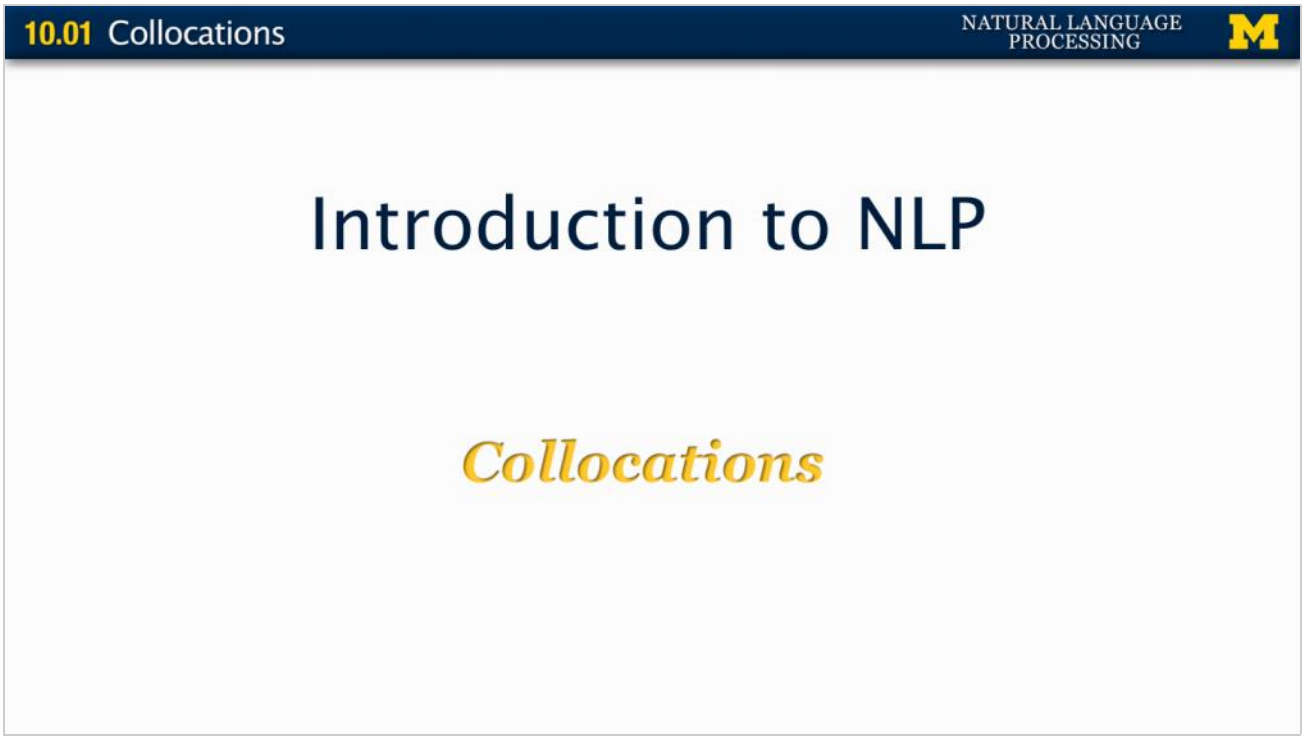
Week 10

Thursday, August 11, 2016 15:20

Collocation

10.1



Collocations (phrases)

- Dictionary definitions
  - Meaning of words in isolation
- "Know a word by the company that it keeps"
  - Firth 1935
- Examples
  - dead end
  - strong tea
  - Benazir Bhutto
  - Fabry disease

not about death
not "powerful"
names of people disease

Collocations

- Properties
  - Common use
  - No general syntactic or semantic rules
  - Important for non-native speakers
- Collocation acquisition
  - Important for NLP

Types Of Multiword Sequences

- Idioms
- Free-word combinations
- Collocations

10.01 Collocations NATURAL LANGUAGE PROCESSING

Examples

different from the mean from the original word

Idioms	Collocations	Free-word combinations
To kick the bucket	To trade actively	To take the bus
Dead end	Table of contents	The end of the road
To catch up	Orthogonal projection	To buy a house

Properties

- Arbitrariness: substitutions are usually not allowed:
  - Make an effort vs. *make an exertion
  - Running commentary vs. *running discussion
  - Commit treason vs. *commit treachery
- Language- and dialect-specific
  - Régler la circulation = direct traffic
  - Russian, German, Serbo-Croatian: direct translation of regulate is used
  - AE: set the table, make a decision
  - BE: lay the table, take a decision
  - "semer le désarroi" - "to sow disarray" - "to wreak havoc"
- Common in technical language
- Recurrent in context

10.01 Collocations NATURAL LANGUAGE PROCESSING

Uses

- Disambiguation (e.g, "bank"/"loan","river")
- Translation
- Generation

Types of Collocations

- Grammatical
  - come to, put on; afraid that, fond of, by accident, witness to
- Semantic
  - only certain synonyms
- Flexible
  - find/discover/notice by chance

Some collocations are flexible

Base-Collocator Pairs

- Base – bears most of the meaning of the collocation. Writers think of the base first. Foreign language speakers search by base. For decoding purposes, it is more appropriate to store the collocation under the collocator.

Base	Collocator	Example
Noun	verb	Set the table
Noun	adjective	Warm greetings
Verb	adverb	Struggle desperately
Adjective	adverb	Sound asleep
Verb	preposition	Put on

Extracting Collocations

- Most-common bigrams?
- Drop function words?
- Look at POS sequences?

not the right approach
only a few occasions
pretty unique ...
don't fit into POS tagging framework

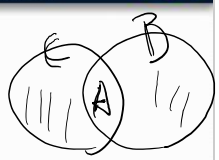
Extracting Collocations

- Mutual information

$$I(x;y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

- Larger means stronger
- What if I(x;y) = 0?
- What if I(x;y) < 0?

Yule's Coefficient



A – frequency of pairs involving both W and X

B – frequency of pairs involving W only

C – frequency of pairs involving X only

D – frequency of pairs involving neither

$$Y = \frac{AD - BC}{AD + BC}$$

$$-1 \leq Y \leq 1$$

Example

	W	w			
X	A=800	C=180		A	800
x	B=160	D=80		B	160
				C	180
				D	80
				AD-BC	35200
				AD+BC	92800
					0.38

Example From The Hansard Corpus (Brown, Lai, And Mercer) – "Prime"

French word	Mutual information
sein	5.63
bureau	5.63
trudeau	5.34
premier	5.25
résidence	5.12
intention	4.57
no	4.53
session	4.34

Flexible And Rigid Collocations

- Example (from Smadja): "free" and "trade"

Total	p-5	p-4	p-3	p-2	p-1	p+1	p+2	p+3	p+4	p+5
8031	7	6	13	5	7918	0	12	20	26	24

Xtract (Smadja)

- The Dow Jones Industrial Average
- The NYSE's composite index of all its

Xtract (Smadja)

- The Dow Jones Industrial Average
- The NYSE's composite index of all its listed common stocks fell *NUMBER* to *NUMBER*

Translating Collocations

- Brush up a lesson, repasser une leçon
- Bring about/осуществлять
- Hansards:
  - late spring
  - fin du printemps
  - Atlantic Canada Opportunities Agency
  - Agence de promotion économique du Canada atlantique

Links

- Sample phrasal collocations
  - http://en.wiktionary.org/wiki/Appendix:Collocations_of_do,_have,_make,_and_take
- List of English language idioms
  - http://en.wikipedia.org/wiki/List_of_English-language_idioms
- Idiomsite
  - http://www.idiomsite.com/

example website containing collocations

10.01 Collocations NATURAL LANGUAGE PROCESSING

NLP

Information Retrieval

NLP

10.02 Information Retrieval
NATURAL LANGUAGE PROCESSING

Introduction to NLP

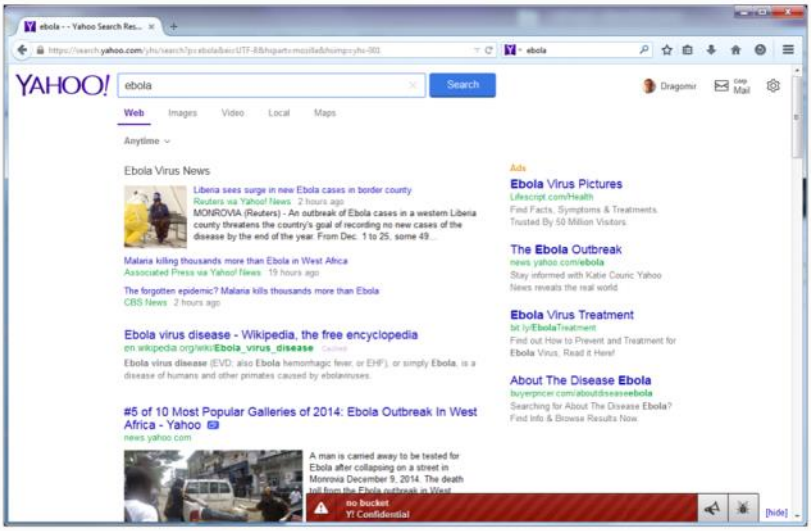
Information Retrieval

Introduction

- People search the Web daily
- Search engines
  - Google
  - Bing
  - Baidu
  - Yandex
- Information Retrieval is about search engines

10.02 Information Retrieval
NATURAL LANGUAGE PROCESSING

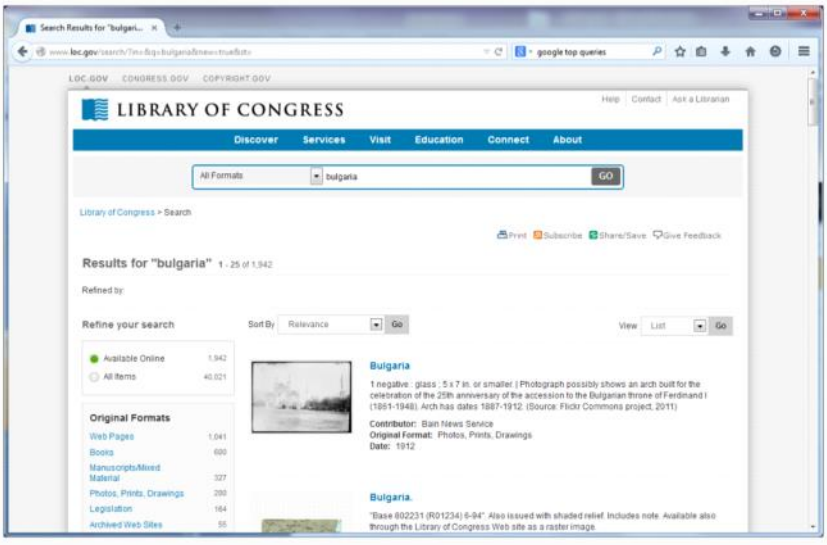
Yahoo Search



Amazon Search

10.02 Information Retrieval
NATURAL LANGUAGE PROCESSING

Library of Congress Search



Examples Of Search Engines

- Conventional (library catalog)
  - Search by keyword, title, author, etc.
- Text-based (Lexis-Nexis, Google, Yahoo!)
  - Search by keywords. Limited search using queries in natural language.
- Image-based
  - shapes, colors, keywords
- Question answering systems (ask.com)
  - Search in (restricted) natural language
- Clustering systems (Vivísimo, Clusty)
- Research systems (Lemur, Nutch)

HC : not very good

Sample Queries

- How to get rid of stretch marks
- Dodge
- Kourtney Kardashian
- How many calories are in pumpkn pie
- Angelina Jolie and Brad Pitt
- How to vote
- Derek Jeter
- Interstellar trailer
- What is Ebola?

https://www.google.com/trends/topcharts

The Size Of The World Wide Web

- The size of the indexed world wide web pages (by 2014)
  - Indexed by Google: about 45 Billion pages
  - Indexed by Bing: about 25 Billion pages

http://www.worldwidewebsize.com/

10.02 Information Retrieval
NATURAL LANGUAGE PROCESSING

Web Statistics

- Twitter hits 400 million tweets per day
  - June, 2012. Dick Costolo, CEO at Twitter
- Over 2.5 billion photos uploaded to Facebook each month (2010)
  - blog.facebook.com
- Google's clusters process a total of more than 20 petabytes of data per day.
  - 2008. Jeffrey Dean from Google

Challenges

- Dynamically generated content
- New pages get added all the time
- The size of the blogosphere doubles every 6 months

pages are not static

Characteristics Of User Queries

- Sessions
  - users revisit their queries
- Very short queries
  - typically 2 words long
- A large number of typos
- A small number of popular queries
  - A long tail of infrequent ones
- Almost no use of advanced query operators
  - with the exception of double quotes

Information Retrieval

- Baseline Process
  - Given a collection of documents
  - And a user's query
  - Find the most relevant documents

10.02 Information Retrieval
NATURAL LANGUAGE PROCESSING

Key Terms Used in IR

- Query
  - a representation of what the user is looking for – can be a list of words or a phrase.
- Document
  - an information entity that the user wants to retrieve
- Collection
  - a set of documents
- Index
  - a representation of information that makes querying easier
- Term
  - word or concept that appears in a document or a query

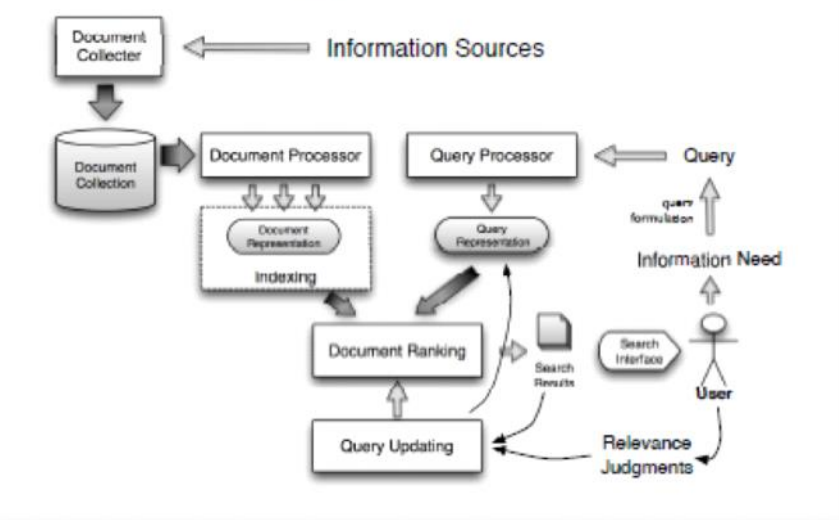
Documents

- Not just printed paper
- Can be records, pages, sites, images, people, movies
- Document encoding (Unicode)
- Document representation
- Document preprocessing (e.g., removing metadata)
- Words, terms, types, tokens

Search Engine Architecture

- Decide what to index
- Collect it
- Index it (efficiently)
- Keep the index up to date
- Provide user-friendly query facilities

Search Engine Architecture



Document Representations

- Term-document matrix (m x n) m terms, n doc
- Document-document matrix (n x n) → similarity of doc.
- Typical example in a medium-sized collection
  - n=3,000,000 documents
  - m=50,000 terms
- Typical example on the Web
  - n=30,000,000,000
  - m=1,000,000
- Boolean vs. integer-valued matrices

Storage Issues

- Imagine a medium-sized collection with n=3,000,000 and m=50,000
- How large a term-document matrix will be needed?
- Is there any way to do better? Any heuristic?

Inverted Index

- Instead of an incidence vector, use a posting table
  - VERMONT: D1, D2, D6
  - MASSACHUSETTS: D1, D5, D6, D7
- Use linked lists to be able to insert new document postings in order and to remove existing postings.
- Can be used to compute document frequency
- Keep everything sorted! This gives you a logarithmic improvement in access.

Basic Operations On Inverted Indexes

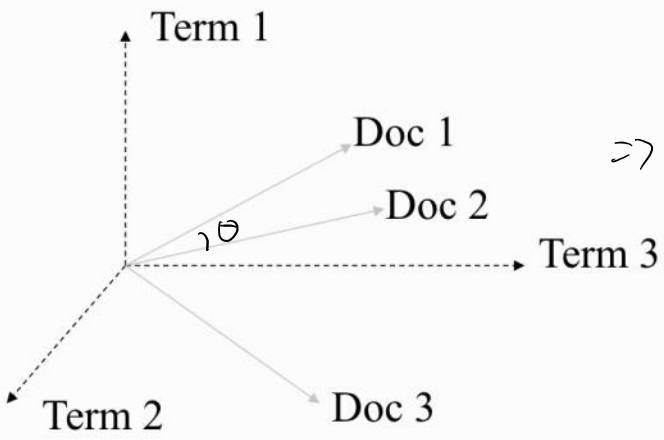
- Conjunction (AND)
  - iterative merge of the two postings: O(x+y)
- Disjunction (OR)
  - very similar
- Negation (NOT)
  - can we still do it in O(x+y)?
  - Example: VERMONT AND NOT MASSACHUSETTS
  - Example: MASSACHUSETTS OR NOT VERMONT
- Recursive operations
- Optimization
  - start with the smallest sets

The Vector Model

Term as dimension

=> can find similarity of docs.

angle



Queries as Documents

- Advantages:
  - Mathematically easier to manage
- Problems:
  - Different lengths
  - Syntactic differences
  - Repetitions of words (or lack thereof)

Take query as doc.

Vector Queries

- Each document is represented as a vector
- Non-efficient representation
- Dimensional compatibility

W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	W_{10}
C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}

The Matching Process

- Document space
- Matching is done between a document and a query (or between two documents)
- Distance vs. similarity measures.
  - Euclidean distance (define)
  - Manhattan distance (define)
  - Word overlap
  - Jaccard coefficient

10.02 Information Retrieval NATURAL LANGUAGE PROCESSING

Similarity Measures

- The Cosine measure (normalized dot product)

$$\sigma(D,Q) = \frac{|D \cap Q|}{\sqrt{|D| \cdot |Q|}} = \frac{\Sigma (d_i \cdot q_i)}{\sqrt{\Sigma (d_i)^2} \cdot \sqrt{\Sigma (q_i)^2}}$$

- The Jaccard coefficient

$$\sigma(D,Q) = \frac{|D \cap Q|}{|D \cup Q|}$$

Exercise

- Compute the cosine scores
  - $\sigma(D_1, D_2)$
  - $\sigma(D_1, D_3)$
- for the documents
  - $D_1 = <1,3>$
  - $D_2 = <100,300>$
  - $D_3 = <3,1>$
- Compute the corresponding Euclidean distances, Manhattan distances, and Jaccard coefficients.

Some of them match
Some don't.

How to deal with

Phrase-based Queries

- Examples
  - "New York City"
  - "Ann Arbor"
  - "Barack Obama"
- We don't want to match
  - York is a city in New Hampshire

Positional Indexing

- Keep track of all words and their positions in the documents
- To find a multi-word phrase, look for the matching words appearing next to each other

10.02 Information Retrieval NATURAL LANGUAGE PROCESSING

Document Ranking

- Compute the similarity between the query and each of the documents
- Use cosine similarity
- Use TF*IDF weighting → weight words
- Return the top K matches to the user

IDF: Inverse Document Frequency

- Motivation
- Example

N: number of documents
d_k: number of documents containing term k
f_{ik}: absolute frequency of term k in document i
w_{ik}: weight of term k in document i

$$\text{idf}_k = \log_2(N/d_k) + 1 = \log_2 N - \log_2 d_k + 1$$

10.02 Information Retrieval
NATURAL LANGUAGE PROCESSING

NLP

10.3

NLP

10.03 Evaluation of IR
NATURAL LANGUAGE PROCESSING

Introduction to NLP

Evaluation of IR

Evaluation

- Size of index
- Speed of indexing
- Speed of retrieval
- Accuracy
- Timeliness
- Ease of use
- Expressiveness of search language

speed.

Contingency Table

	retrieved	not retrieved	
relevant	$w=tp$ (+ true pos)	$x=fn$	$n_1 = w + x$
not relevant	$y=fp$	$z=tn$	
	$n_2 = w + y$		N

Precision and Recall

Recall: $\frac{w}{w+x}$ relevant

Precision: $\frac{w}{w+y}$ retrieved

Issues

- Why not use accuracy A=(w+z)/N?
- Average precision
- Report when P=R → Don't expand P in the cost of R.
- F measure:
  - $F=(\beta^2+1)PR/(\beta^2P+R)$
- F1 measure:
  - F1 = 2/(1/R+1/P) : harmonic mean of P and R

z >> w. or reverse. then the A is misleading.

Sample TREC query

TREC
a Text Retrieval Conference

<top>
<num> Number: 305
<title> Most Dangerous Vehicles

<desc> Description:
Which are the most crashworthy, and least crashworthy, passenger vehicles?

<narr> Narrative:
A relevant document will contain information on the crashworthiness of a given vehicle or vehicles that can be used to draw a comparison with other vehicles. The document will have to describe/compare vehicles, not drivers. For instance, it should be expected that vehicles preferred by 16-25 year-olds would be involved in more crashes, because that age group is involved in more crashes. I would view number of fatalities per 100 crashes to be more revealing of a vehicle's crashworthiness than the number of crashes per 100,000 miles, for example.
</top>

title
description
narratives

LA031689-0177	LA042790-0172
FT922-1008	LA021790-0136
LA090190-0126	LA092289-0167
LA101190-0218	LA111189-0013
LA082690-0158	LA120189-0179
LA112590-0109	LA020490-0021
FT944-136	LA122989-0063
LA020590-0119	LA091389-0119
FT944-5300	LA072189-0048
LA052190-0048	FT944-15615
LA051689-0139	LA091589-0101
FT944-9371	LA021289-0208
LA032390-0172	

10.03 Evaluation of IR
NATURAL LANGUAGE PROCESSING

<DOCNO> LA031689-0177 </DOCNO>
<DOCID> 31701 </DOCID>
<DATE><P>March 16, 1989, Thursday, Home Edition </P></DATE>
<SECTION><P>Business; Part 4; Page 1; Column 5; Financial Desk </P></SECTION>
<LENGTH><P>586 words </P></LENGTH>
<HEADLINE><P>AGENCY TO LAUNCH STUDY OF FORD BRONCO II AFTER HIGH RATE OF ROLL-OVER ACCIDENTS </P></HEADLINE>
<BYLINE><P>By LINDA WILLIAMS, Times Staff Writer </P></BYLINE>
<TEXT>
<P>The federal government's highway safety watchdog said Wednesday that the Ford Bronco II appears to be involved in more fatal roll-over accidents than other vehicles in its class and that it will seek to determine if the vehicle itself contributes to the accidents. </P>
<P>The decision to do an engineering analysis of the Ford Motor Co. utility-sport vehicle grew out of a federal accident study of the Suzuki Samurai, said Tim Hurd, a spokesman for the National Highway Traffic Safety Administration. NHTSA looked at Samurai accidents after Consumer Reports magazine charged that the vehicle had basic design flaws. </P>
<P>Several Fatalities </P>
<P>However, the accident study showed that the "Ford Bronco II appears to have a higher number of single-vehicle, first event roll-overs, particularly those involving fatalities," Hurd said. The engineering analysis of the Bronco, the second of three levels of investigation conducted by NHTSA, will cover the 1984-1989 Bronco II models, the agency said. </P>
<P>According to a Fatal Accident Reporting System study included in the September report on the Samurai, 43 Bronco II single-vehicle roll-overs caused fatalities, or 19 of every 100,000 vehicles. There were eight Samurai fatal roll-overs, or 6 per 100,000; 13 involving the Chevrolet S10 Blazers or GMC Jimmy, or 6 per 100,000, and six fatal Jeep Cherokee roll-overs, for 2.5 per 100,000. After the accident report, NHTSA declined to investigate the Samurai. </P>
...
</TEXT>
<GRAPHIC><P> Photo, The Ford Bronco II "appears to have a higher number of single-vehicle, first event roll-overs," a federal official said. </P></GRAPHIC>
<SUBJECT>
<P>TRAFFIC ACCIDENTS; FORD MOTOR CORP; NATIONAL HIGHWAY TRAFFIC SAFETY ADMINISTRATION; VEHICLE INSPECTIONS; RECREATIONAL VEHICLES; SUZUKI MOTOR CO; AUTOMOBILE SAFETY </P>
</SUBJECT>
</DOC>

TREC (cont'd)

- http://trec.nist.gov/tracks.html
- http://trec.nist.gov/presentations/presentations.html

Most Used Reference Collections

- Generic retrieval
  - OHSUMED, CRANFIELD, CACM
- Text classification
  - Reuters, 20newsgroups
- Question answering
  - TREC-QA
- Web
  - DOTGOV, wt100g
- Blogs
  - Buzzmetrics datasets
- TREC ad hoc collections, 2-6 GB
- TREC Web collections, 2-100GB

Comparing Two Systems

- Comparing A and B
- One query?
- Average performance?
- Need: A to consistently outperform B

[Example from James Allan]

The Sign Test

- Example 1:
  - A > B (12 times)
  - A = B (25 times)
  - A < B (3 times)
  - p < 0.035 (significant at the 5% level)
- Example 2:
  - A > B (18 times)
  - A < B (9 times)
  - p < 0.122 (not significant at the 5% level)
- External link:
  - http://www.fon.hum.uva.nl/Service/Statistics/Sign_Test.html

from statistics community
based on hypothesis testing tech.
Null Hypothesis: the two sys are equally good.
link to sign test calculator

Other Tests

- Student t-test: takes into account the actual performances, not just which system is better
  - http://www.fon.hum.uva.nl/Service/Statistics/Student_t_Test.html

Other Tests

- Student t-test: takes into account the actual performances, not just which system is better
  - http://www.fon.hum.uva.nl/Service/Statistics/Student_t_Test.html
  - http://www.socialresearchmethods.net/kb/stat_t.php
- Wilcoxon Matched-Pairs Signed-Ranks Test
  - http://www.fon.hum.uva.nl/Service/Statistics/Signed_Rank_Test.html

10.03 Evaluation of IR
NATURAL LANGUAGE PROCESSING

NLP

NLP

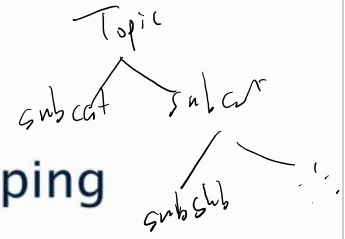
10.04 Text Classification
NATURAL LANGUAGE PROCESSING

Introduction to NLP

Text Classification

Classification

- Assigning documents to predefined categories
  - topics, languages, users
- A given set of classes C
  - Given x, determine its class in C
- Hierarchical vs. flat
- Overlapping (soft) vs non-overlapping (hard)



Classification

- Ideas: manual classification using rules
  - e.g., Columbia AND University → Education
    Columbia AND "South Carolina" → Geography
- Popular techniques
  - generative (k-nn, Naïve Bayes) vs. discriminative (SVM, regression)
- Generative
  - model joint prob p(x,y) and use Bayesian prediction to compute p(y|x)
- Discriminative
  - model p(y|x) directly.

Representations For Document Classification (And Clustering)

- Typically: vector-based
  - Words: "cat", "dog", etc.
  - Features: document length, author name, etc.
- Each document is represented as a vector in an *n*-dimensional space
- Similar documents appear nearby in the vector space (distance measures are needed)

Naïve Bayesian classifiers

- Naïve Bayesian classifier

$$P(d \in C \mid F_1, F_2, \ldots F_k) = \frac{P(F_1, F_2, \ldots F_k \mid d \in C) P(d \in C)}{P(F_1, F_2, \ldots F_k)}$$

- Assuming statistical independence

$$P(d \in C \mid F_1, F_2, \ldots F_k) = \frac{\prod_{j=1}^{k} P(F_j \mid d \in C) P(d \in C)}{\prod_{j=1}^{k} P(F_j)}$$

- Features = words (or phrases) typically

Issues with Naïve Bayes

- Where do we get the values $P(d \in C)$
  - use maximum likelihood estimation (N_i/N)
- Same for the conditionals
  - these are based on a multinomial generator and the MLE estimator is ($T_{ji}/\Sigma T_{ji}$)
- Smoothing is needed
  - why
  - Laplace smoothing ($(T_{ji}+1)/\Sigma(T_{ji}+1)$)
- Implementation
  - how to avoid floating point underflow

take log

Spam Recognition

Return-Path: <ig_esq@rediffmail.com>
X-Sieve: CMU Sieve 2.2
From: "Ibrahim Galadima" <ig_esq@rediffmail.com>
Reply-To: galadima_esq@netpiper.com
To: webmaster@aclweb.org
Subject: Gooday

DEAR SIR

FUNDS FOR INVESTMENTS

THIS LETTER MAY COME TO YOU AS A SURPRISE SINCE I HAD NO PREVIOUS CORRESPONDENCE WITH YOU

I AM THE CHAIRMAN TENDER BOARD OF INDEPENDENT NATIONAL ELECTORAL COMMISSION INEC I GOT YOUR CONTACT IN THE COURSE OF MY SEARCH FOR A RELIABLE PERSON WITH WHOM TO HANDLE A VERY CONFIDENTIAL TRANSACTION INVOLVING THE ! TRANSFER OF FUND VALUED AT TWENTY ONE MILLION SIX HUNDRED THOUSAND UNITED STATES DOLLARS US$20M TO A SAFE FOREIGN ACCOUNT

SpamAssassin

- http://spamassassin.apache.org/
- http://spamassassin.apache.org/tests_3_3_x.html
- Examples:
  - body Incorporates a tracking ID number
  - body HTML and text parts are different
  - header Date: is 3 to 6 hours before Received: date
  - body HTML font size is huge
  - header Attempt to obfuscate words in Subject:
  - header Subject =~ /^urgent(?:[\s\W]*(dollar) | .{1,40}(?:alert| response| assistance| proposal| reply| warning| noti(?:ce| fication)| greeting| matter))/i

Feature Selection: The X^2 Test

- For a term *t*:

		I_t	
		0	1
C	0	k_{00}	k_{01}
	1	k_{10}	k_{11}

- C=class, i_t = feature
- Testing for independence: $P(C=0,I_t=0)$ should be equal to $P(C=0)\ P(I_t=0)$
  - $P(C=0) = (k_{00}+k_{01})/n$
  - $P(C=1) = 1-P(C=0) = (k_{10}+k_{11})/n$
  - $P(I_t=0) = (k_{00}+K_{10})/n$
  - $P(I_t=1) = 1-P(I_t=0) = (k_{01}+k_{11})/n$

Feature Selection: The X^2 Test

$$X^2 = \frac{n(k_{11}k_{00} - k_{10}k_{01})^2}{(k_{11}+k_{10})(k_{01}+k_{00})(k_{11}+k_{01})(k_{10}+k_{00})}$$

- High values of X^2 indicate lower belief in independence.
- In practice, compute X^2 for all words and pick the top k among them.

10.04 Text Classification NATURAL LANGUAGE PROCESSING

Feature Selection: Mutual Information

- No document length scaling is needed
- Documents are assumed to be generated according to the multinomial model
- Measures amount of information: if the distribution is the same as the background distribution, then MI=0
- X = word; Y = class

$$MI(X,Y) = \sum_x \sum_y P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

Well-known Datasets

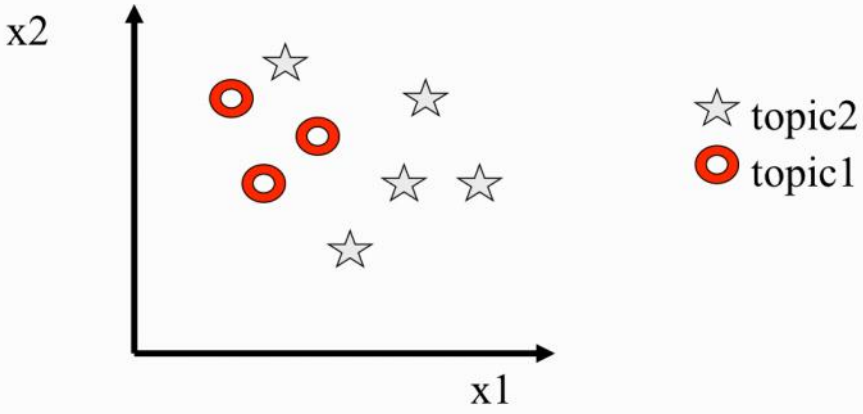
- 20 newsgroups
  - http://qwone.com/~jason/20Newsgroups/
- Reuters-21578
  - http://www.daviddlewis.com/resources/testcollections/reuters21578/
  - Cats: grain, acquisitions, corn, crude, wheat, trade...
- WebKB
  - http://www-2.cs.cmu.edu/~webkb/
  - course, student, faculty, staff, project, dept, other
- RCV1
  - http://www.daviddlewis.com/resources/testcollections/rcv1/
  - Larger Reuters corpus

Evaluation Of Text Classification

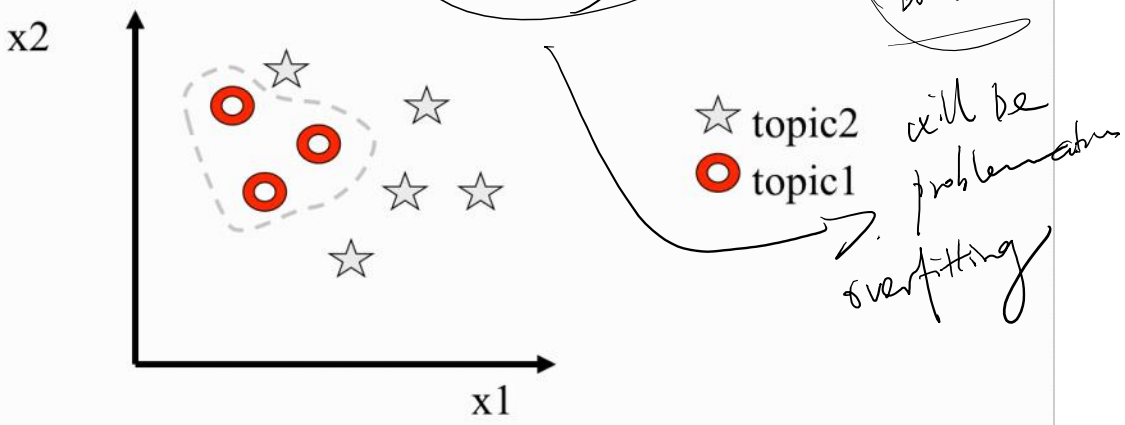
- Microaveraging
  - average over classes
- Macroaveraging
  - uses pooled table

pooled table mentioned before (I think)

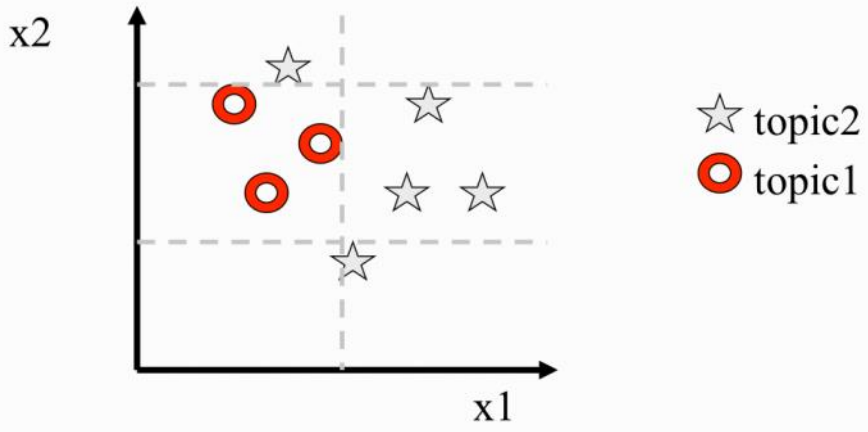
Vector Space Classification



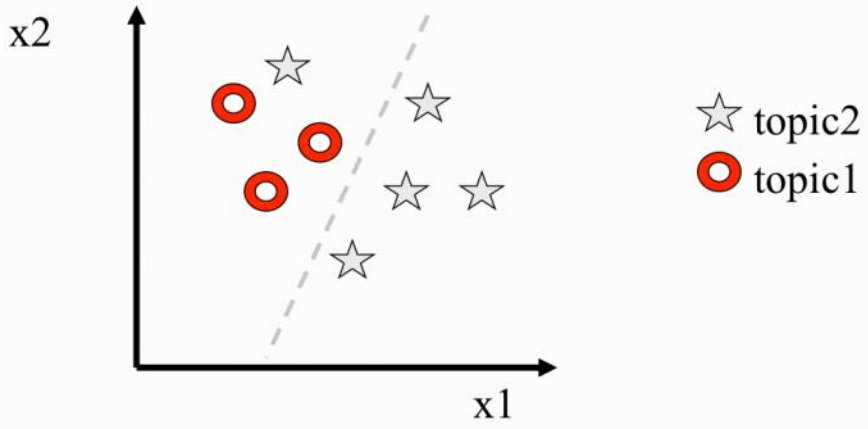
Decision Surfaces



Decision Trees



Linear Boundary



Vector Space Classifiers

build centroids for each class.

- Using centroids
- Boundary
  - line that is equidistant from two centroids

10.04 Text Classification NATURAL LANGUAGE PROCESSING

Linear Separators

- Two-dimensional line:
  $w_1x_1+w_2x_2=b$ is the linear separator
  $w_1x_1+w_2x_2>b$ for the positive class
- In n-dimensional spaces:

$$\vec{w}^T\vec{x} = b$$

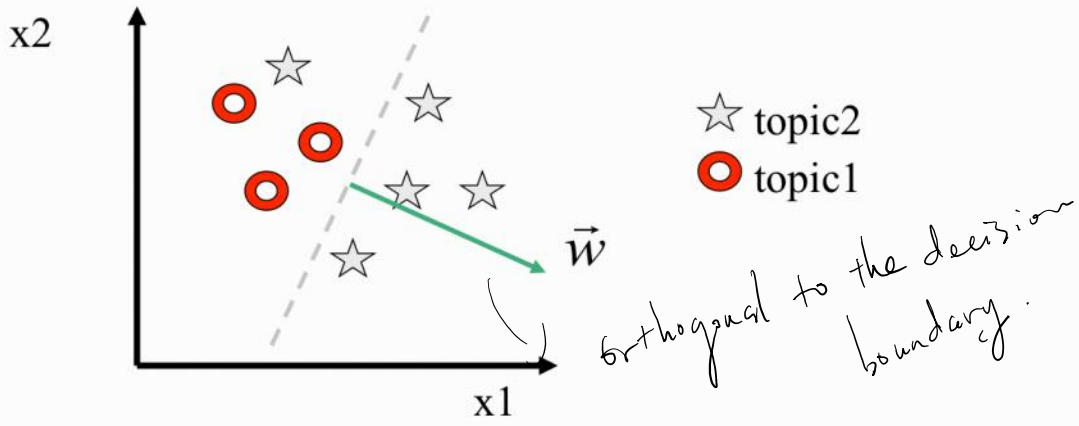
w_1, w_2 weight

b — bias

if $b=0$ — no prior distinction between two classes.

$< b$ negative class

Decision Boundary



Example

- Bias b=0
- Document is "A D E H"
- Its score will be

  $0.6*1+0.4*1+0.4*1+(-0.5)*1 = 0.9>0$

w_i	x_i	w_i	x_i
0.6	A	-0.7	G
0.5	B	-0.5	H
0.5	C	-0.3	I
0.4	D	-0.2	J
0.4	E	-0.2	K
0.3	F	-0.2	L

Perceptron Algorithm

ML

Perceptron Algorithm

optimize weight

Input: $S = ((\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)), \vec{x}_i \in \Re^N, y_i \in \{-1, 1\}$

$\eta \in \Re$ → Determine how fast to converge

Algorithm:

$\vec{w}_0 = \vec{0}, k = 0$

FOR $i = 1$ TO n

IF $y_i(\vec{w}_k \bullet \vec{x}_i) \le 0$ (y_i has a diff sign)

$\vec{w}_{k+1} = \vec{w}_k + \eta y_i \vec{x}_i$

$k = k + 1$

END

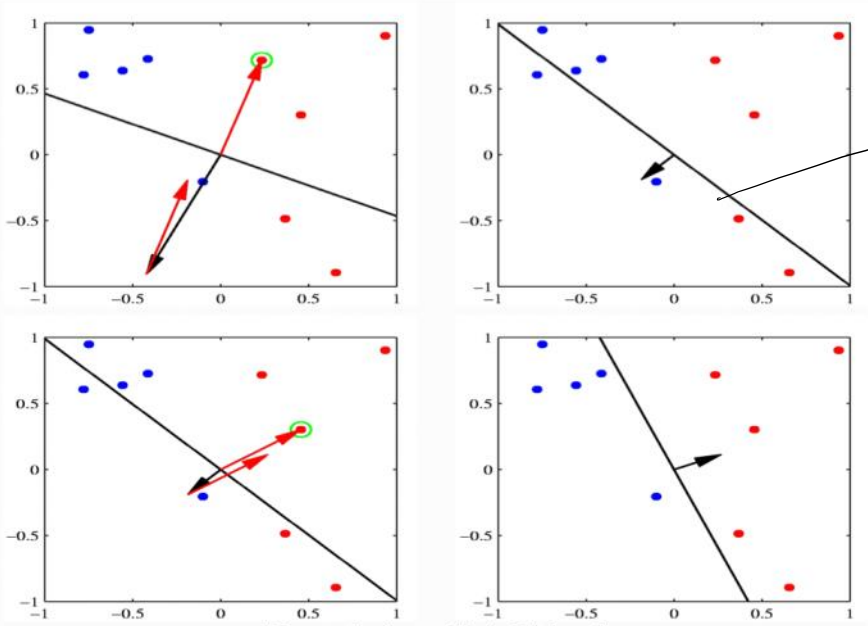
END

(the diff sign) ↓ stops after 1a iteration

goal: learn weights

Output: $\vec{w}_k$

10.04 Text Classification NATURAL LANGUAGE PROCESSING



[Example from Chris Bishop]

Add the misclassified dot to the decision boundary each step

Generative Models: knn

- Assign each element to the closest cluster
- K-nearest neighbors

$$score(c, d_q) = b_c + \sum_{d \in kNN(d_q)} s(d_q, d)$$

- Very easy to program
- Issues:
  - choosing k, b?
- Demo:
  - http://www-2.cs.cmu.edu/~zhuxj/courseproject/knndemo/KNN.html

10.04 Text Classification NATURAL LANGUAGE PROCESSING

NLP

NLP

10.05 Text Clustering
NATURAL LANGUAGE PROCESSING

Introduction to NLP

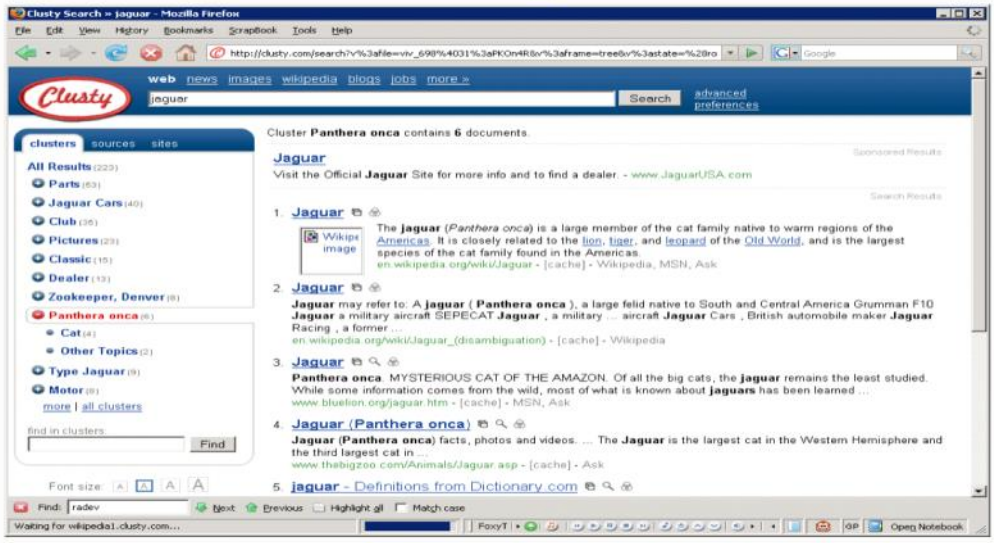
Text Clustering

Clustering

- Exclusive/overlapping clusters
- Hierarchical/flat clusters

- The cluster hypothesis
  - Documents in the same cluster are relevant to the same query
  - How do we use it in practice?

Example



k-means

know # of clusters in advance.

- Iteratively determine which cluster a point belongs to, then adjust the cluster centroid, then repeat
- Needed: small number k of desired clusters
- hard decisions

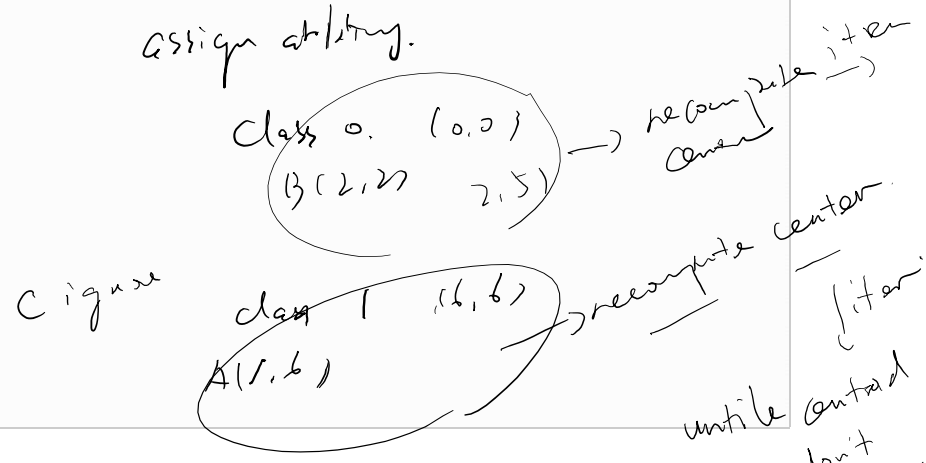
10.05 Text Clustering NATURAL LANGUAGE PROCESSING

k-means

1 initialize cluster centroids to arbitrary vectors
2 **while** further improvement is possible **do**
3 **for** each document *d* **do**
4 find the cluster *c* whose centroid is ***closest*** to *d*
5 assign *d* to cluster *c*
6 **end for**
7 **for** each cluster *c* **do**
8 recompute the centroid of cluster *c* based on its documents
9 **end for**
10 **end while**

Example

- Cluster the following vectors into two groups:
  - A = <1,6>
  - B = <2,2>
  - C = <4,0>
  - D = <3,3>
  - E = <2,5>
  - F = <2,1>



Demos

- http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html
- http://cgm.cs.mcgill.ca/~godfried/student_projects/bonnef_k-means
- http://www.cs.washington.edu/research/imagedatabase/demo/kmcluster
- http://www.cc.gatech.edu/~dellaert/FrankDellaert/Software.html
- http://www-2.cs.cmu.edu/~awm/tutorials/kmeans11.pdf
- http://web.archive.org/web/20110223234358/http://www.ece.neu.edu/groups/rpl/projects/kmeans/

Evaluation of Clustering

- Purity
  - considering the majority class in each cluster
- RAND index
  - See next slide

Purity

- Three clusters

XXXOO 3/5

OOOX% 3/5

%%%%XX 4/6

- Purity: overall
  - (3+3+4)/16=62.5%

read extend

See as a series of decision about pairs of items

True Positive: correctly assigns a pair that belong to same cluster, and to the same cluster

Rand Index

- Accuracy when preserving object-object relationships.
- RI=(TP+TN)/(TP+FP+FN+TN)
- In the example:

get point when 2 obj that are truly in one cluster assigned same

- RI=(TP+TN)/(TP+FP+FN+TN)
- In the example:

$$TP + FP = \binom{5}{2} + \binom{5}{2} + \binom{6}{2} = 35$$

$$TP = \binom{3}{2} + \binom{3}{2} + \binom{4}{2} + \binom{2}{2} = 13$$

$$FP = 35 - 13 = 22$$

Rand Index

	Same cluster	
Same class	TP=13	FN=21
	FP=22	TN=64

RI = (TP+TN)/(TP+TN+FP+FN)=(13+64)/(13+64+22+21)=0.64

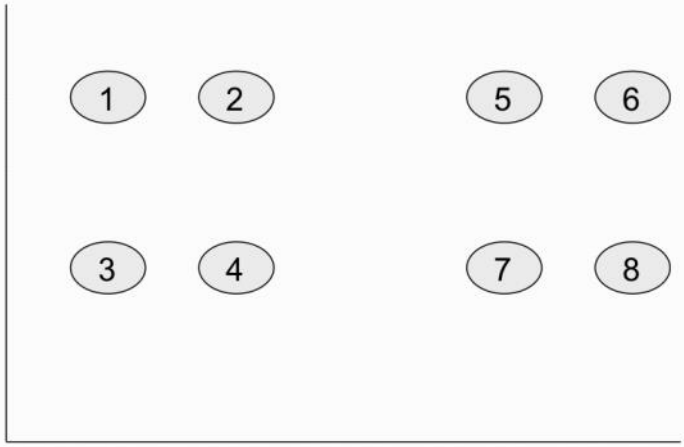
10.05 Text Clustering NATURAL LANGUAGE PROCESSING

Hierarchical Clustering Methods

- Single-linkage
  - One common pair is sufficient
  - disadvantages: long chains
- Complete-linkage
  - All pairs have to match
  - Disadvantages: too conservative
- Average-linkage

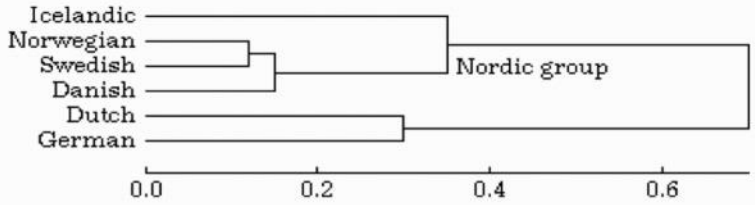
- Average-linkage

Hierarchical Clustering



Hierarchical Agglomerative Clustering Dendrograms

E.g., language similarity:



http://odur.let.rug.nl/~kleiweg/clustering/clustering.html

Clustering Using Dendrograms

Example: cluster the following sentences:

A B C B A
A D C C A D E
C D E F C D A
E F G F D A
A C D A B A

REPEAT
- Compute pairwise similarities
- Identify closest pair
- Merge pair into single node

UNTIL only one node left

Q: what is the equivalent Venn diagram representation?

10.05 Text Clustering
NATURAL LANGUAGE PROCESSING

NLP

10.6

10.06 Information Retrieval Toolkits
NATURAL LANGUAGE PROCESSING

Introduction to NLP

Information Retrieval Toolkits

Open Source IR Toolkits

- Smart (Cornell)
- MG (RMIT & Melbourne, Australia; Waikato, New Zealand),
- Lemur (CMU/Univ. of Massachusetts)
- Terrier (Glasgow)
- Clairlib (University of Michigan)
- Lucene/SOLR (Apache)

10.06 Information Retrieval Toolkits
NATURAL LANGUAGE PROCESSING

Smart

- The most influential IR system/toolkit
- Developed at Cornell since 1960's
- Vector space model with lots of weighting options
- Written in C
- The Cornell/AT&T groups have used the Smart system to achieve top TREC performance

MG

- A highly efficient toolkit for retrieval of text and images
- Developed by people at Univ. of Waikato, Univ. of Melbourne, and RMIT in 1990's
- Written in C, running on Unix
- Vector space model with lots of compression and speed up tricks
- People have used it to achieve good TREC performance

10.06 Information Retrieval Toolkits
NATURAL LANGUAGE PROCESSING

Lemur/Indri

- An IR toolkit emphasizing language models
- Developed at CMU and Univ. of Massachusetts in 2000's
- Written in C++, highly extensible
- Vector space and probabilistic models including language models
- Achieving good TREC performance with a simple language model

Lucene

- Open Source IR toolkit
- Initially developed by Doug Cutting in Java
- Now has been ported to some other languages
- Good for building IR/Web applications
- Many applications have been built using Lucene (e.g., Nutch and SOLR)

10.06 Information Retrieval Toolkits
NATURAL LANGUAGE PROCESSING

NLP