-step

$$P_\beta(z|d,w) = \frac{P(z)[P(d|z)P(w|z)]^\beta}{\sum_{z'} P(z')[P(d|z')P(w|z')]^\beta}$$

Tempered EM Steps

1) Split data into training and validation sets
2) Set β to 1
3) Perform EM on training set until performance on validation set decreases
4) Decrease β by setting it to ηβ, where η <1, and go back to step 3
5) Stop when decreasing β gives no improvement

Example: Identifying Authoritative Documents



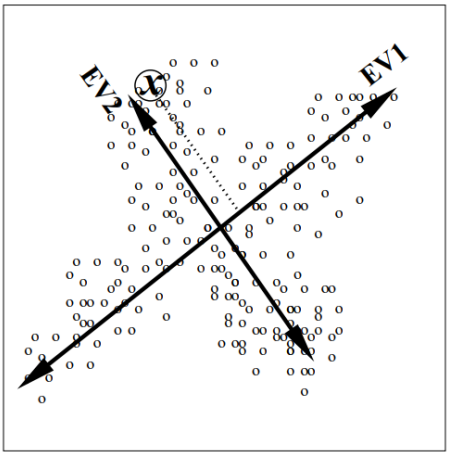
HITS

- Hubs and Authorities
  - Each webpage has an authority score *x* and a hub score *y*
  - Authority – value of content on the page to a community
    - likelihood of being cited
  - Hub – value of links to other pages
    - likelihood of citing authorities
  - A good hub points to many good authorities
  - A good authority is pointed to by many good hubs
- Principal components correspond to different communities
  - Identify the principal eigenvector of co-citation matrix

27

HITS Drawbacks

- Uses only the largest eigenvectors, not necessary the only relevant communities
- Authoritative documents in smaller communities may be given no credit
- Solution: Probabilistic HITS

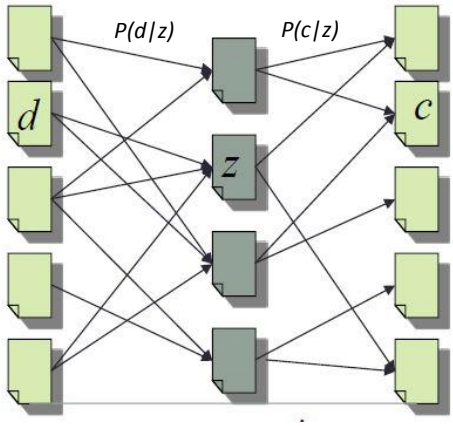


28

pHITS

$$P(d,c) = \sum_{z} P(z)P(c|z)P(d|z)$$

Documents



Citations

Communities

29

Interpreting pHITS Results

- Explain *d* and *c* in terms of the latent variable "community"
- Authority score: P(c|z)
  - Probability of a document being cited from within community *z*
- Hub Score: P(d|z)
  - Probability that a document *d* contains a reference to community *z*.
- Community Membership: P(z|c).
  - Classify documents

30

Joint Model of pLSA and pHITS

- Joint probabilistic model of document content (pLSA) and connectivity (pHITS)
  - Able to answer questions on both structure and content
  - Model can use evidence about link structure to make predictions about document content, and vice versa
  - Reference flow – connection between one topic and another
- Maximize log-likelihood function

$$\mathcal{L} = \sum_j \left[\alpha \sum_i \frac{N_{ij}}{\sum_{i'} N_{i'j}} log \sum_k P(w_i|z_k) P(z_k|d_j) + (1-\alpha) \sum_l \frac{A_{lj}}{\sum_{l'} A_{l'j}} log \sum_k P(c_l|z_k) P(z_k|d_j) \right]$$

31

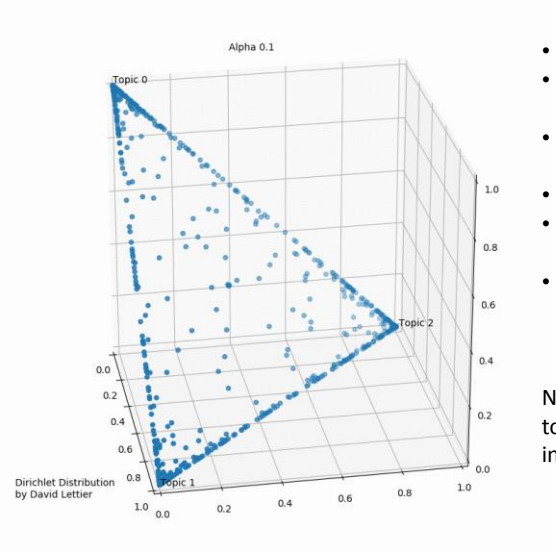
pLSA: Main Deficiencies

- Incomplete in that it provides no probabilistic model at the document level i.e. no proper priors are defined.
- Each document is represented as a list of numbers (the mixing proportions for topics), and there is no generative probabilistic model for these numbers, thus:
  1. The number of parameters in the model grows linearly with the size of the corpus, leading to overfitting
  2. It is unclear how to assign probability to a document outside of the training set
- Latent Dirichlet allocation (LDA) captures the exchangeability of both words *and* documents using a Dirichlet distribution, allowing a coherent generative process for test data

32

Latent Dirichlet Allocation (LDA)

LDA: Dirichlet Distribution



- A 'distribution of distributions'
- Multivariate distribution whose components all take values on (0,1) and which sum to one.
- Parameterized by the vector α, which has the same number of elements (k) as our multinomial parameter θ.
- Generalization of the beta distribution into multiple dimensions
- The alpha hyperparameter controls the mixture of topics for a given document
- The beta hyperparameter controls the distribution of words per topic

Note: Ideally we want our composites to be made up of only a few topics and our parts to belong to only some of the topics. With this in mind, alpha and beta are typically set below one.

LDA: Dirichlet Distribution (cont'd)

- A k-dimensional Dirichlet random variable θ can take values in the (k-1)-simplex (a k-vector θ lies in the (k-1)-simplex if $\theta_i \geq 0, \sum_{i=1}^{k} \theta_i = 1$) and has the following probability density on this simplex:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \dots \theta_k^{\alpha_k - 1},$$

  where the parameter α is a k-vector with components $\alpha_i > 0$ and where $\Gamma(x)$ is the Gamma function.
- The Dirichlet is a convenient distribution on the simplex:
  - In the exponential family
  - Has finite dimensional sufficient statistics
  - Conjugate to the multinomial distribution

35

LDA: Generative Process

LDA assumes the following generative process for each document w in a corpus D:

1. Choose $N \sim Poisson(\xi)$.
2. Choose $\theta \sim Dir(\alpha)$.
3. For each of the N words w_N:
   a. Choose a topic $z_n \sim Multinomial(\theta)$.
   b. Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Example: Assume a group of articles that can be broken down by three topics described by the following words:

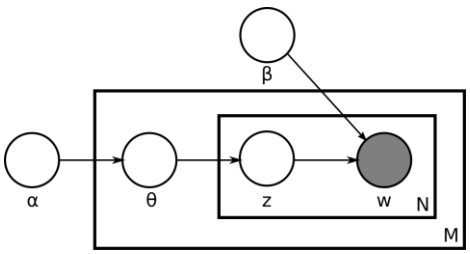
- Animals: dog, cat, chicken, nature, zoo
- Cooking: oven, food, restaurant, plates, taste
- Politics: Republican, Democrat, Congress, ineffective, divisive

To generate a new document that is 80% about animals and 20% about cooking:

- Choose the length of the article (say, 1000 words)
- Choose a topic based on the specified mixture (~800 words will coming from topic 'animals')
- Choose a word based on the word distribution for each topic

36

LDA: Model (Plate Notation)



α is the parameter of the Dirichlet prior on the per-document topic distribution,

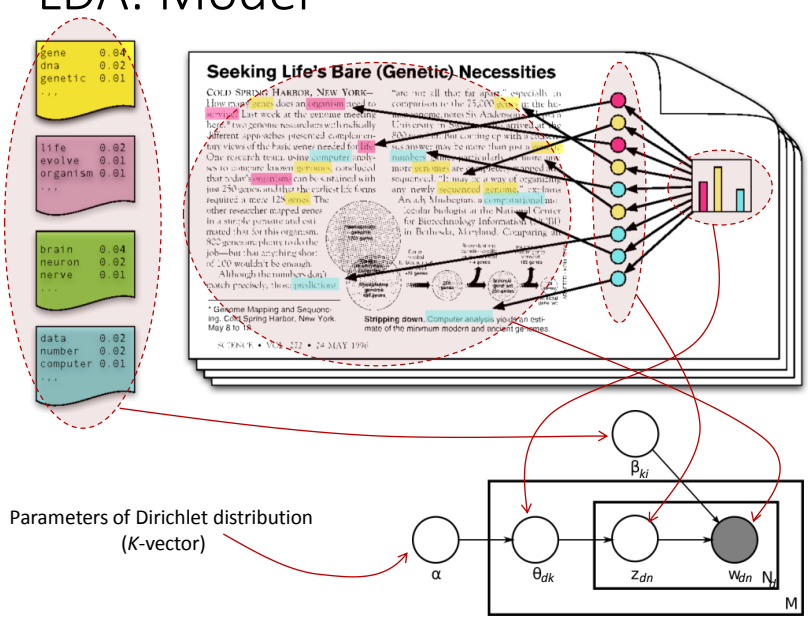
β is the parameter of the Dirichlet prior on the per-topic word distribution,

θ_M is the topic distribution for document *M,*

z_{MN} is the topic for the *N-th* word in document *M,* and

w_{MN} is the word.

LDA: Model



Parameters of Dirichlet distribution (*K*-vector)

1 … topic … *K*

1 ⋮ doc ⋮ *M*

θ_{dk}

1 … *n*th word … N_d

1 ⋮ doc ⋮ *M*

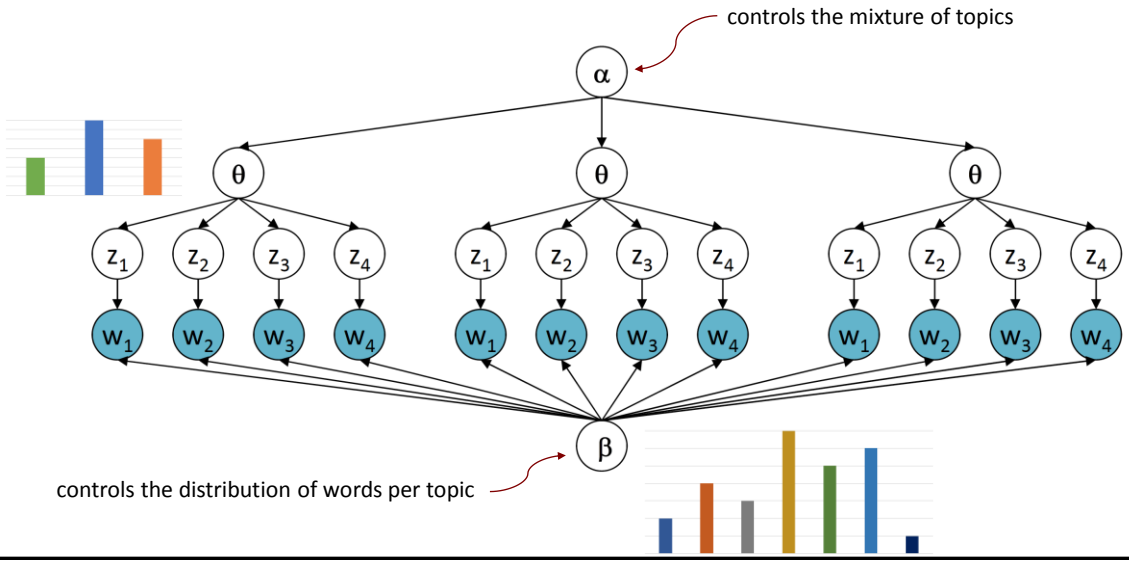
$z_{dn} = \{1, \dots, K\}$

1 … word idx … *V*

1 ⋮ topic ⋮ *M*

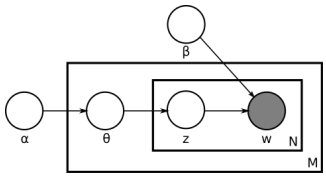
$\beta_{ki} = p(w|z)$

LDA: Model (cont'd)



39

LDA: Model (cont'd)



Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics z, and a set of N words w is given by:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta),$$

where $p(z_n|\theta)$ is θ_i for the unique i such that $z_n^i = 1$. Integrating over θ and summing over z, we obtain the marginal distribution of a document:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta_d.$$

Finally, taking the products of the marginal probabilities of single documents, we obtain the probability of a corpus:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta) \right) d\theta_d.$$

40

LDA: Exchangeability

- A finite set of random variables $\{x_1, \dots, x_N\}$ is said to be exchangeable if the joint distribution is invariant to permutation. If π is a permutation of the integers from 1 to N:

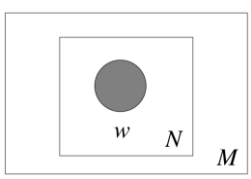
$$p(x_1, \dots, x_N) = p(x_{\pi_1}, \dots, x_{\pi_N})$$

- An infinite sequence of random numbers is infinitely exchangeable if every finite sequence is exchangeable
- We assume that words are generated by topics and that those topics are infinitely exchangeable within a document
- By De Finetti's Theorem:

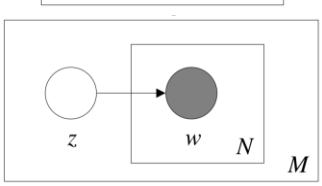
$$p(w, z) = \int p(\theta) \left(\prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n) \right) d\theta$$

41

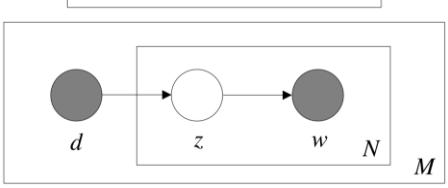
LDA vs. other latent variable models



Unigram model: $p(w) = \prod_{n=1}^{N} p(w_n)$



Mixture of unigrams: $p(w) = \sum_z p(z) \prod_{n=1}^{N} p(w_n|z)$

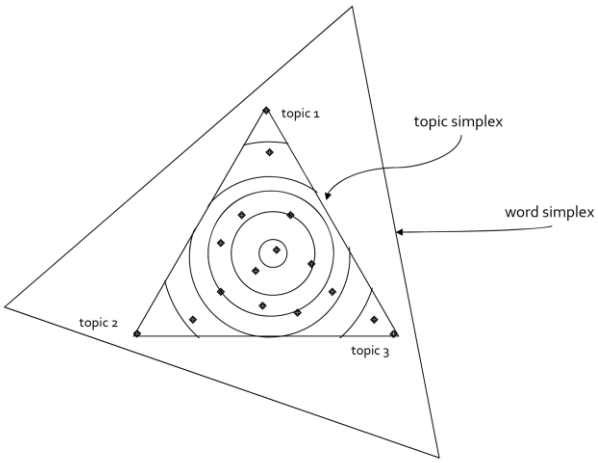


pLSI: $p(d, w_n) = p(d) \sum_z p(w_n|z) p(z|d)$

42

LDA: Geometric Interpretation

- Topic simplex for three topics embedded in the word simplex for three words
- Corners of the word simplex correspond to the three distributions where each word has probability one
- Corners of the topic simplex correspond to three different distributions over words
- Mixture of unigrams places each document at one of the corners of the topic simplex
- pLSI induces an empirical distribution on the topic simplex denoted by diamonds
- LDA places a smooth distribution on the topic simplex denoted by contour lines



43

LDA: Goal of Inference

LDA inputs: Set of words per document for each document in a corpus

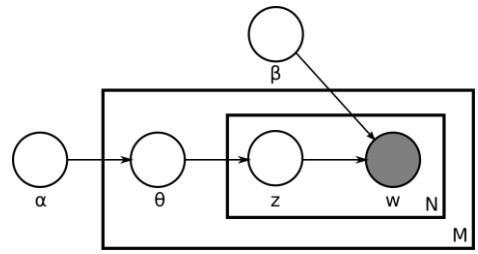
LDA outputs: Corpus-wide topic vocabulary distributions

Topic assignments per word

Topic proportions per document

44

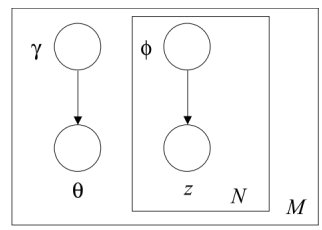
LDA: Inference



The key inferential problem we need to solve with LDA is that of computing the posterior distribution of the hidden variables given a document:

$$p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)}$$

This formula is intractable to compute in general (the integral cannot be solved in closed form), so to normalize the distribution we marginalize over the hidden variables:



$$p(w|\alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta$$

45

LDA: Variational Inference

- Basic idea: make use of Jensen's inequality to obtain an adjustable lower bound on the log likelihood
- Consider a family of lower bounds indexed by a set of variational parameters chosen by an optimization procedure that attempts to find the tightest possible lower bound
- Problematic coupling between θ and β arises due to edges between θ, z and w. By dropping these edges and the w nodes, we obtain a family of distributions on the latent variables characterized by the following variational distribution:

$$q(\theta, z|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n)$$

where γ and $(\phi_1, \dots, \phi_n)$ and the free variational parameters.

46

LDA: Variational Inference (cont'd)

- With this specified family of probability distributions, we set up the following optimization problem to determine θ and ϕ:

$$(\gamma^*, \phi^*) = argmin_{(\gamma,\phi)} D(q(\theta, z|\gamma, \phi) \parallel p(\theta, z|w, \alpha, \beta))$$

- The optimizing values of these parameters are found by minimizing the KL divergence between the variational distribution and the true posterior $p(\theta, z|w, \alpha, \beta)$
- By computing the derivatives of the KL divergence and setting them equal to zero, we obtain the following pair of update equations:

$$\phi_{ni} \propto \beta_{iw_n} \exp\{E_q[\log(\theta_i)\,|\gamma]\}$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}$$

- The expectation in the multinomial update can be computed as follows:

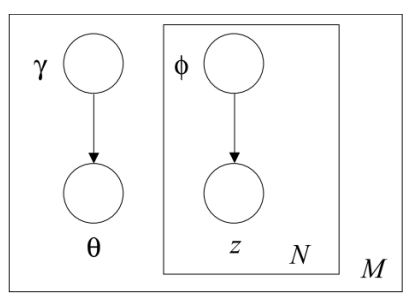
$$E_q[\log(\theta_i)\,|\gamma] = \Psi(\gamma_i) - \Psi(\prod_{j=1}^{k} \gamma_j)$$

where Ψ is the first derivative of the $\log\Gamma$ function.

47

LDA: Variational Inference (cont'd)

```
initialize φ^0_ni := 1/k for all i and n
initialize γ_i := α_i + N/k for all i
repeat
    for n = 1 to N
        for i = 1 to k
            φ^{t+1}_ni := β_{iw_n} exp(Ψ(γ^t_i))
        normalize φ^{t+1}_n to sum to 1.
    γ^{t+1} := α + Σ_{n=1}^{N} φ^{t+1}_n
until convergence
```



48

LDA: Parameter Estimation

- Given a corpus of documents $D = \{w_1, w_2 \dots, w_M\}$, we wish to find α and β that maximize the marginal log likelihood of the data:

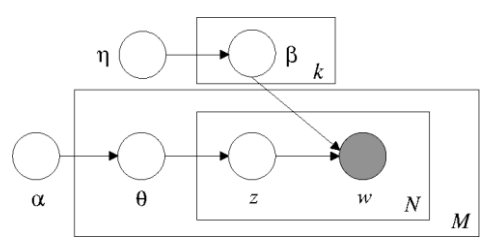
$$\ell(\alpha, \beta) = \sum_{d=1}^{M} logp(w_d|\alpha, \beta)$$

- Variational EM yields the following iterative algorithm:
  1. (E-step) For each document, find the optimizing values of the variational parameters $\{\gamma_d^*, \phi_d^* : d \in D\}$
  2. (M-step) Maximize the resulting lower bound on the log likelihood with respect to the model parameters α and β

  These two steps are repeated until the lower bound on the log likelihood converges.

LDA: Smoothing

- Introduces Dirichlet smoothing on β to avoid the zero frequency word problem
- Fully Bayesian approach:



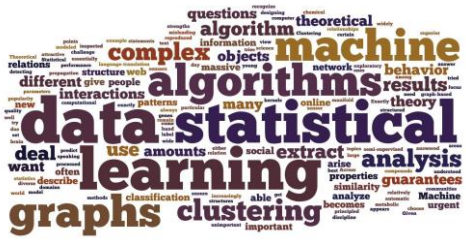
$$q(\beta_{1:k}, z_{1:M}, \theta_{1:M}|\lambda, \phi, \gamma) = \sum_{i=1}^{k} Dir(\beta_i|\lambda_i) \prod_{d=1}^{M} q_d(\theta_d, z_d|\phi_d, \gamma_d)$$

where $q_d(\theta, z\,|\phi, \gamma)$ is the variational distribution defined for LDA. We require an additional update for the new variational parameter λ:

$$\lambda_{ij} = \eta + \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni}^* w_{dn}^j$$

Topic Model Applications

- Information Retrieval
- Visualization
- Computer Vision
  - Document = image, word = “visual word”
- Bioinformatics
  - Genomic features, gene sequencing, diseases
- Modeling networks
  - cities, social networks



51

pLSA / LDA Libraries

- gensim (Python)
- MALLET (Java)
- topicmodels (R)
- Stanford Topic Modeling Toolbox

References

David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation. JMLR, 2003.

David Cohn and Huan Chang. Learning to probabilistically identify Authoritative documents. ICML, 2000.

David Cohn and Thomas Hoffman. The Missing Link - A Probabilistic Model of Document Content and Hypertext Connectivity. NIPS, 2000.

Thomas Hoffman. Probabilistic Latent Semantic Analysis. UAI-99, 1999.

Thomas Hofmann. Probabilistic Latent Semantic Indexing. SIGIR-99, 1999.