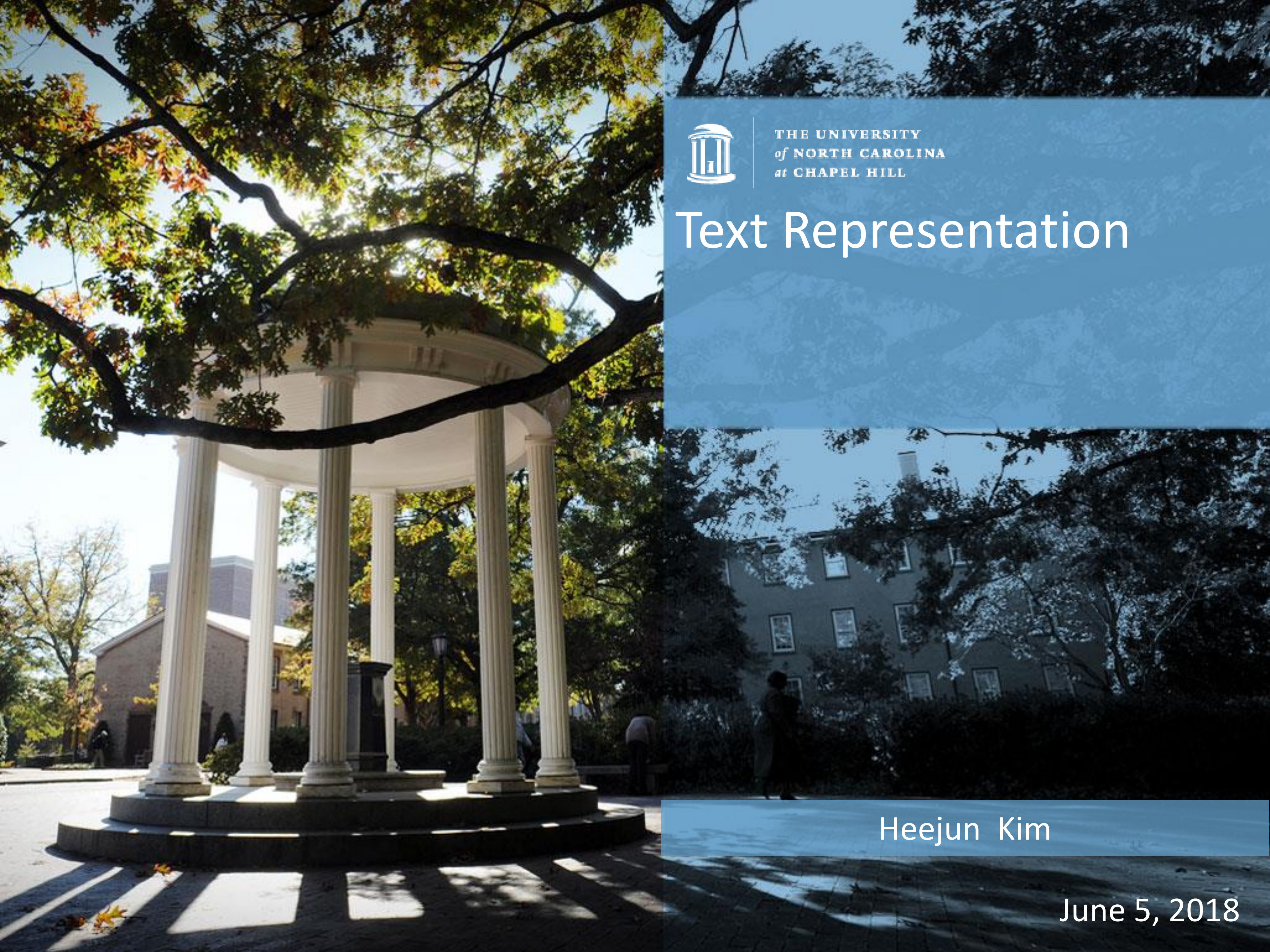


THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Text Representation

Heejun Kim

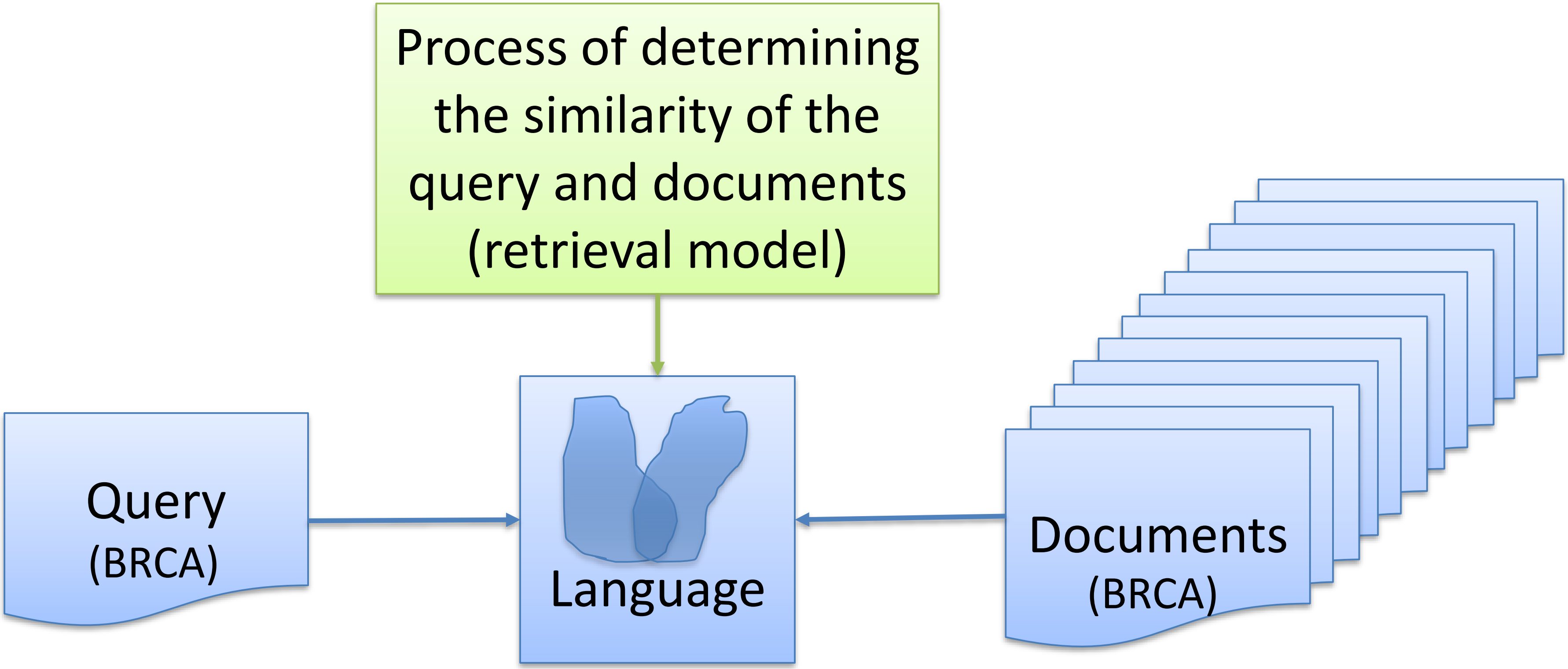
June 5, 2018

Outline for Text Representation

- Bag of Words Representation
- Indexing
- Text Processing
- Vector Space Model
- Term Weighting

Bag of Words Representation

How to Find a Relevant Documents?



Process of determining the similarity of the query and documents (retrieval model)

Query (BRCA)

Language

Documents (BRCA)

Topic Classification

BRCA

- Doc1: Genotype-Phenotype Correlations in BRCA Mutation Carriers
- Doc2: Breast cancer following ovarian cancer in BRCA mutation carriers
- Doc3: Breast cancer, BRCA mutations, and attitudes regarding pregnancy
- Doc4: Surgical management of breast cancer in BRCA-mutation carriers
- Doc5: Cancer risk management decision making for BRCA women

- Doc6: Inverse association between cancer and neurodegenerative disease
- Doc7: Molecular neurodegeneration: basic biology and disease pathways
- Doc8: Mechanisms of neurodegeneration and axonal dysfunction
- Doc9: Dysfunction of neuronal calcium signaling in neuroinflammation and neurodegeneration
- Doc10: Epigenetic mechanisms of neurodegeneration in Huntington's disease

Neurodegeneration

Bag of Words Text Representation

Genotype-Phenotype Correlations in BRCA Mutation Carriers

Breast cancer following ovarian cancer in BRCA mutation carriers

Breast cancer, BRCA mutations, and attitudes regarding pregnancy

Surgical management of breast cancer in BRCA-mutation carriers

Cancer risk management decision making for BRCA women

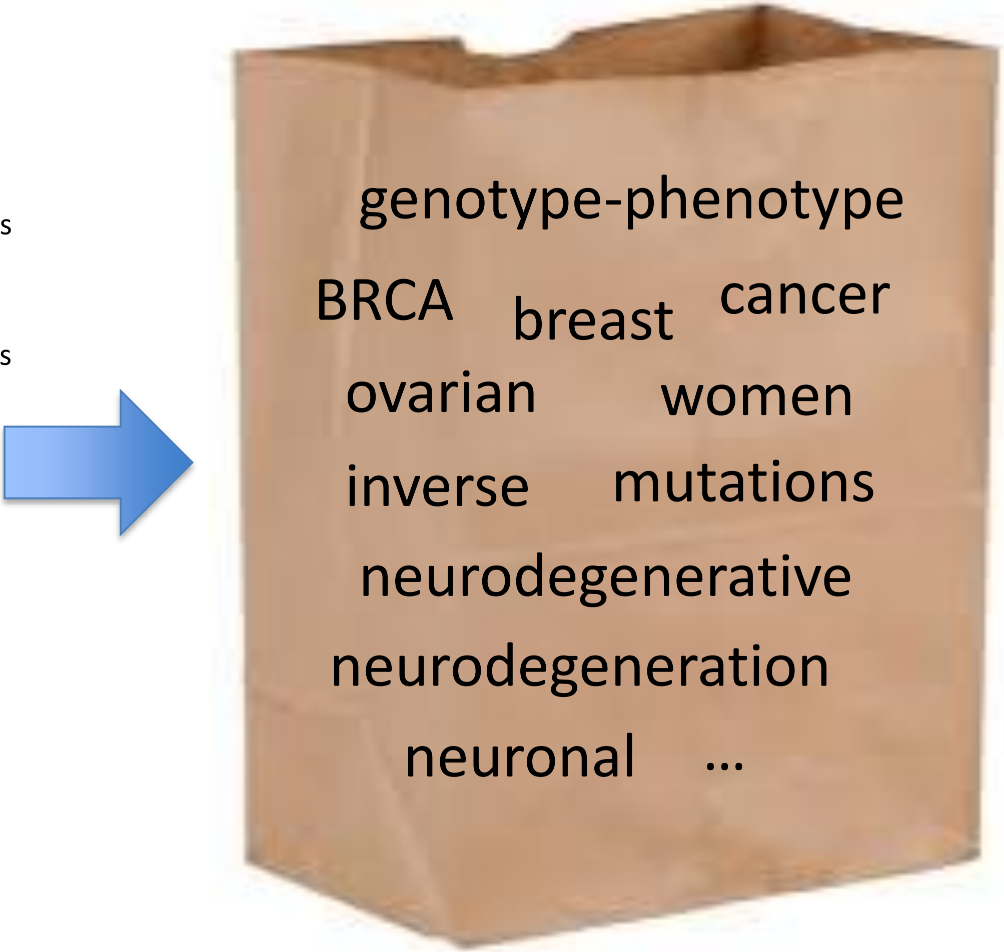
Inverse association between cancer and neurodegenerative disease

Molecular neurodegeneration: basic biology and disease pathways

Mechanisms of neurodegeneration and axonal dysfunction

Dysfunction of neuronal calcium signaling in neuroinflammation and neurodegeneration

Epigenetic mechanisms of neurodegeneration in Huntington's disease



genotype-phenotype

BRCA breast cancer

ovarian women

inverse mutations

neurodegenerative

neurodegeneration

neuronal …

Bag of Words Text Representation

- Features correspond to terms in the vocabulary
  - vocabulary: the set of distinct terms appearing in at least one training instance
  - remember that all training instances and all test instances must have the same representation!
- Position information and word order is lost
  - BRCA mutation carriers = mutation BRCA carriers
- Simple, but often effective

Tokenization

- Token
  - A unit of text analysis. Usually a word or other atomic parse element (i.e., symbol, term, etc.) between white spaces
- Tokenization
  - Splitting text into terms of tokens

Text Representation

features

concept

instances

	brca	breast	cancer	mutation	neuro degeneration	neuronal	neuro degenerative	Label
Doc1	1	0	0	1	0	0	0	BRCA
Doc2	1	1	2	1	0	0	0	BRCA
Doc3	1	1	1	0	0	0	0	BRCA
Doc4	0	1	1	1	0	0	0	BRCA
Doc5	1	0	1	0	0	0	0	BRCA
Doc6	0	0	1	0	0	0	1	AD
Doc7	0	0	0	0	1	0	0	AD
Doc8	0	0	0	0	1	0	0	AD
Doc9	0	0	0	0	1	1	0	AD
Doc10	0	0	0	0	1	0	0	AD

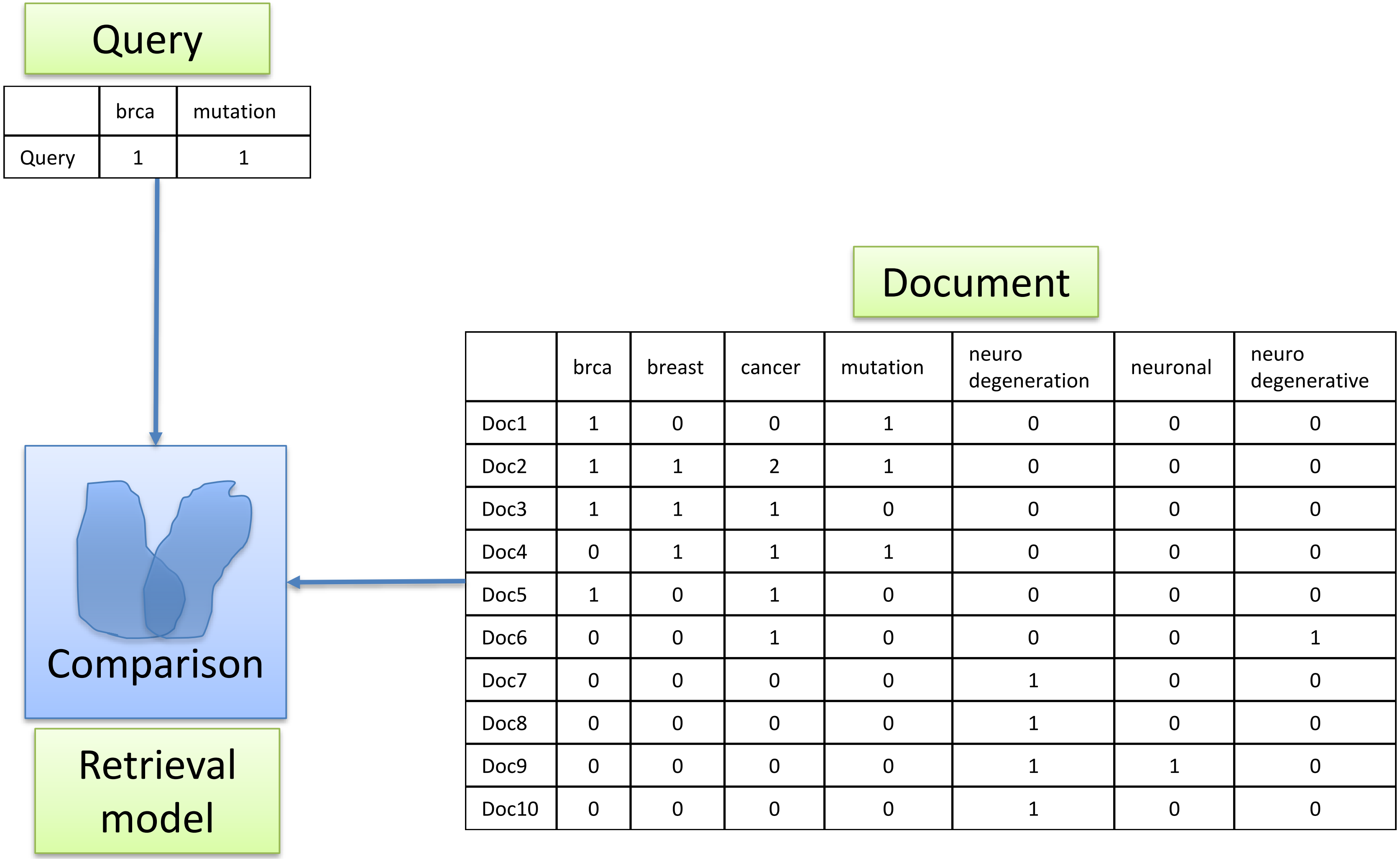
* AD stands for Alzheimer’s disease.

Indexing

Indexing

- Index: facilitates quickly finding the documents that match the query
- Query language: defines how users can describe their information needs to the system (e.g., boolean queries)
- Document representation: determines what goes in the index (e.g., term-occurrences, term-frequencies, etc.)
- Retrieval model: decides whether a document is relevant to the query (and possibly its degree of relevance)

Indexing



Query

	brca	mutation
Query	1	1

Document

	brca	breast	cancer	mutation	neuro degeneration	neuronal	neuro degenerative
Doc1	1	0	0	1	0	0	0
Doc2	1	1	2	1	0	0	0
Doc3	1	1	1	0	0	0	0
Doc4	0	1	1	1	0	0	0
Doc5	1	0	1	0	0	0	0
Doc6	0	0	1	0	0	0	1
Doc7	0	0	0	0	1	0	0
Doc8	0	0	0	0	1	0	0
Doc9	0	0	0	0	1	1	0
Doc10	0	0	0	0	1	0	0

Comparison

Retrieval model

McGraw-Hill Computer Science Series

Ahuja: Design and Analysis of Computer Communication Networks
Allen: Anatomy of LISP
Barbacci and Siewiorek: The Design and Analysis of Instruction Set Processors
Bell and Newell: Computer Structures: Readings and Examples
Donovan: Systems Programming
Gear: Computer Organization and Programming
Givone: Introduction to Switching Circuit Theory
Goodman and Hedetniemi: Introduction to the Design and Analysis of Algorithms
Hamacher, Vranesic, and Zaky: Computer Organization
Hamming: Introduction to Applied Numerical Analysis
Hayes: Computer Architecture and Organization
Hellerman: Digital Computer System Principles
Hellerman and Conroy: Computer System Performance
Katzan: Microprogramming Primer
Keller: A First Course in Computer Programming Using PASCAL
Liu: Elements of Discrete Mathematics
Liu: Introduction to Combinatorial Mathematics
MacEwen: Introduction to Computer Systems: Using the PDP-11 and Pascal
Madnick and Donovan: Operating Systems
Manna: Mathematical Theory of Computation
Newman and Sproull: Principles of Interactive Computer Graphics
Nilsson: Problem-Solving Methods in Artificial Intelligence
Payne: Introduction to Simulation: Programming Techniques and Methods of Analysis
Rice: Matrix Computations and Mathematical Software
Salton and McGill: Introduction to Modern Information Retrieval
Shooman: Software Engineering: Design, Reliability, and Management
Siewiorek, Bell, and Newell: Computer Structures: Principles and Examples
Stone: Introduction to Computer Organization and Data Structures
Stone and Siewiorek: Introduction to Computer Organization and Data Structures: PDP-11 Edition
Tonge and Feldman: Computing: An Introduction to Procedures and Procedure-Followers
Tremblay and Bunt: An Introduction to Computer Science: An Algorithmic Approach
Tremblay and Bunt: An Introduction to Computer Science: An Algorithmic Approach, Short Edition
Tremblay and Manohar: Discrete Mathematical Structures with Applications to Computer Science
Tremblay and Sorenson: An Introduction to Data Structures with Applications
Tucker: Programming Languages
Wiederhold: Database Design

McGraw-Hill Advanced Computer Science Series

Davis and Lenat: Knowledge-Based Systems in Artificial Intelligence
Kogge: The Architecture of Pipelined Computers
Lindsay, Buchanan, Feigenbaum, and Lederberg: Applications of Artificial Intelligence for Organic Chemistry: The Dendral Project
Nilsson: Problem-Solving Methods in Artificial Intelligence
Wulf, Levin, and Harbison: Hydra/C.mmp: An Experimental Computer System

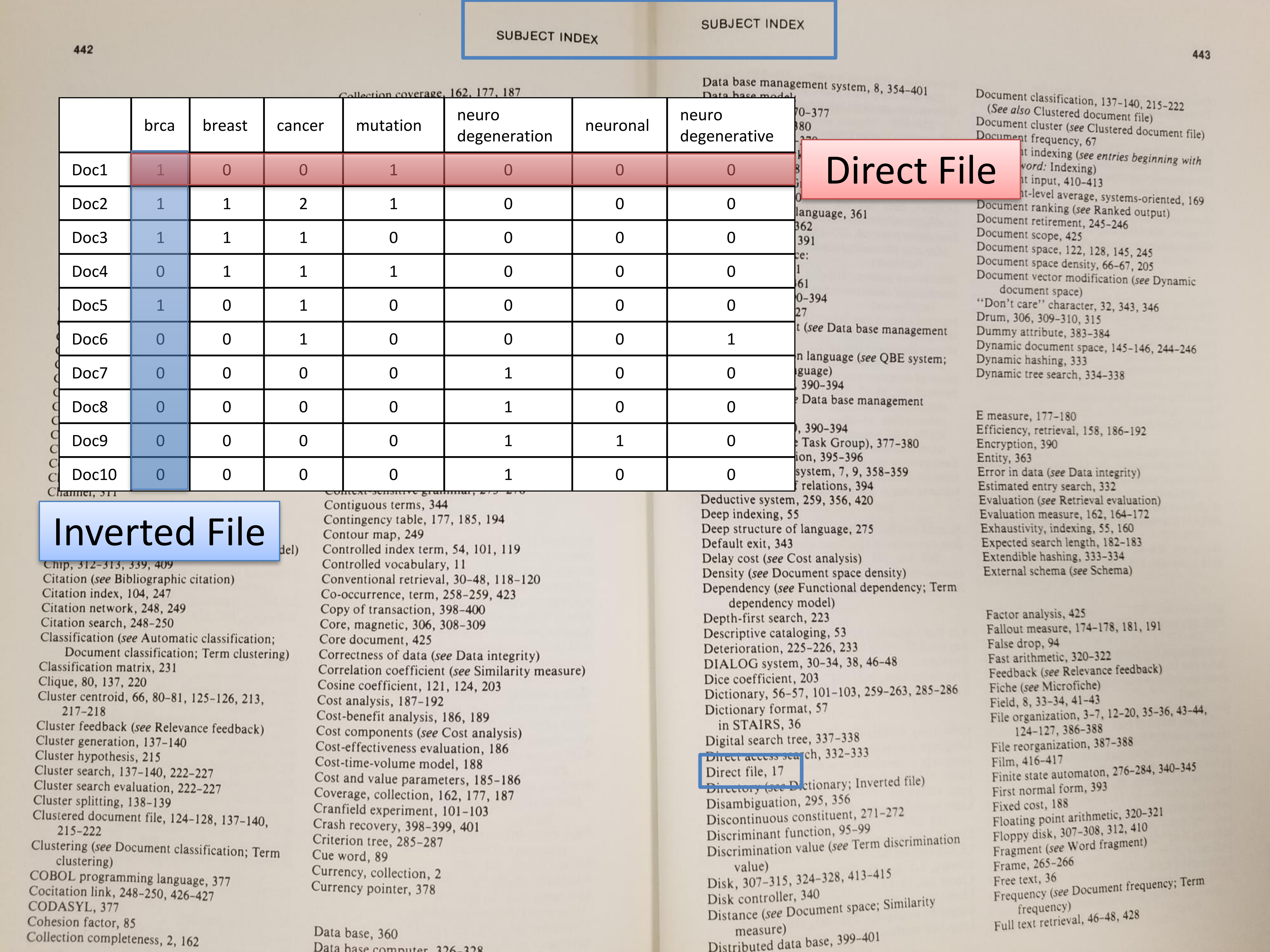
Introduction to Modern Information Retrieval

Gerard Salton
Professor of Computer Science
Cornell University

Michael J. McGill
Associate Professor of Information Studies
Syracuse University

McGraw-Hill Book Company

New York St. Louis San Francisco Auckland Bogotá Hamburg
Johannesburg London Madrid Mexico Montreal New Delhi
Panama Paris São Paulo Singapore Sydney Tokyo Toronto



	brca	breast	cancer	mutation	neuro degeneration	neuronal	neuro degenerative
Doc1	1	0	0	1	0	0	0
Doc2	1	1	2	1	0	0	0
Doc3	1	1	1	0	0	0	0
Doc4	0	1	1	1	0	0	0
Doc5	1	0	1	0	0	0	0
Doc6	0	0	1	0	0	0	1
Doc7	0	0	0	0	1	0	0
Doc8	0	0	0	0	1	0	0
Doc9	0	0	0	0	1	1	0
Doc10	0	0	0	0	1	0	0

Direct File

Inverted File

Chip, 312–313, 339, 409
Citation (*see* Bibliographic citation)
Citation index, 104, 247
Citation network, 248, 249
Citation search, 248–250
Classification (*see* Automatic classification;
Document classification; Term clustering)
Classification matrix, 231
Clique, 80, 137, 220
Cluster centroid, 66, 80–81, 125–126, 213,
217–218
Cluster feedback (*see* Relevance feedback)
Cluster generation, 137–140
Cluster hypothesis, 215
Cluster search, 137–140, 222–227
Cluster search evaluation, 222–227
Cluster splitting, 138–139
Clustered document file, 124–128, 137–140,
215–222
Clustering (*see* Document classification; Term
clustering)
COBOL programming language, 377
Cocitation link, 248–250, 426–427
CODASYL, 377
Cohesion factor, 85
Collection completeness, 2, 162

Collection coverage, 162, 177, 187

Contiguous terms, 344
Contingency table, 177, 185, 194
Contour map, 249
Controlled index term, 54, 101, 119
Controlled vocabulary, 11
Conventional retrieval, 30–48, 118–120
Co-occurrence, term, 258–259, 423
Copy of transaction, 398–400
Core, magnetic, 306, 308–309
Core document, 425
Correctness of data (*see* Data integrity)
Correlation coefficient (*see* Similarity measure)
Cosine coefficient, 121, 124, 203
Cost analysis, 187–192
Cost-benefit analysis, 186, 189
Cost components (*see* Cost analysis)
Cost-effectiveness evaluation, 186
Cost-time-volume model, 188
Cost and value parameters, 185–186
Coverage, collection, 162, 177, 187
Cranfield experiment, 101–103
Crash recovery, 398–399, 401
Criterion tree, 285–287
Cue word, 89
Currency, collection, 2
Currency pointer, 378

Data base, 360

Data base management system, 8, 354–401

language, 361

t (*see* Data base management

n language (*see* QBE system;

, 390–394

Data base management

, 390–394
Task Group), 377–380
ion, 395–396
system, 7, 9, 358–359
relations, 394
Deductive system, 259, 356, 420
Deep indexing, 55
Deep structure of language, 275
Default exit, 343
Delay cost (*see* Cost analysis)
Density (*see* Document space density)
Dependency (*see* Functional dependency; Term
dependency model)
Depth-first search, 223
Descriptive cataloging, 53
Deterioration, 225–226, 233
DIALOG system, 30–34, 38, 46–48
Dice coefficient, 203
Dictionary, 56–57, 101–103, 259–263, 285–286
Dictionary format, 57
in STAIRS, 36
Digital search tree, 337–338
Direct access search, 332–333
Direct file, 17
Directory (*see* Dictionary; Inverted file)
Disambiguation, 295, 356
Discontinuous constituent, 271–272
Discriminant function, 95–99
Discrimination value (*see* Term discrimination
value)
Disk, 307–315, 324–328, 413–415
Disk controller, 340
Distance (*see* Document space; Similarity
measure)
Distributed data base, 399–401

Document classification, 137–140, 215–222
(*See also* Clustered document file)
Document cluster (*see* Clustered document file)
Document frequency, 67
t indexing (*see entries beginning with*
word: Indexing)
t input, 410–413
t-level average, systems-oriented, 169
Document ranking (*see* Ranked output)
Document retirement, 245–246
Document scope, 425
Document space, 122, 128, 145, 245
Document space density, 66–67, 205
Document vector modification (*see* Dynamic
document space)
"Don't care" character, 32, 343, 346
Drum, 306, 309–310, 315
Dummy attribute, 383–384
Dynamic document space, 145–146, 244–246
Dynamic hashing, 333
Dynamic tree search, 334–338

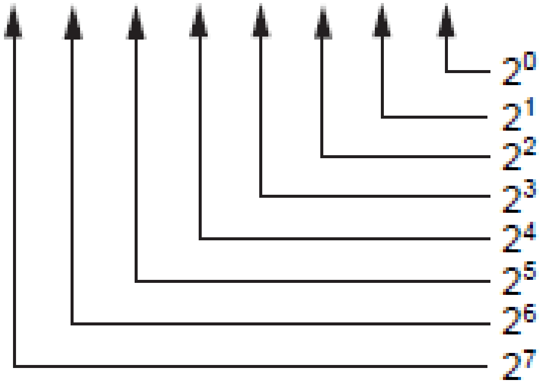
E measure, 177–180
Efficiency, retrieval, 158, 186–192
Encryption, 390
Entity, 363
Error in data (*see* Data integrity)
Estimated entry search, 332
Evaluation (*see* Retrieval evaluation)
Evaluation measure, 162, 164–172
Exhaustivity, indexing, 55, 160
Expected search length, 182–183
Extendible hashing, 333–334
External schema (*see* Schema)

Factor analysis, 425
Fallout measure, 174–178, 181, 191
False drop, 94
Fast arithmetic, 320–322
Feedback (*see* Relevance feedback)
Fiche (*see* Microfiche)
Field, 8, 33–34, 41–43
File organization, 3–7, 12–20, 35–36, 43–44,
124–127, 386–388
File reorganization, 387–388
Film, 416–417
Finite state automaton, 276–284, 340–345
First normal form, 393
Fixed cost, 188
Floating point arithmetic, 320–321
Floppy disk, 307–308, 312, 410
Fragment (*see* Word fragment)
Frame, 265–266
Free text, 36
Frequency (*see* Document frequency; Term
frequency)
Full text retrieval, 46–48, 428

How a Computer Stores Data?

- Computers store data in **binary** format
  - A **binary digit** has two possible values: **0** or **1**
- Binary digits are called **bits**
- The values of binary digits are powers of 2

1 1 1 1 1 1 1 1

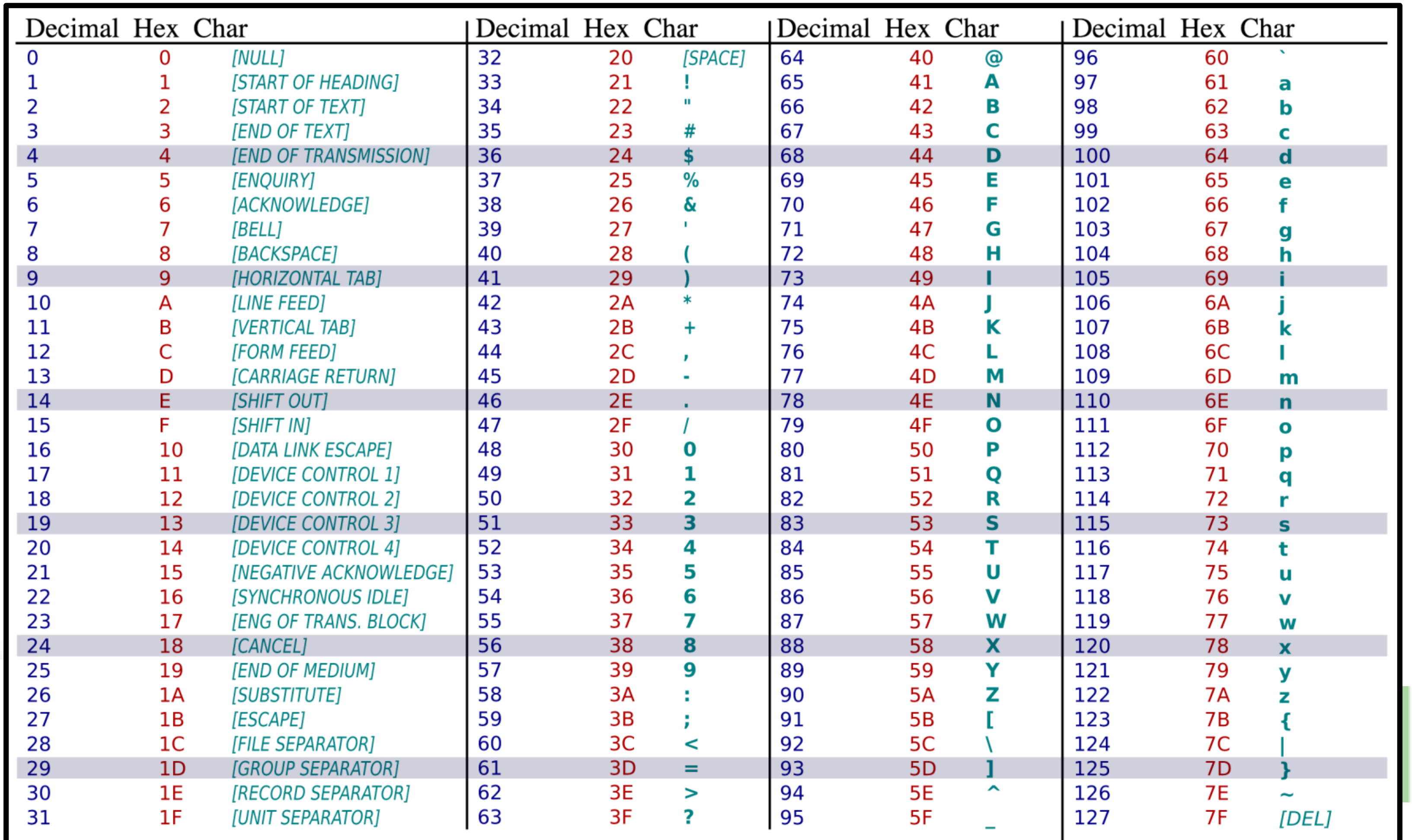


How a Computer Stores Data?

- Bits are grouped 8-at-a-time to form **bytes**
- **00000000 = 0**
- **00000001 = 1**
- **00000010 = 2**
- **00000011 = 3**
- …

What about Text?

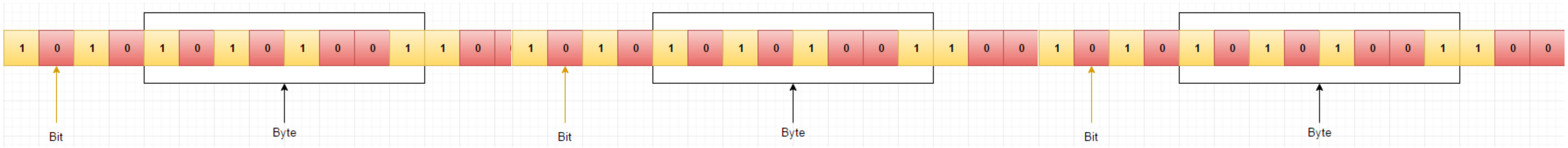
- Each character is mapped to an integer
- e.g., ASCII: 7 bits per character (128 unique codes)

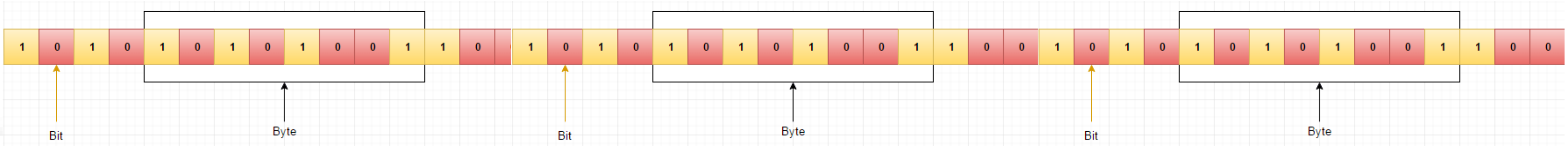


Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

How a Computer Stores Data?

- Computers cannot use intuitive multi-dimensional structure (e.g., table). Data should be sequentially stored in memory.
- Data in computers are like a list of lists.





How a Computer Stores Data?

	brca	breast	cancer	mutation	neuro degeneration	neuronal	neuro degenerative
Doc1	1	0	0	1	0	0	0
Doc2	1	1	2	1	0	0	0
Doc3	1	1	1	[illegible]	[illegible]	0	0
Doc4	0	1	1	[illegible]	0	0	0
Doc5	1	0	1	[illegible]	0	0	0
Doc6	0	0	1	[illegible]	[illegible]	0	1
Doc7	0	0	0	0	1	0	0
Doc8	0	0	0	0	1	0	0
Doc9	0	0	0	0	1	1	0
Doc10	0	0	0	0	1	0	0

Doc1	1	0	0	1	0	0	0	Doc2	1	1	2	1	0	0	0
Doc3	1	1	1	0	0	0	0	Doc4	0	1	1	1	0	0	0
Doc5	1	0	1	0	0	0	0								

brca	1	1	1	0	1	0	0	0	0	0	breast	0	1	1	1	0	0	0	0	0	0
cancer	0	2	1	1	1	1	0	0	0	0	Mutat.	1	1	0	1	0	0	0	0	0	0

Direct File

- A file where documents themselves provide the main order of the files.
- A relevance judgement must be made for each and every document in a direct file.

	brca	breast	cancer	mutation	neuro degeneration	neuronal	neuro degenerative
Doc1	1	0	0	1	0	0	0
Doc2	1	1	2	1	0	0	0
Doc3	1	1	1	0	0	0	0
Doc4	0	1	1	1	0	0	0
Doc5	1	0	1	0	0	0	0
Doc6	0	0	1	0	0	0	1
Doc7	0	0	0	0	1	0	0
Doc8	0	0	0	0	1	0	0
Doc9	0	0	0	0	1	1	0
Doc10	0	0	0	0	1	0	0

Inverted File

- A file where values for each term for the documents are recorded.
- A relevance judgement can be made by using each topic term as a key to the corresponding documents.

Transposed Matrix

	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10
brca	1	1	1	0	1	0	0	0	0	0
breast	0	1	1	1	0	0	0	0	0	0
cancer	0	2	1	1	1	1	0	0	0	0
mutation	1	1	0	1	0	0	0	0	0	0
neuro degeneration	0	0	0	0	0	0	1	1	1	1
neuronal	0	0	0	0	0	0	0	0	1	0
neuro degenerative	0	0	0	0	0	1	0	0	0	0

Update the Inverted File

- When new records are added, both the file and index must be changed.
- Updating can be expensive, and might outweigh the gains in the search process.

	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10
brca	1	1	1	0	1	0	0	0	0	0
breast	0	1	1	1	0	0	0	0	0	0
cancer	0	2	1	1	1	1	0	0	0	0
mutation	1	1	0	1	0	0	0	0	0	0
neuro degeneration	0	0	0	0	0	0	1	1	1	1
neuronal	0	0	0	0	0	0	0	0	1	0
neuro degenerative	0	0	0	0	0	1	0	0	0	0
alzheimer	0	0	0	0	0	1	1	0	1	0

Update the Inverted File

- The order of indexed terms also should be updated.

	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10
brca	1	1	1	0	1	0	0	0	0	0
breast	0	1	1	1	0	0	0	0	0	0
cancer	0	2	1	1	1	1	0	0	0	0
mutation	1	1	0	1	0	0	0	0	0	0
neuro degeneration	0	0	0	0	0	0	1	1	1	1
neuronal	0	0	0	0	0	0	0	0	1	0
neuro degenerative	0	0	0	0	0	1	0	0	0	0
alzheimer	0	0	0	0	0	1	1	0	1	0

Update the Inverted File

- The order of indexed terms also should be updated.

	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10
alzheimer	0	0	0	0	0	1	1	0	1	0
brca	1	1	1	0	1	0	0	0	0	0
breast	0	1	1	1	0	0	0	0	0	0
cancer	0	2	1	1	1	1	0	0	0	0
mutation	1	1	0	1	0	0	0	0	0	0
neuro degene ration	0	0	0	0	0	0	1	1	1	1
neuronal	0	0	0	0	0	0	0	0	1	0
neuro degene rative	0	0	0	0	0	1	0	0	0	0

Boolean Expressions

- What if there are more than one term in the query?
  - AND (both terms should appear)
  - OR (one of terms should appear)
  - NOT (the term should not appear)

Order of Boolean Operations

- Usually, all equivalent operators are performed from left to right.
- APPLE AND ORANGE OR BANANA

Terms	Documents reference numbers
APPLE	1 3 5 7
ORANGE	2 3 4 5 6
BANANA	4 6 8
GRAPE	3 7 9 11

3 5

3 4 5 6 8

Example from Salton's Book

Order of Boolean Operations

- Parentheses can change the order. Operations within parentheses normally have priority.
- (APPLE AND ORANGE) OR (BANANA AND ORANGE)

Terms	Documents reference numbers
APPLE	1 3 5 7
ORANGE	2 3 4 5 6
BANANA	4 6 8
GRAPE	3 7 9 11

3 5

4 6

3 4 5 6

Example from Salton’s Book

Adjacency Operations + Frequency

- If you are looking for “BRCA” (beast cancer), what will be more relevant?
  - – Challenges of treating incidental synchronous bilateral **breast cancer** with differing tumor biology.
  - – Radioresistance of the **breast** tumor is highly correlated to its level of **cancer** stem cell and its clinical implication for breast irradiation.

Adjacency Operations + Frequency

- How can we utilize the adjacency in retrieving relevant documents?
  - Add information about the location of terms within each document
  - Doc 102: Radioresistance of the **breast** tumor is highly correlated to its level of **cancer** stem cell and its clinical implication for **breast** irradiation.
  - e.g., breast: 102, 2[4, 21] , cancer: 102, 1[13]
  - *docid, tf [pos_1, pos_2, ..., pos_tf]*

Manual Indexing vs. Automatic Indexing

- Manual Indexing (e.g., PubMed): indexing process usually done by experts using controlled vocabulary, taxonomy, thesaurus or ontology
  - "biomedical subject specialists who analyze the subject content of articles and index the concepts that are discussed, using the Medical Subject Headings (MeSH) controlled vocabulary; and computer and information specialists who develop and maintain the various systems, including the retrieval system." (NLM)
- Automatic Indexing (e.g., Google): a process in which computers scan documents against controlled vocabulary, taxonomy, thesaurus or ontology and build indexes. The resource are often built automatically.

Ontologies

- Conceptual structure consisting of vocabularies that are descriptive of a domain/topical area
  - Which provides us a view of the key topics in a domain
  - Which provides a way to understand relationships among topics
  - Which can be applied in data indexing, annotation, integration, retrieval, and analysis

* Next few pages are courtesy of Dr. Javed Mostafa

Medical Subject Heading (MeSH)

- “MeSH is the National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity.”
- “MeSH descriptors are arranged in both an alphabetic and a hierarchical structure. ” (NLM)

Medical Subject Heading List: Breast Cancer

- Take the example of breast cancer in the context of MeSH. It has the following major terms:
    - Breast Neoplasms
    - Breast Cancer
    - Breast Carcinoma
    - Breast Tumors
    - Cancer of Breast
    - Malignant Neoplasm of Breast
    - Malignant Tumor of Breast
    - Mammary Neoplasm, Human

MeSH Definition: Breast Neoplasms

- Scope Note: Tumors or cancer of the human BREAST.
- Annotation: human only; BREAST NEOPLASMS, MALE is also available; for animal, index MAMMARY NEOPLASMS, ANIMAL or MAMMARY NEOPLASMS, EXPERIMENTAL; coordinate IM with histological type of neoplasm (IM)

MeSH Ontology Structure

- Neoplasms [C04]Neoplasms by Site [C04.588]
  - Abdominal Neoplasms [C04.588.033]
  - Anal Gland Neoplasms [C04.588.083]
  - Bone Neoplasms [C04.588.149]
  - **Breast Neoplasms [C04.588.180]**
    - Breast Carcinoma In Situ [C04.588.180.130]
    - Breast Neoplasms, Male [C04.588.180.260]
    - Carcinoma, Ductal, Breast [C04.588.180.390]
    - Carcinoma, Lobular [C04.588.180.437]
    - Hereditary Breast and Ovarian Cancer Syndrome [C04.588.180.483]
    - Inflammatory Breast Neoplasms [C04.588.180.576]
    - Unilateral Breast Neoplasms [C04.588.180.682]
    - Triple Negative Breast Neoplasms [C04.588.180.788]
  - Digestive System Neoplasms [C04.588.274]
  - Endocrine Gland Neoplasms [C04.588.322]
  - Eye Neoplasms [C04.588.364]

Other Resources

- ICD9/ICD10
- SNOMED CT
- LOINC
- UMLS
- …

Text Processing

N-gram

- N-gram: a sequence of n tokens from a given text
- Hepatitis B virus reactivation in breast cancer patients undergoing chemotherapy
- Unigram: [“Hepatitis”, “B”, “virus”, “reactivation”, “breast”, “cancer”, “patients”, “undergoing”, “chemotherapy”]
- Bigram: [“Hepatitis B”, “B virus”, “virus reactivation”, “reactivation breast”, “breast cancer”, “cancer patients”, “patients undergoing”, “undergoing chemotherapy”]
- Triagram: [“Hepatitis B virus”, “B virus reactivation”, “virus reactivation breast”, “reactivation breast cancer”, “breast cancer patients”, “cancer patients undergoing”, …]

Text Processing

- Down-casing: converting text to lower-case
- Tokenization: splitting text into terms or tokens

Text Processing: original text

The purpose of this study was to investigate decision patterns to reduce the risks of BRCA-related breast and gynecologic cancers in carriers of BRCA pathogenic variants. We found a change in risk-reducing (RR) management patterns after December 2012, when the National Health Insurance System (NHIS) of Korea began to pay for BRCA testing and risk-reducing salpingo-oophorectomy (RRSO) in pathogenic-variant carriers. The study group consisted of 992 patients, including 705 with breast cancer (BC), 23 with ovarian cancer (OC), and 254 relatives of high-risk patients who underwent BRCA testing at the National Cancer Center of Korea from January 2008 to December 2016.

Text Processing: down-casing

the purpose of this study was to investigate decision patterns to reduce the risks of brca-related breast and gynecologic cancers in carriers of brca pathogenic variants. we found a change in risk-reducing (rr) management patterns after december 2012, when the national health insurance system (nhis) of korea began to pay for brca testing and risk-reducing salpingo-oophorectomy (rrso) in pathogenic-variant carriers. the study group consisted of 992 patients, including 705 with breast cancer (bc), 23 with ovarian cancer (oc), and 254 relatives of high-risk patients who underwent brca testing at the national cancer center of korea from january 2008 to december 2016.

Text Processing: tokenization

```
['the', 'purpose', 'of', 'this', 'study', 'was', 'to', 'investigate',
'decision', 'patterns', 'to', 'reduce', 'the', 'risks', 'of', 'brca-
related', 'breast', 'and', 'gynecologic', 'cancers', 'in', 'carriers', 'of',
'brca', 'pathogenic', 'variants', '.', 'we', 'found', 'a', 'change', 'in',
'risk-reducing', '(', 'rr', ')', 'management', 'patterns', 'after',
'december', '2012', ',', 'when', 'the', 'national', 'health',
'insurance', 'system', '(', 'nhis', ')', 'of', 'korea', 'began', 'to', 'pay',
'for', 'brca', 'testing', 'and', 'risk-reducing', 'salpingo-
oophorectomy', '(', 'rrso', ')', 'in', 'pathogenic-variant', 'carriers',
'.', 'the', 'study', 'group', 'consisted', 'of', '992', 'patients', ',',
'including', '705', 'with', 'breast', 'cancer', '(', 'bc', ')', ',’, …]
```

Text Processing: removing stopwords

```
['purpose', 'study', 'investigate', 'decision', 'patterns',
'reduce', 'risks', 'brca-related', 'breast', 'gynecologic',
'cancers', 'carriers', 'brca', 'pathogenic', 'variants', '.',
'found', 'change', 'risk-reducing', '(', 'rr', ')', 'management',
'patterns', 'december', '2012', ',', 'national', 'health',
'insurance', 'system', '(', 'nhis', ')', 'korea', 'began', 'pay',
'brca', 'testing', 'risk-reducing', 'salpingo-oophorectomy', '(',
'rrso', ')', 'pathogenic-variant', 'carriers', '.', 'study', 'group',
'consisted', '992', 'patients', ',', 'including', '705', 'breast',
'cancer', '(', 'bc', ')', ‘,’…]
```

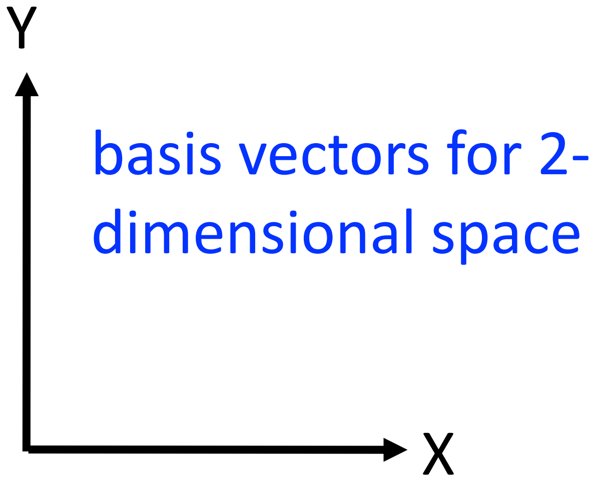
Text Processing:
in Python

```
import nltk
from nltk.corpus import stopwords
Import codecs
…
file = codecs.open(“training.txt", "r", encoding='utf-8')
lines = file.readlines()
for text in lines:
        lower_text = text.lower()
        temp_tokens = nltk.word_tokenize(lower_text)
        filtered_tokens = [w for w in tokens if not w in
                         stopwords.words('english')]
```

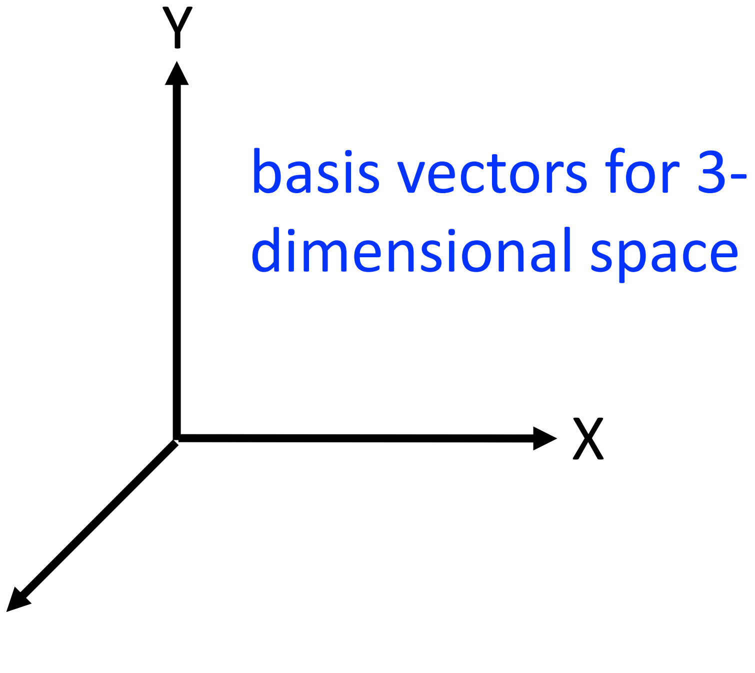
Vector Space

What is a Vector Space?

- Formally, a vector space is defined by a set of linearly independent basis vectors
- The basis vectors correspond to the dimensions or directions of the vector space



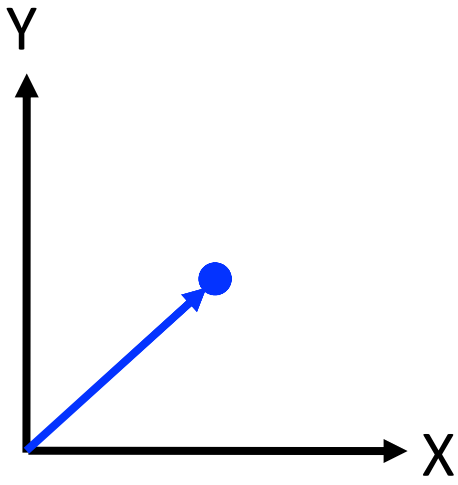
basis vectors for 2-dimensional space

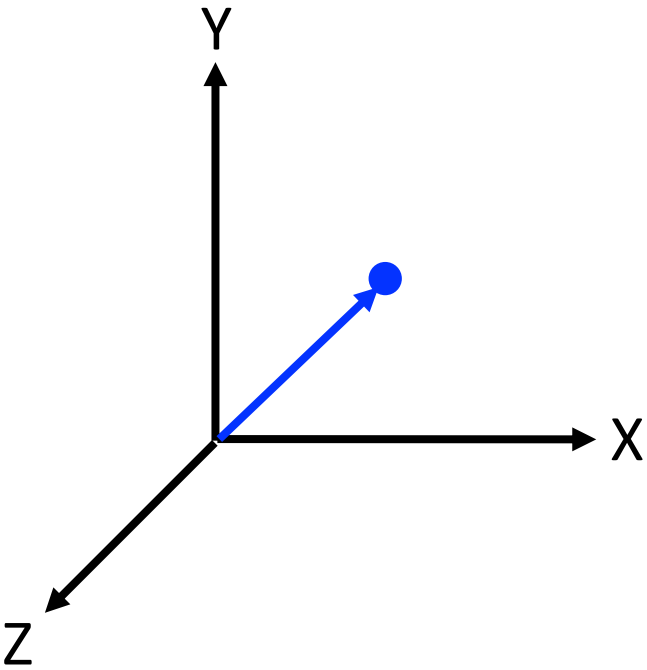


basis vectors for 3-dimensional space

What is a Vector?

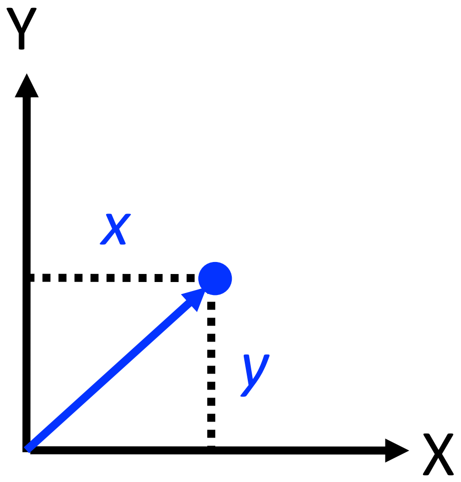
- A vector is a point in a vector space and has length (from the origin to the point) and direction

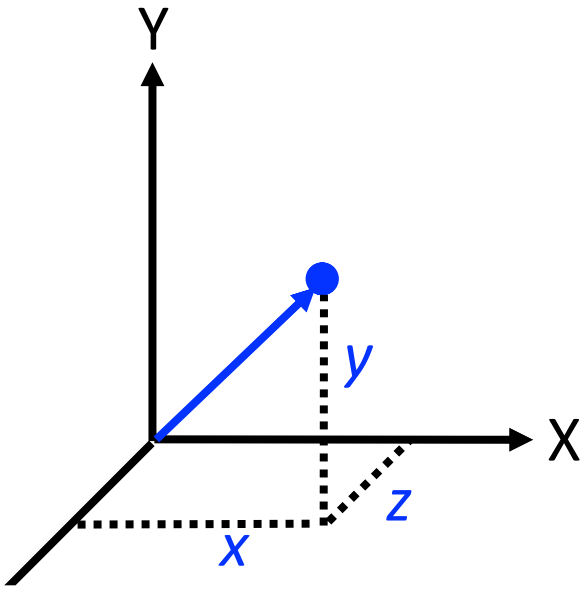




What is a Vector?

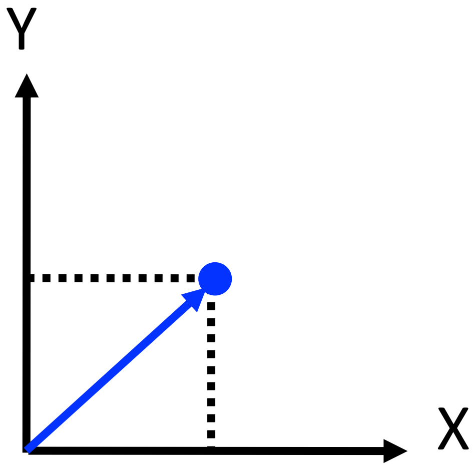
- A 2-dimensional vector can be written as [*x*,*y*]
- A 3-dimensional vector can be written as [*x*,*y*,*z*]



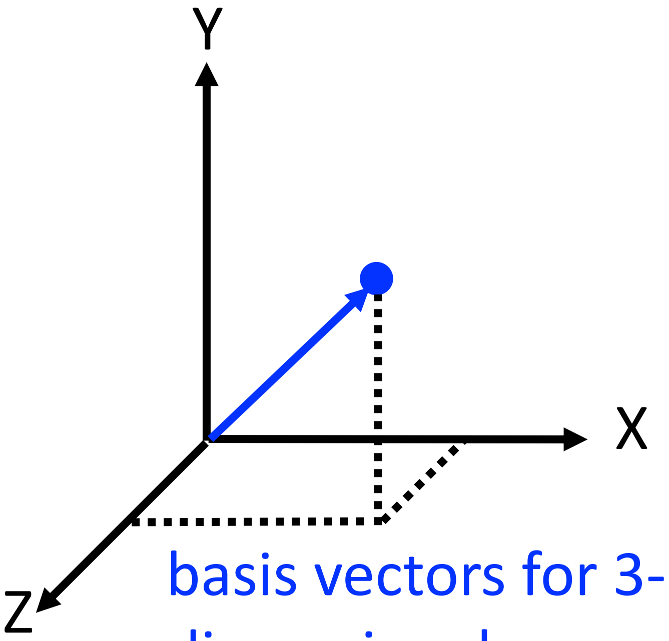


What is a Vector Space?

- The basis vectors are linearly independent because knowing a vector’s value along one dimension doesn’t say anything about its value along another dimension



basis vectors for 2-dimensional space



basis vectors for 3-dimensional space

Binary Text Representation

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10	sentiment
1	0	1	0	1	0	0	1	1	0	positive
0	1	0	1	1	0	1	1	0	0	negative
0	1	0	1	1	0	1	0	0	0	negative
0	0	1	0	1	1	0	1	1	1	positive
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	0	1	1	0	0	1	0	1	positive

- Terms as features
- Bag of words representation: no word order
- 1 = the term appears in the text and 0 = the term does not appear in the text

Vector Space Representation

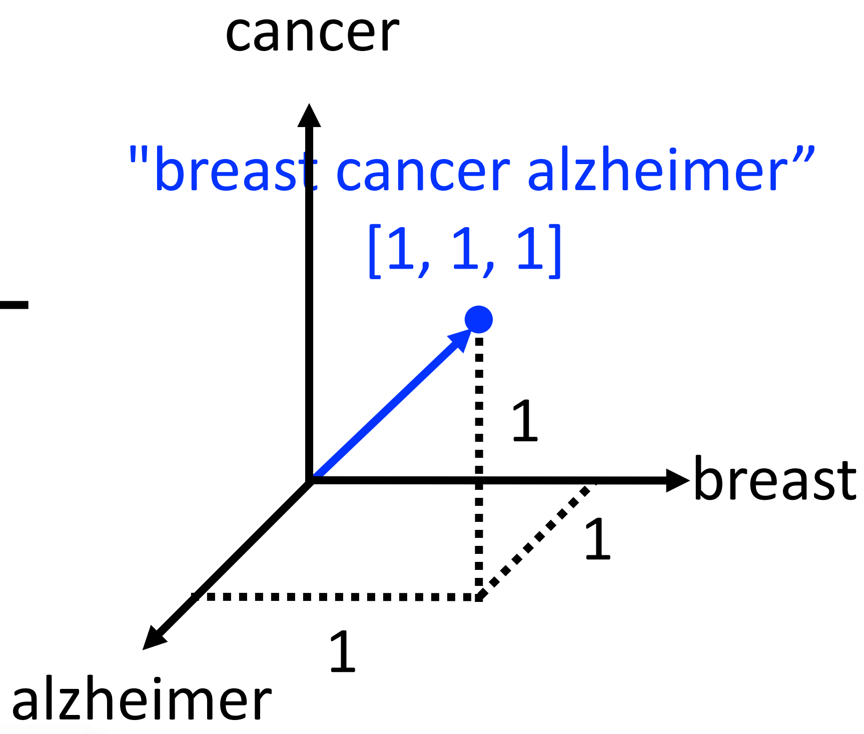
- Let V denote the set of features in our feature representation
- Any arbitrary instance can be represented as a vector in |V|-dimensional space
- For simplicity, let’s assume three features: breast, cancer, alzheimer (i.e., |V| = 3)
- Why? Because it’s easy to visualize 3-D space

Vector Space Representation

with binary weights

- 1 = the term appears at least once
- 0 = the term does not appear

	breast	*cancer*	*alzheimer*
i_1	1	1	1

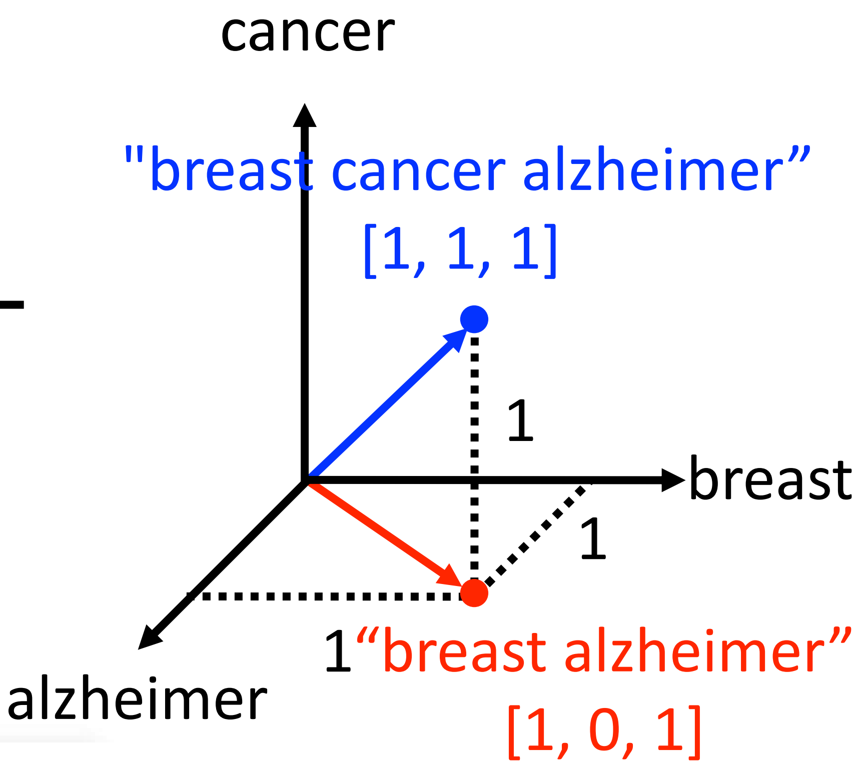


Vector Space Representation

with binary weights

- 1 = the term appears at least once
- 0 = the term does not appear

	breast	*cancer*	*alzheimer*
i_1	1	1	1
i_2	1	0	1

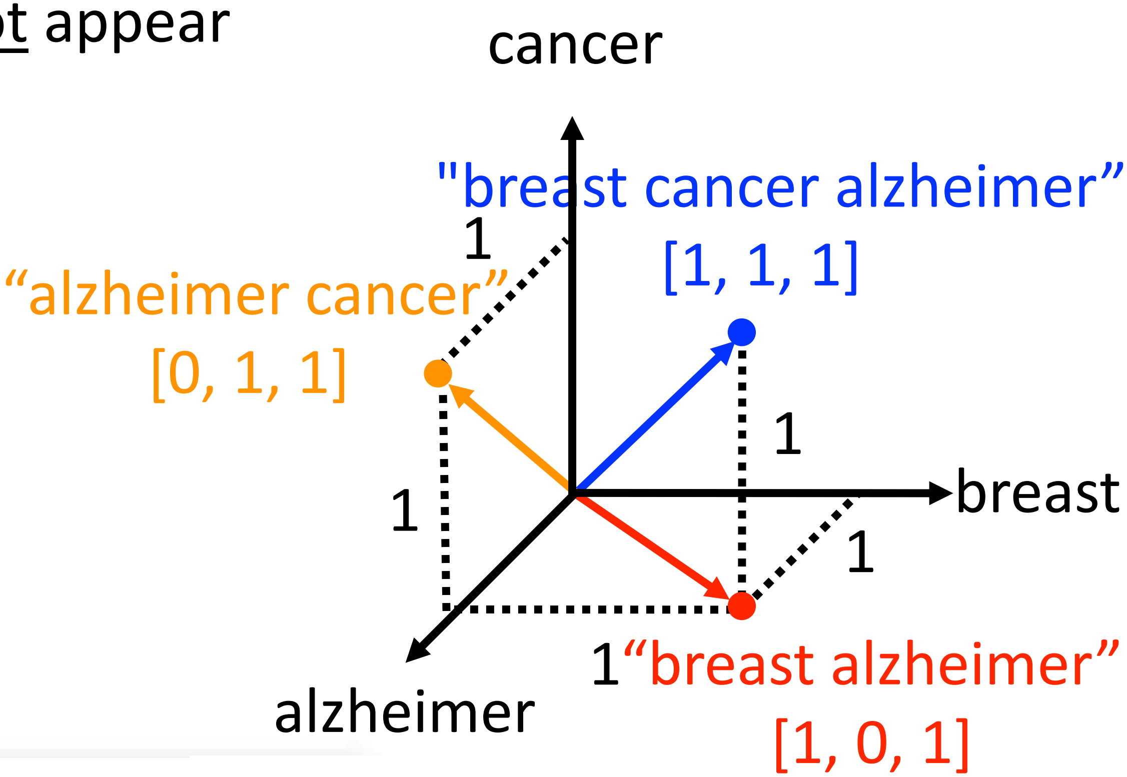


Vector Space Representation

with binary weights

- 1 = the term appears at least once
- 0 = the term does not appear

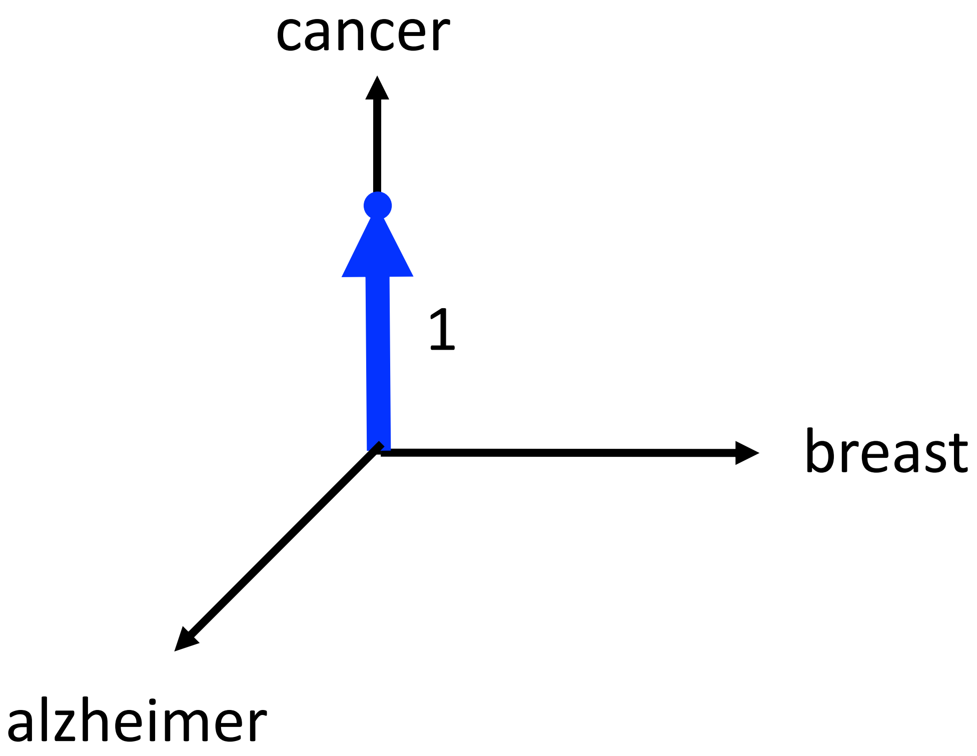
	breast	*cancer*	*alzheimer*
i_1	1	1	1
i_2	1	0	1
i_3	0	1	1



Vector Space Representation

with binary weights

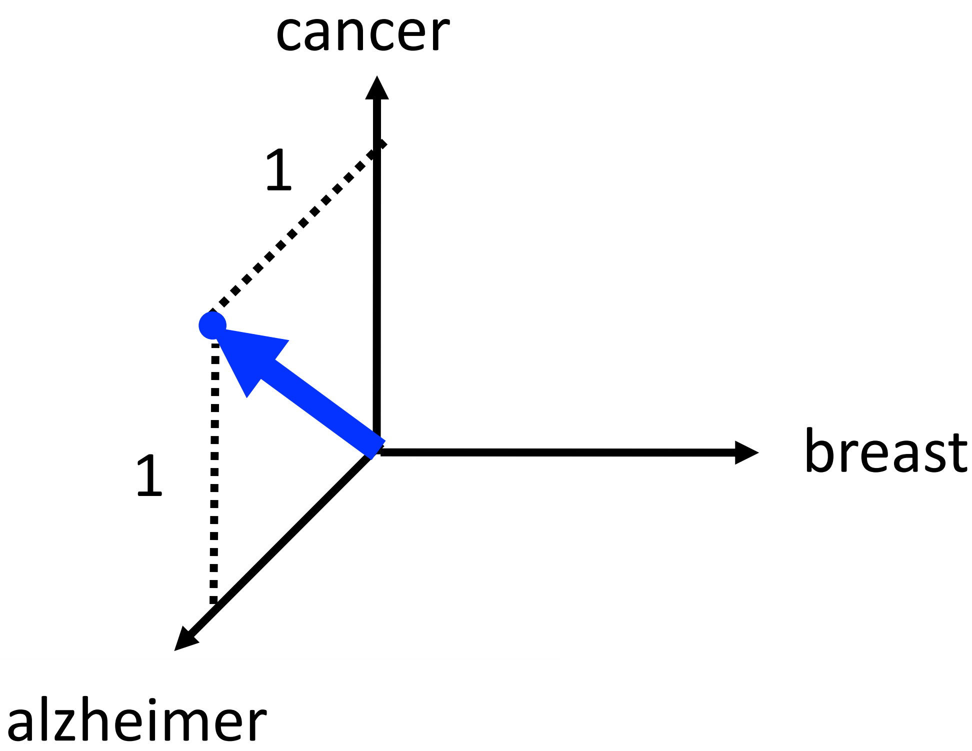
- What span(s) of text does this vector represent?



Vector Space Representation

with binary weights

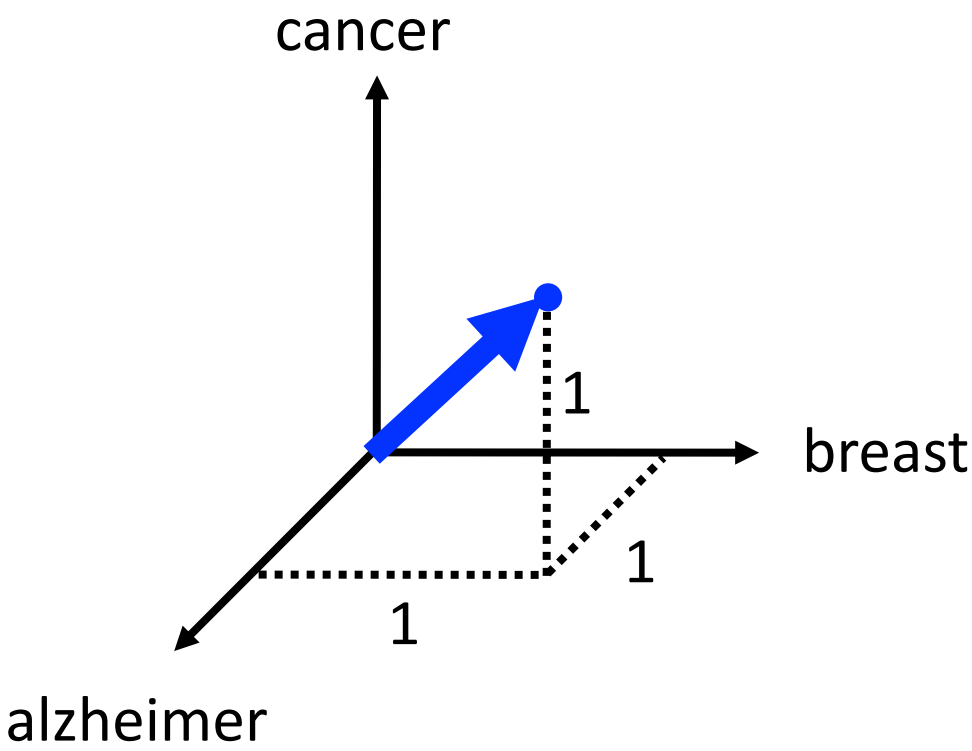
- What span(s) of text does this vector represent?



Vector Space Representation

with binary weights

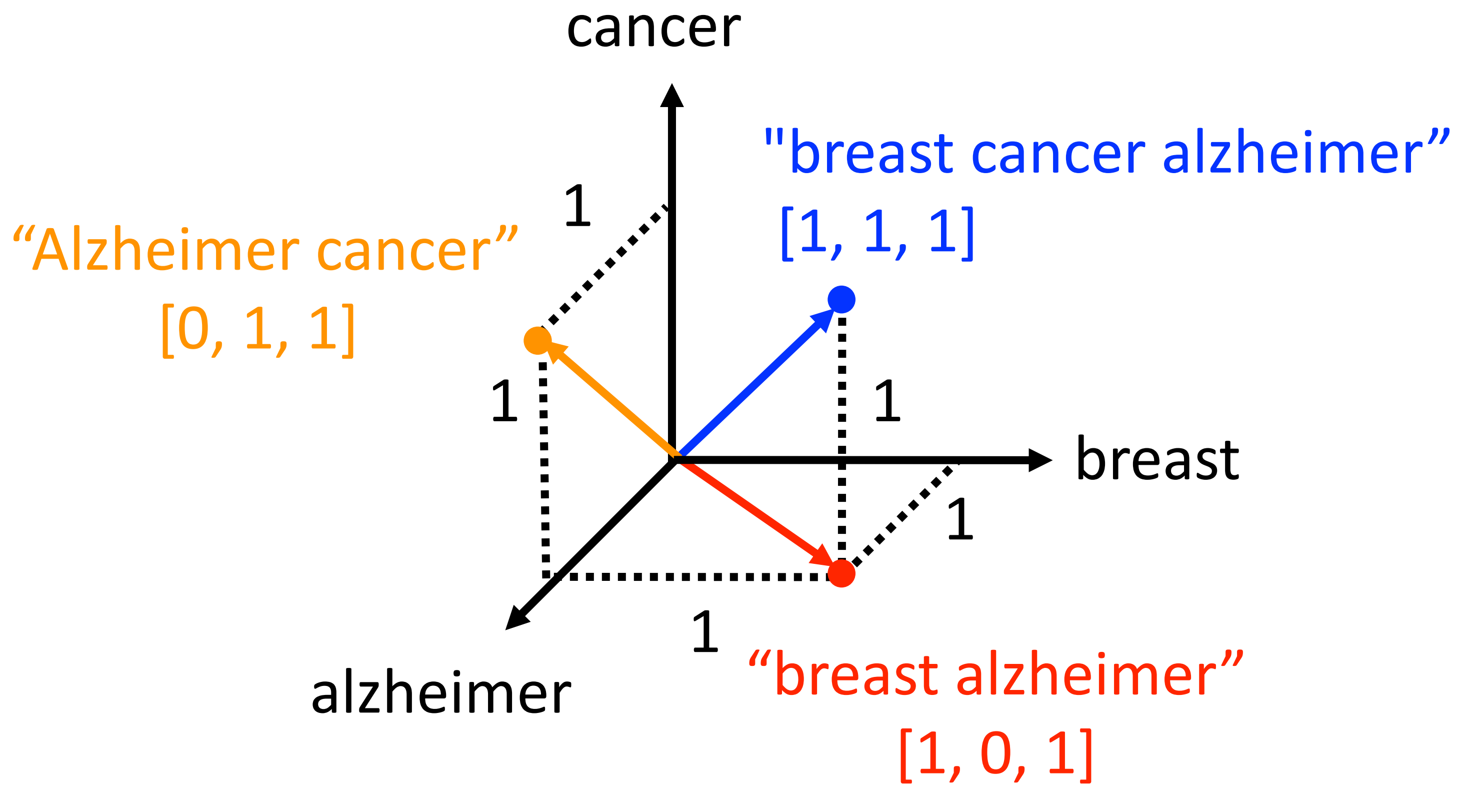
- What span(s) of text does this vector represent?



Vector Space Representation

with binary weights

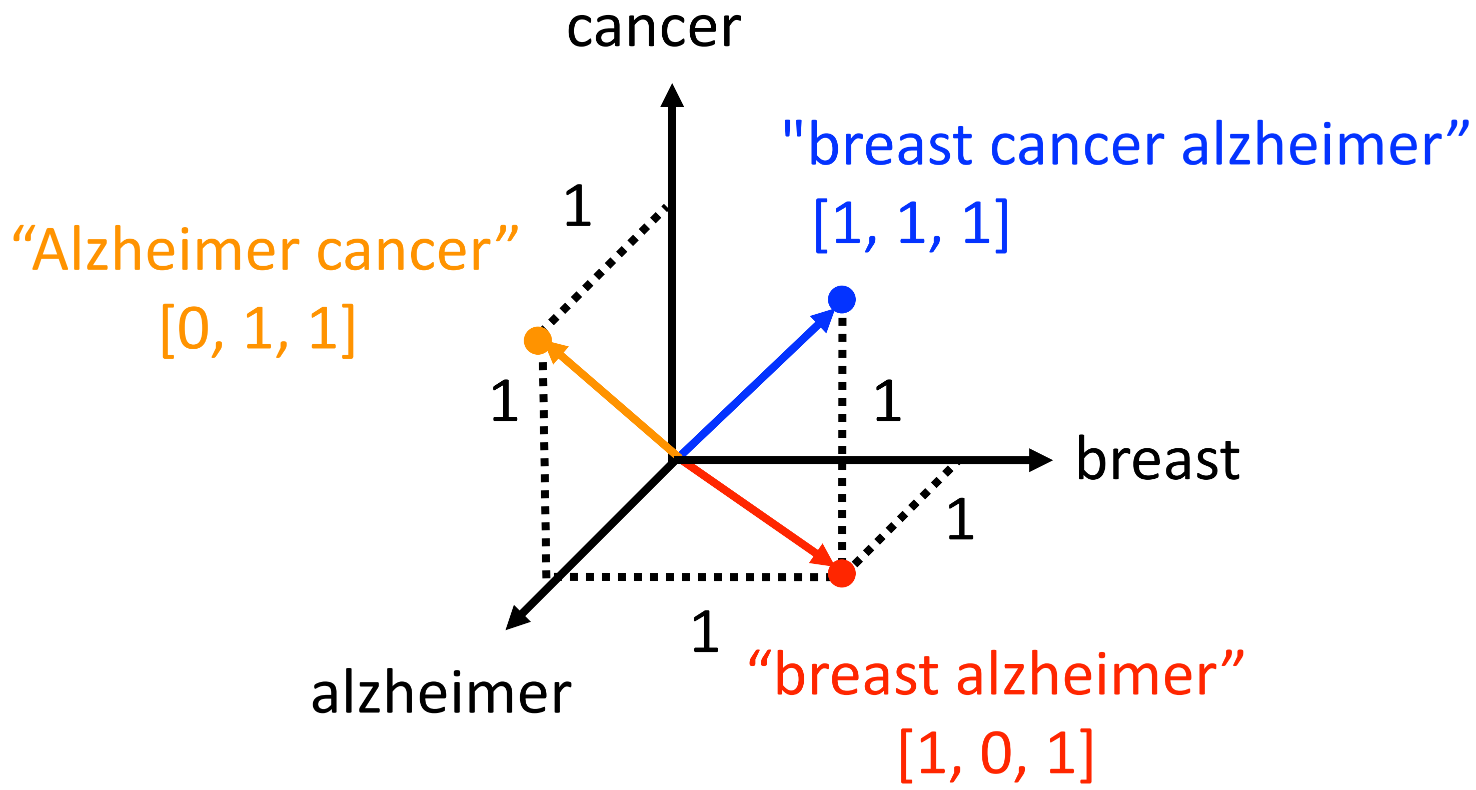
- Any arbitrary span of text can be represented as a vector in |V|-dimensional space



Vector Space Representation

with binary weights

- How can we use a vector-space representation to compute similarity or distance?



Vector Space Representation

with binary weights

- How can we use a vector-space representation to compute similarity or distance?
- Euclidean distance:

$$D(x,y) = \sqrt{\left(\sum_{i=1}^{|\mathcal{V}|}(x_i - y_i)^2\right)}$$

$x = (x_1, x_2, x_3)$ $y = (y_1, y_2, y_3)$

Euclidean Distance

	x	y	$(x_i - y_i)^2$
Breast	1	1	0
Cancer	1	1	0
alzheimer	1	1	0
$D(x,y) = \sqrt{\left(\sum_{i=1}^{\vert\mathcal{V}\vert}(x_i - y_i)^2\right)}$			0

"breast cancer alzheimer" vs. "breast cancer alzheimer"

Euclidean Distance

	x	y	$(x_i - y_i)^2$
Breast	1	1	0
Cancer	1	1	0
alzheimer	1	0	1
$D(x,y) = \sqrt{\left(\sum_{i=1}^{\lvert\mathcal{V}\rvert}(x_i - y_i)^2\right)}$			1

“breast cancer alzheimer” vs. “breast cancer”

Euclidean Distance

	x	y	$(x_i - y_i)^2$
Breast	1	0	1
Cancer	1	1	0
alzheimer	1	0	1
$D(x,y) = \sqrt{\left(\sum_{i=1}^{\lvert\mathcal{V}\rvert}(x_i - y_i)^2\right)}$			1.41

“breast cancer alzheimer” vs. “cancer”

Binary Text Representation

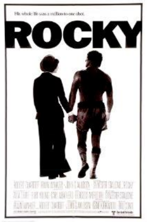
w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_10	sentiment
1	0	1	0	1	0	0	1	1	0	positive
0	1	0	1	1	0	1	1	0	0	negative
0	1	0	1	1	0	1	0	0	0	negative
0	0	1	0	1	1	0	1	1	1	positive
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	0	1	1	0	0	1	0	1	positive

- Is this a good (bag of words) representation?
- Can we do better?

Term Weighting

Term-Weighting

what are the most important terms?

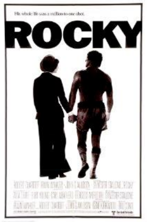


- Movie: Rocky (1976)
- Plot:

Rocky Balboa is a struggling boxer trying to make the big time. Working in a meat factory in Philadelphia for a pittance, he also earns extra cash as a debt collector. When heavyweight champion Apollo Creed visits Philadelphia, his managers want to set up an exhibition match between Creed and a struggling boxer, touting the fight as a chance for a "nobody" to become a "somebody". The match is supposed to be easily won by Creed, but someone forgot to tell Rocky, who sees this as his only shot at the big time. Rocky Balboa is a small-time boxer who lives in an apartment in Philadelphia, Pennsylvania, and his career has so far not gotten off the canvas. Rocky earns a living by collecting debts for a loan shark named Gazzo, but Gazzo doesn't think Rocky has the viciousness it takes to beat up deadbeats. Rocky still boxes every once in a while to keep his boxing skills sharp, and his ex-trainer, Mickey, believes he could've made it to the top if he was willing to work for it. Rocky, goes to a pet store that sells pet supplies, and this is where he meets a young woman named Adrian, who is extremely shy, with no ability to talk to men. Rocky befriends her. Adrain later surprised Rocky with a dog from the pet shop that Rocky had befriended. Adrian's brother Paulie, who works for a meat packing company, is thrilled that someone has become interested in Adrian, and Adrian spends Thanksgiving with Rocky. Later, they go to Rocky's apartment, where Adrian explains that she has never been in a man's apartment before. Rocky sets her mind at ease, and they become lovers. Current world heavyweight boxing champion Apollo Creed comes up with the idea of giving an unknown a shot at the title. Apollo checks out the Philadelphia boxing scene, and chooses Rocky. Fight promoter Jergens gets things in gear, and Rocky starts training with Mickey. After a lot of training, Rocky is ready for the match, and he wants to prove that he can go the distance with Apollo. The 'Italian Stallion', Rocky Balboa, is an aspiring boxer in downtown Philadelphia. His one chance to make a better life for himself is through his boxing and Adrian, a girl who works in the local pet store. Through a publicity stunt, Rocky is set up to fight Apollo Creed, the current heavyweight champion who is already set to win. But Rocky really needs to triumph, against all the odds...

Term-Frequency

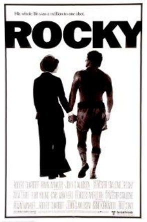
how important is a term?



rank	term	freq.
1	a	22
2	rocky	19
3	to	18
4	the	17
5	is	11
6	and	10
7	in	10
8	for	7
9	his	7
10	he	6
11	adrian	6
12	with	6
13	who	6
14	that	5
15	apollo	5
16	creed	5
17	philadelphia	5
18	has	4
19	pet	4
20	boxing	4
21	up	4
22	an	4
23	boxer	4
24	s	3
25	balboa	3
26	it	3
27	heavyweigh	3
28	champion	3
29	fight	3
30	become	3

Term-Frequency

how important is a term?



rank	term	freq.	rank	term	freq.
1	a	22	16	creed	5
2	rocky	19	17	philadelphia	5
3	to	18	18	has	4
4	the	17	19	pet	4
5	is	11	20	boxing	4
6	and	10	21	up	4
7	in	10	22	an	4
8	for	7	23	boxer	4
9	his	7	24	s	3
10	he	6	25	balboa	3
11	adrian	6	26	it	3
12	with	6	27	heavyweigh	3
13	who	6	28	champion	3
14	that	5	29	fight	3
15	apollo	5	30	become	3

Inverse Document Frequency (IDF)

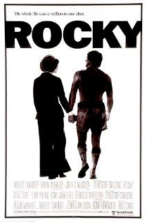
how important is a term?

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

- *N* = number of training set instances
- *df*$_t$ = number of training set instances where term *t* appears

Inverse Document Frequency (IDF)

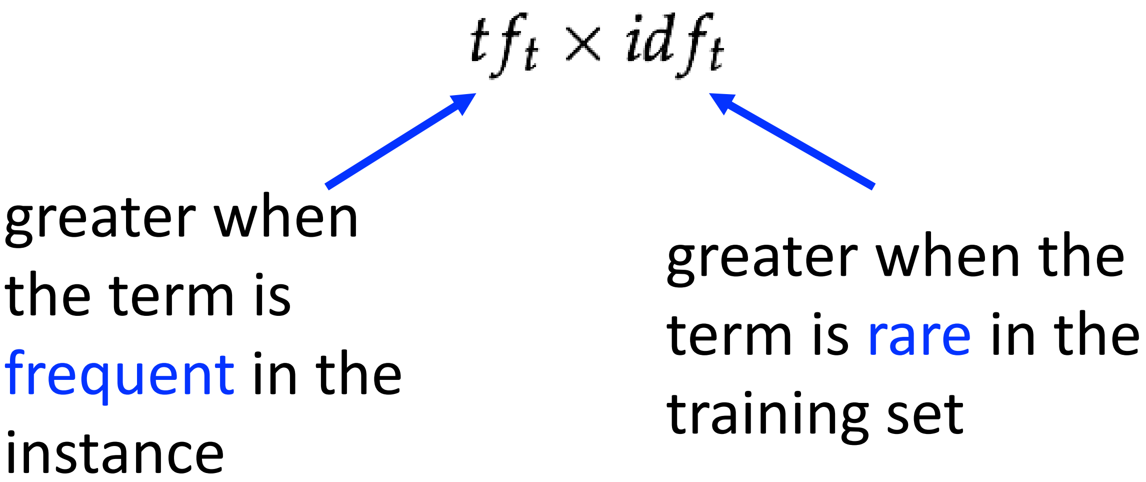
how important is a term?



rank	term	idf	rank	term	idf
1	doesn	11.66	16	creed	6.84
2	adrain	10.96	17	paulie	6.82
3	viciousness	9.95	18	packing	6.81
4	deadbeats	9.86	19	boxes	6.75
5	touting	9.64	20	forgot	6.72
6	jergens	9.35	21	ease	6.53
7	gazzo	9.21	22	thanksgivin	6.52
8	pittance	9.05	23	earns	6.51
9	balboa	8.61	24	pennsylvani	6.50
10	heavyweigh	7.18	25	promoter	6.43
11	stallion	7.17	26	befriended	6.38
12	canvas	7.10	27	exhibition	6.31
13	ve	6.96	28	collecting	6.23
14	managers	6.88	29	philadelphia	6.19
15	apollo	6.84	30	gear	6.18

TF.IDF

how important is a term?



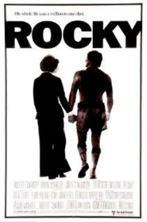
$$tf_t \times idf_t$$

greater when the term is frequent in the instance

greater when the term is rare in the training set

TF.IDF

how important is a term?

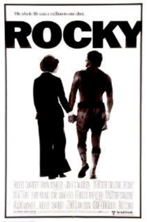


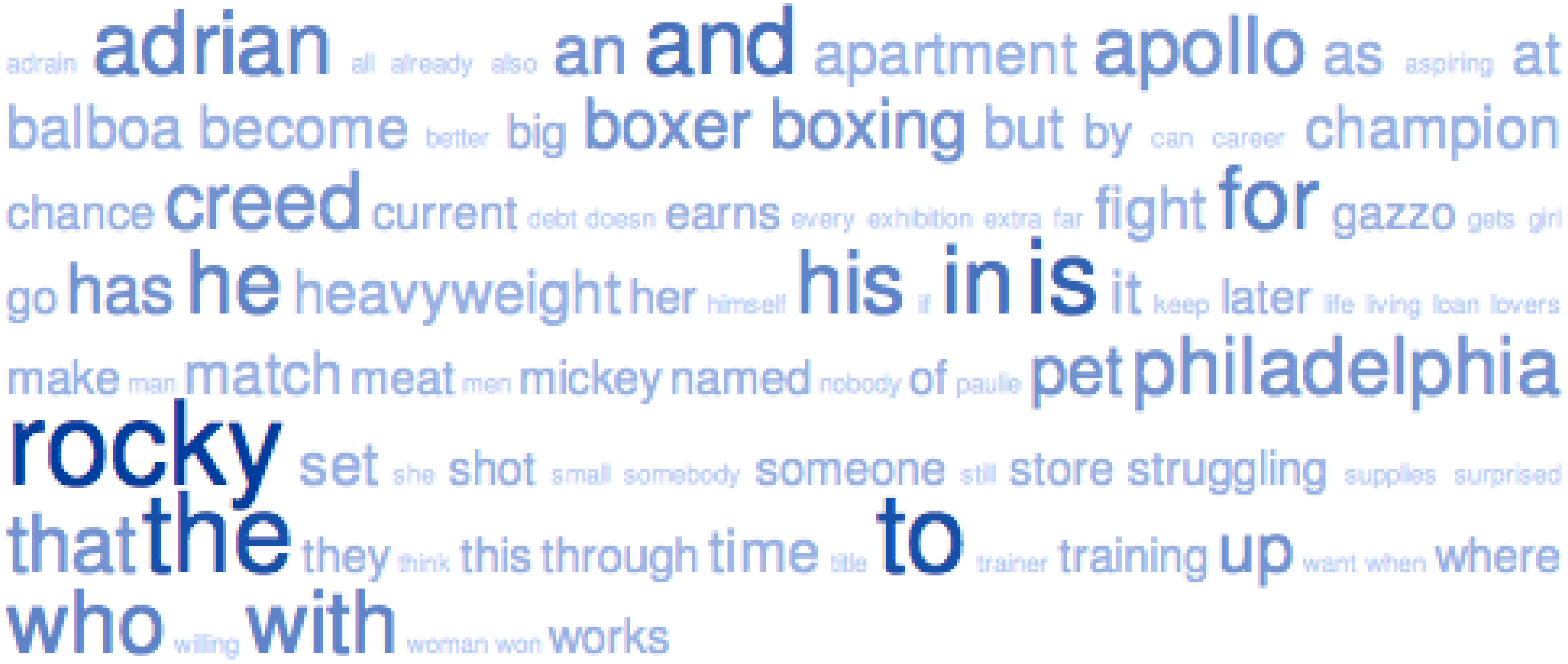
rank	term	tf·idf	rank	term	idf
1	rocky	96.72	16	meat	11.76
2	apollo	34.20	17	doesn	11.66
3	creed	34.18	18	adrain	10.96
4	philadelphia	30.95	19	fight	10.02
5	adrian	26.44	20	viciousness	9.95
6	balboa	25.83	21	deadbeats	9.86
7	boxing	22.37	22	touting	9.64
8	boxer	22.19	23	current	9.57
9	heavyweigh	21.54	24	jergens	9.35
10	pet	21.17	25	s	9.29
11	gazzo	18.43	26	struggling	9.21
12	champion	15.08	27	training	9.17
13	match	13.96	28	pittance	9.05
14	earns	13.01	29	become	8.96
15	apartment	11.82	30	mickey	8.96

TF.IDF/Caricature Analogy

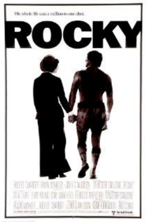
- TF.IDF: accentuates terms that are frequent in the instance, but not frequent in general
- Caricature: exaggerates traits that are characteristic of the person compared to the average

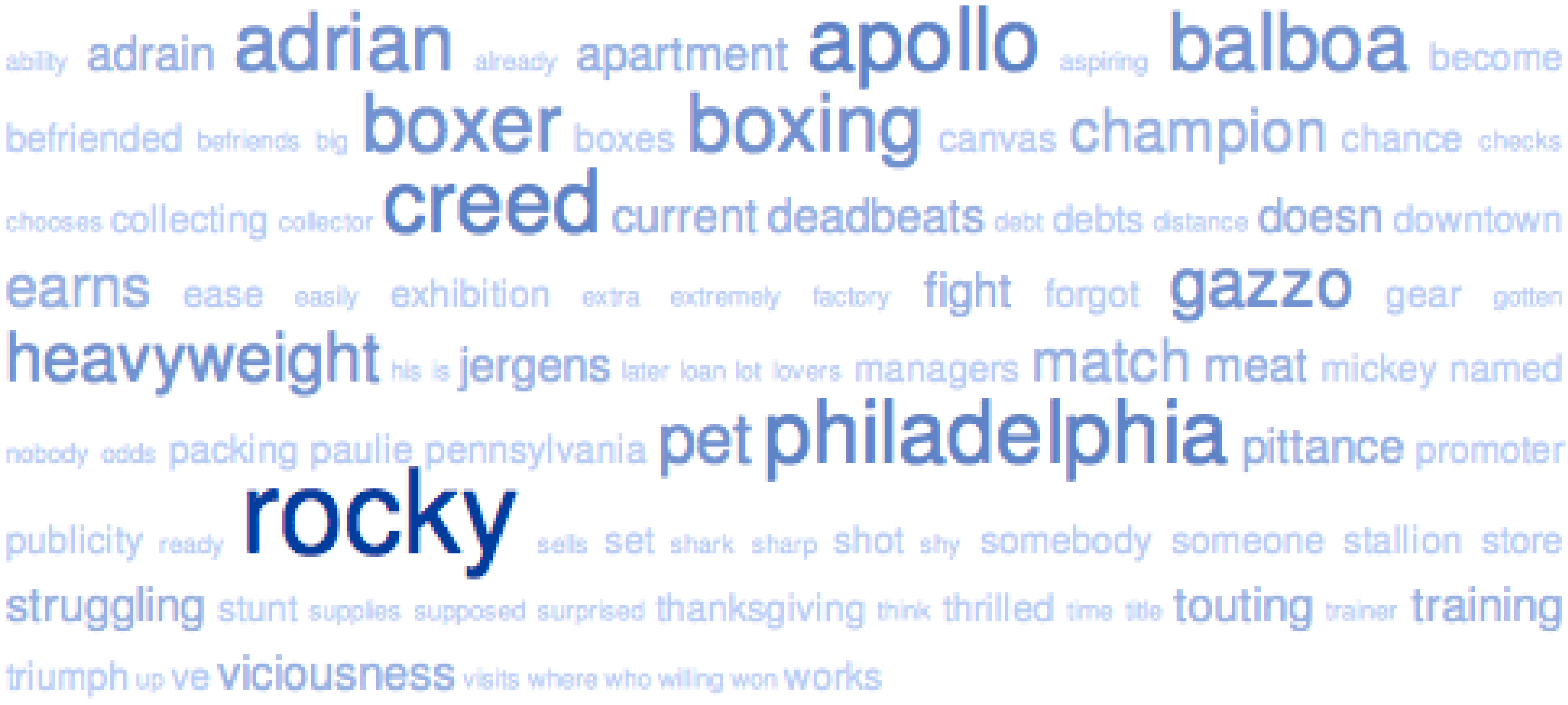
TF, IDF, or TF.IDF?



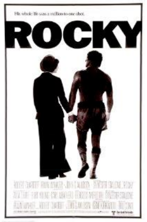


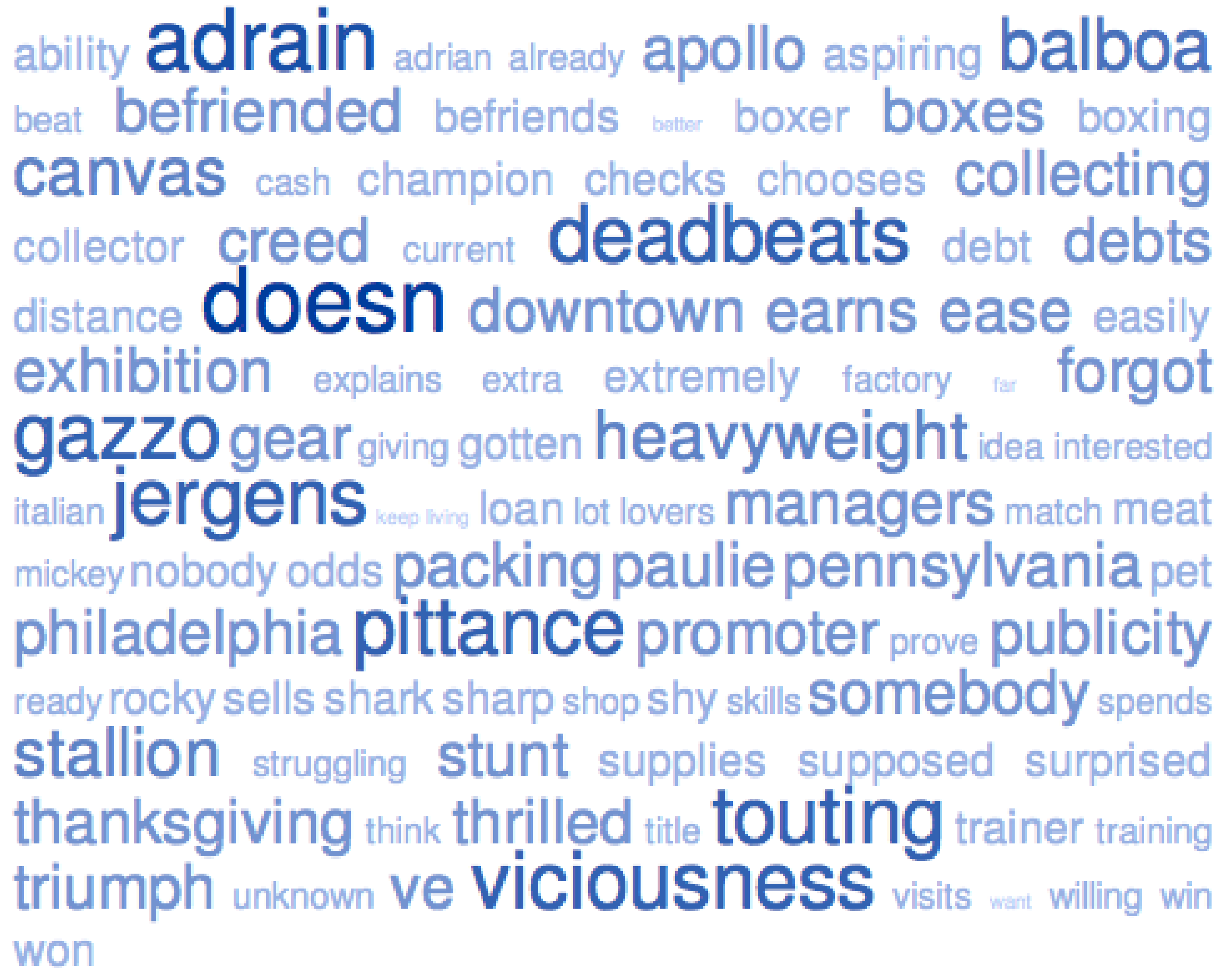
TF, IDF, or TF.IDF?



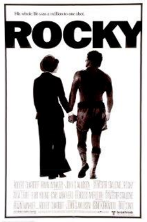


TF, IDF, or TF.IDF?





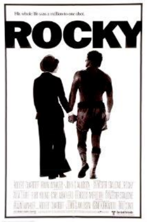
Calculating TF.IDF Weights



$$tf_t \times log\left(\frac{N}{df_t}\right)$$

term	tf	N	df	idf	tf.idf
rocky	19	230721	1420	5.09	96.72
philadelphia	5	230721	473	6.19	30.95
boxer	4	230721	900	5.55	22.19
fight	3	230721	8170	3.34	10.02
mickey	2	230721	2621	4.48	8.96
for	7	230721	117137	0.68	4.75

Putting Everything Together

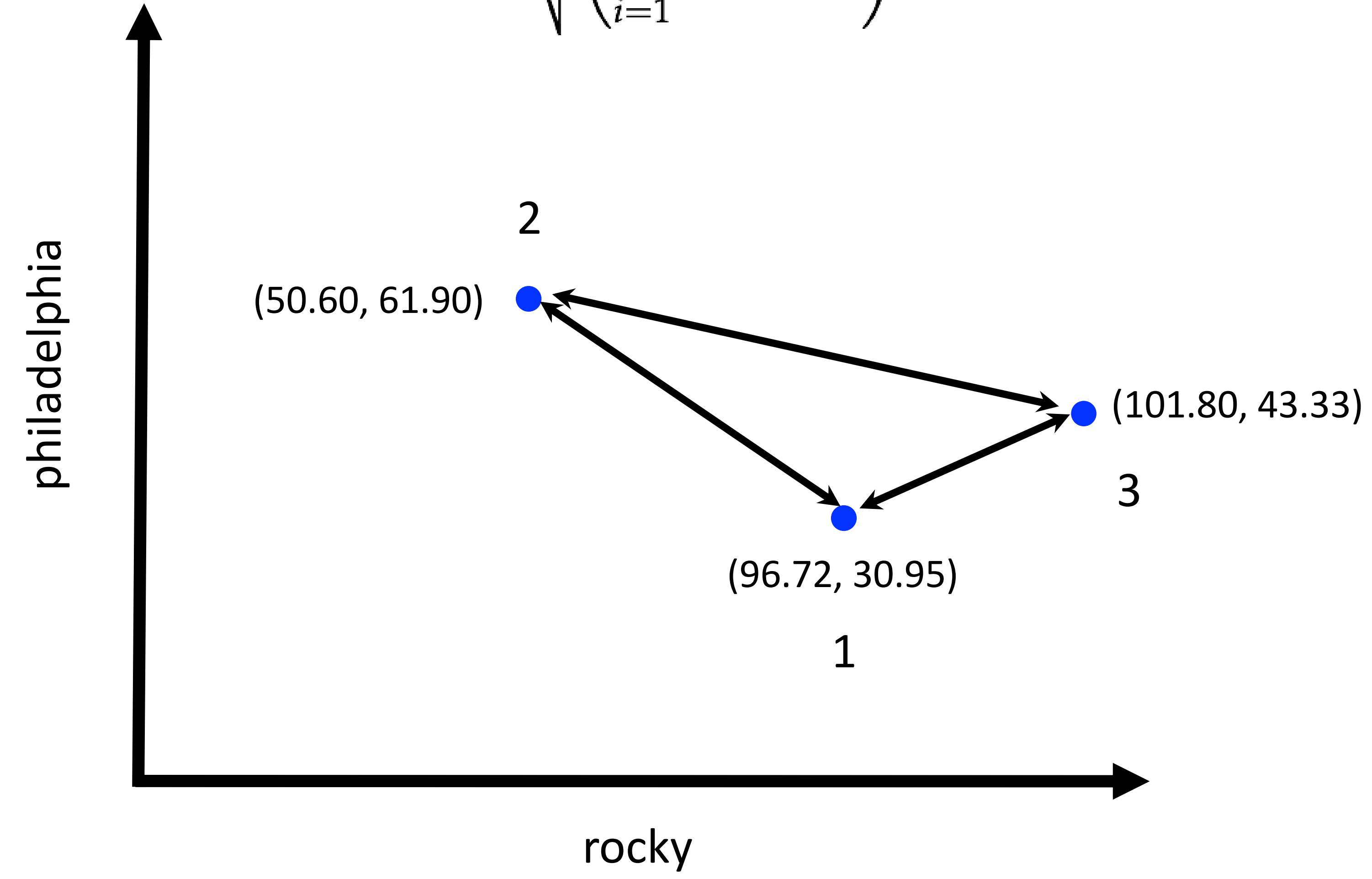


$$tf_t \times log\left(\frac{N}{df_t}\right)$$

	term	tf	N	df	idf	tf.idf
1	rocky	19	230721	1420	5.09	96.72
	philadelphia	5	230721	473	6.19	30.95
2	rocky	10	230721	1420	5.09	50.60
	philadelphia	10	230721	473	6.19	61.90
3	rocky	20	230721	1420	5.09	101.80
	philadelphia	7	230721	473	6.19	43.33

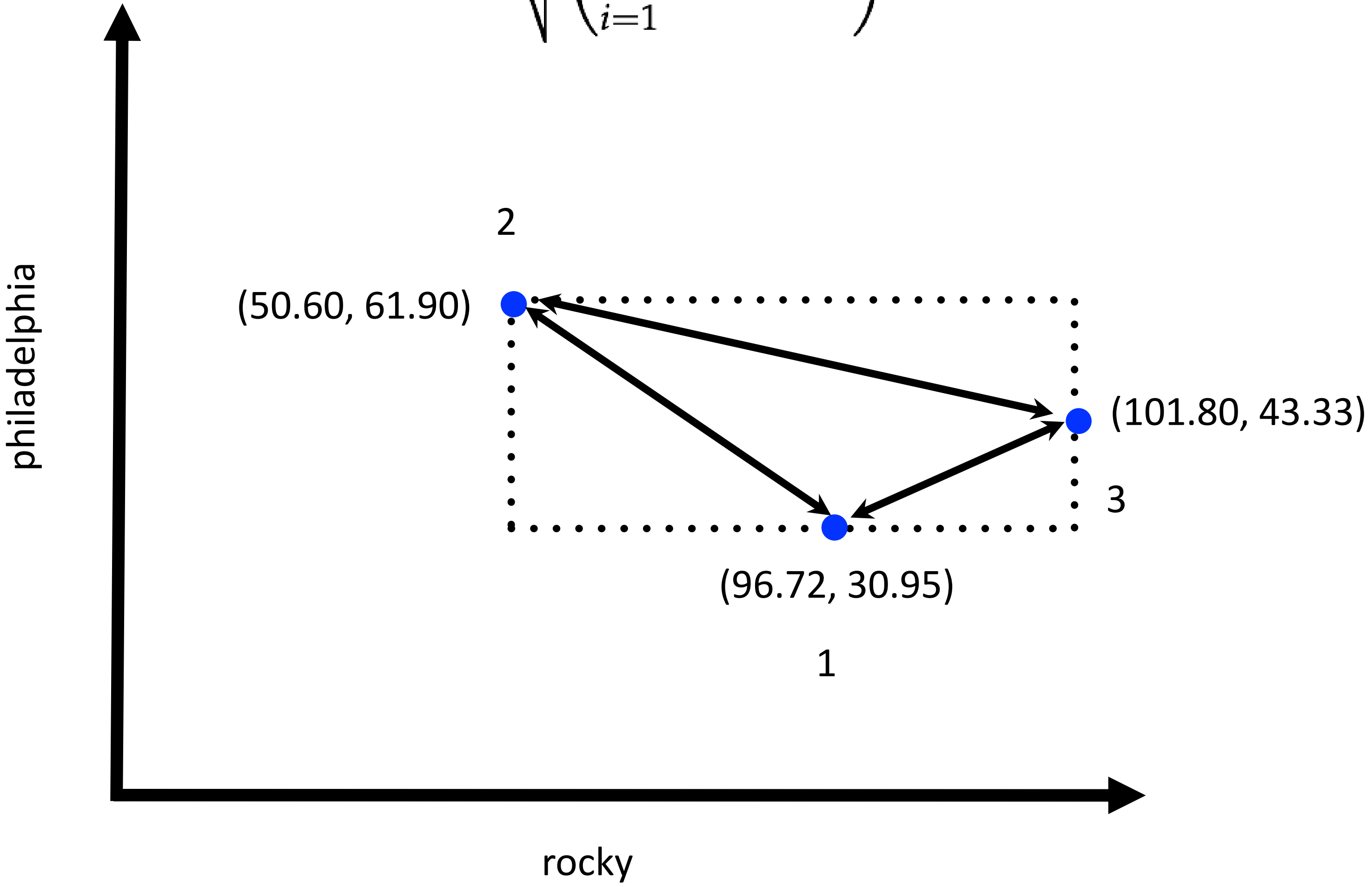
Putting Everything Together

$$D(x,y) = \sqrt{\left(\sum_{i=1}^{|\mathcal{V}|}(x_i - y_i)^2\right)}$$



Putting Everything Together

$$D(x,y) = \sqrt{\left(\sum_{i=1}^{|\mathcal{V}|}(x_i - y_i)^2\right)}$$



Any Questions?

Feature Selection

Next Class