11-755— Spring 2021
Large Scale Multimedia Processing

Lecture 6/6

Multimedia capture and storage

Rita Singh

Carnegie Mellon University

In this lecture

- Digital multimedia: Recording and devices
  - Audio
  - Images
  - Video
  - Text
- Digital multimedia: Processing
  - Audio processing
  - Two generic processing techniques

Representing text

- **This is a cat**
- Represented as the binary string: 01110100 01101000 01101001 01110011 00100000 01101001 01110011 00100000 01100001 00100000 01100011 01100001 01110100
  - This representation is 8-bit ASCII representation
  - Many online tools available for conversion, e.g. http://codebeautify.org/string-binary-converter
- Text to Hex string: 7468697320697320612063617 4
  - text equivalent can be seen using, e.g., http://tomeko.net/online_tools/hex_to_file.php?lang=en (hex to file)

Representing text

2^0

1 byte, 8 bits, 2 nibbles

XXXX XXXX

2^3+2^2+2^1+2^0 = 15

Hexadecimal

- 0-9,A-F counts from 0 to 15 (A=10, B=11, … , F=15)
- Each nibble can represent 0 to 15 (i.e 16 numerical values)
  - 1111 1111 => FF
  - 1111 1110 => FE

32 byte text string = 64 nibbles = **64-digit hexadecimal number**

Representing Text: Character Encodings

- When any key on a keyboard is pressed, a code corresponding to the character is stored

  Example of a code: ASCII

  - abbreviated from **American Standard Code for Information Interchange**, is a character encoding standard for electronic communication.

  - Text characters start at **denary** number 0 in the ASCII code
    - This covers special characters including punctuation, the return key and control characters as well as the number keys, capital letters and lower case letters.
    - the letter ‘a’ = binary number 0110 0001 (denary number 97)
    - the letter ‘b’ = binary number 0110 0010 (denary number 98)
    - the letter ‘c’ = binary number 0110 0011 (denary number 99)

  - 8 bit per character, only 128 characters possible enough to cover English

ASCII Table

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]

Representing Text

- To cover languages with larger alphabets, or accented European languages, we need a more extended code
  - **Unicode (UTF-8, UTF-32 and UTF-16 character encoding)**
  - **Can represent all graphic symbols and languages in the world today**

- ASCII is is faster to process than than multi-byte encoding scheme (fewer bits to process)

- Unicode is a character set of various symbols (see table in the next slides)

  - UTF-8, UTF-16 and UTF-32 are different ways to represent unicode
  - UTF-8 and UTF-16 are variable length encoding
    - number of bytes used depends upon the character.
    - UTF-8 uses 1 to 4 bytes
    - UTF-16 uses either 2 or 4 bytes.
  - UTF-32 is fixed width encoding
  - Uses 4 bytes

- The exact format is indicated by a header at the beginning of a text file. It is not visible in usual text editors

Unicode table

Start	End	Description
0000	1FFF	Alphabets
0000	007F	Basic Latin
0080	00FF	Latin-1 Supplement
0100	017F	Latin Extended-A
0180	024F	Latin Extended-B
0250	02AF	IPA Extensions
02B0	02FF	Spacing Modifier Letters
0300	036F	Combining Diacritical Marks
0370	03FF	Greek
0400	04FF	Cyrillic
0530	058F	Armenian
0590	05FF	Hebrew
0600	06FF	Arabic
0900	097F	Devanagari
0980	09FF	Bengali
0A00	0A7F	Gurmukhi
0A80	0AFF	Gujarati
0B00	0B7F	Oriya
0B80	0BFF	Tamil
0C00	0C7F	Telugu
0C80	0CFF	Kannada
0D00	0D7F	Malayalam
0E00	0E7F	Thai
0E80	0EFF	Lao
0F00	0FBF	Tibetan
10A0	10FF	Georgian
1100	11FF	Hangul Jamo
1E00	1EFF	Latin Extended Additional
1F00	1FFF	Greek Extended
2000	2FFF	Symbols and Punctuation
2000	206F	General Punctuation
2070	209F	Superscripts and Subscripts
20A0	20CF	Currency Symbols
20D0	20FF	Combining Marks for

Unicode table

Symbols

2100	214F	Letterlike Symbols
2150	218F	Number Forms
2190	21FF	Arrows
2200	22FF	Mathematical Operators
2300	23FF	Miscellaneous Technical
2400	243F	Control Pictures
2440	245F	Optical Character Recognition
2460	24FF	Enclosed Alphanumerics
2500	257F	Box Drawing
2580	259F	Block Elements
25A0	25FF	Geometric Shapes
2600	26FF	Miscellaneous Symbols
2700	27BF	Dingbats
3000	33FF	CJK Auxiliary
3000	303F	CJK Symbols and Punctuation
3040	309F	Hiragana
30A0	30FF	Katakana
3100	312F	Bopomofo
3130	318F	Hangul Compatibility Jamo
3190	319F	Kanbun
3200	32FF	Enclosed CJK Letters and Months
3300	33FF	CJK Compatibility
4E00	9FFF	

CJK Unified Ideographs Han characters used in China, Japan, Korea, Taiwan, and Vietnam

AC00	D7A3	Hangul Syllables
D800	DFFF	Surrogates
D800	DB7F	High Surrogates
DB80	DBFF	High Private Use

Unicode table

Surrogates

DC00	DFFF	Low Surrogates
E000	F8FF	Private Use
F900	FFFF	Miscellaneous
F900	FAFF	CJK Compatibility Ideographs
FB00	FB4F	Alphabetic Presentation Forms
FB50	FDFF	Arabic Presentation Forms-A
FE20	FE2F	Combining Half Marks
FE30	FE4F	CJK Compatibility Forms
FE50	FE6F	Small Form Variants
FE70	FEFE	Arabic Presentation Forms-B
FEFF	FEFF	Specials
FF00	FFEF	Halfwidth and Fullwidth Forms
FFF0	FFFF	Specials

UTF-8

- UTF-8 is a variable-width or "multi-byte" encoding format; this means that different characters require different numbers of bytes. In UTF-8, the standard ASCII characters occupy only one byte, and remain untouched by the encoding (i.e., a string of ASCII characters is a legal UTF-8 string). As a tradeoff, however, other Unicode characters occupy two or three bytes.

- In UTF-8, Unicode characters between \u0000 and \u007F occupy a single byte, which has a value of between 0x00 and 0x7F, and which always has its high-order bit set to 0. Characters between \u0080 and \u07FF occupy two bytes, and characters between \u0800 and \uFFFF occupy three bytes. The first byte of a two-byte character always has high-order bits 110, and the first byte of a three-byte character always has high-order bits 1110. Since single-byte characters always have 0 as their high-order bit, the one-, two-, and three-byte characters can easily be distinguished from each other.

- The second and third bytes of two- and three-byte characters always have high-order bits 10, which distinguishes them from one-byte characters, and also distinguishes them from the first byte of a two- or three-byte sequence. This is important because it allows a program to locate the start of a character in a multi-byte sequence.

- The remaining bits in each character (i.e., the bits that are not part of one of the required high-order bit sequences) are used to encode the actual Unicode character data. In the single-byte form, there are seven bits available, suitable for encoding characters up to \u007F. In the two-byte form, there are 11 data bits available, which is enough to encode values to \u07FF, and in the three-byte form there are 16 available data bits, which is enough to encode all 16-bit Unicode characters. Table 11.2 summarizes the UTF-8 encoding.

UTF-8 Encoding

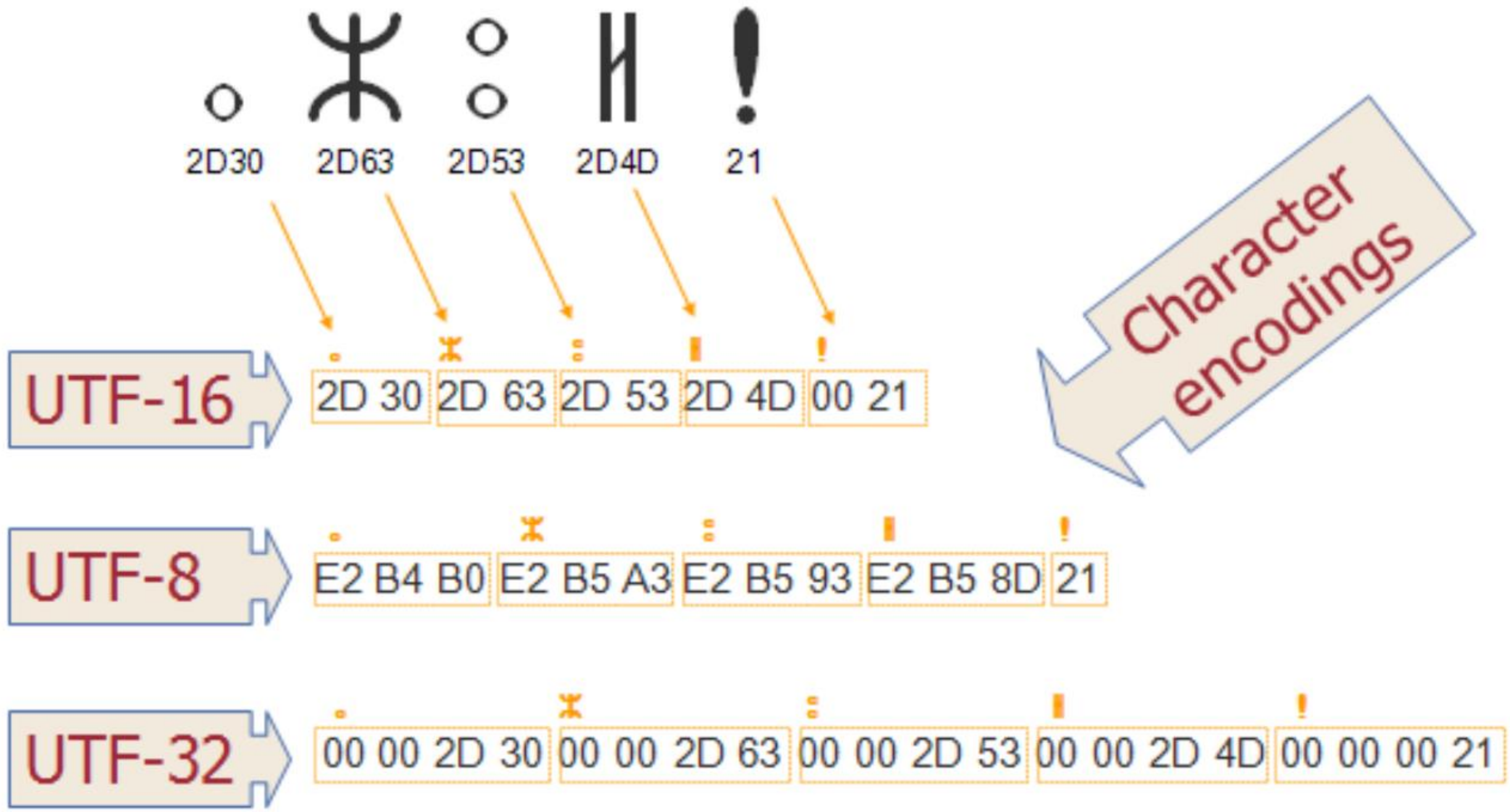
Start Character	End Character	Required Data Bits	Binary Byte Sequence (x = data bits)
\u0000	\u007F	7	0xxxxxxx
\u0080	\u07FF	11	110xxxxx 10xxxxxx
\u0800	\uFFFF	16	1110xxxx 10xxxxxx 10xxxxxx

UTF-8

- UTF8 Desirable features

- All ASCII characters are one-byte UTF-8 characters. A legal ASCII string is a legal UTF-8 string.
  - UTF-8 is compatible with ASCII while UTF-16 is incompatible with ASCII
- Any non-ASCII character (i.e., any character with the high-order bit set) is part of a multi-byte character.
- The first byte of any UTF-8 character indicates the number of additional bytes in the character.
- The first byte of a multi-byte character is easily distinguished from the subsequent bytes. Thus, it is easy to locate the start of a character from an arbitrary position in a data stream.
- It is easy to convert between UTF-8 and Unicode.
- The UTF-8 encoding is relatively compact. For text with a large percentage of ASCII characters, it is more compact than Unicode. In the worst case, a UTF-8 string is only 50% larger than the corresponding Unicode string.

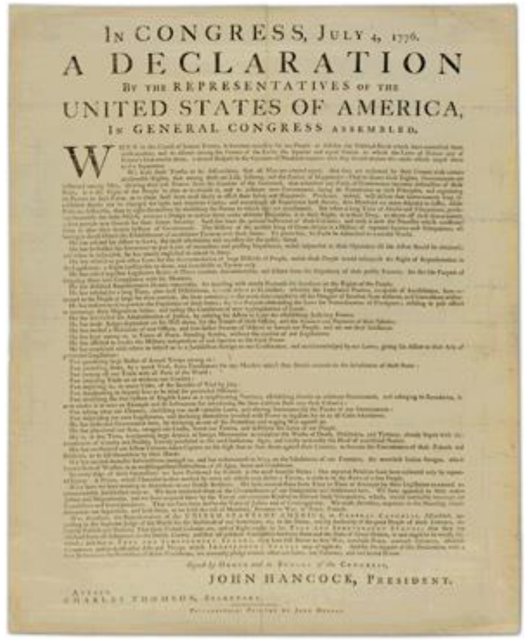
A puzzle to solve



Some simple techniques in text processing

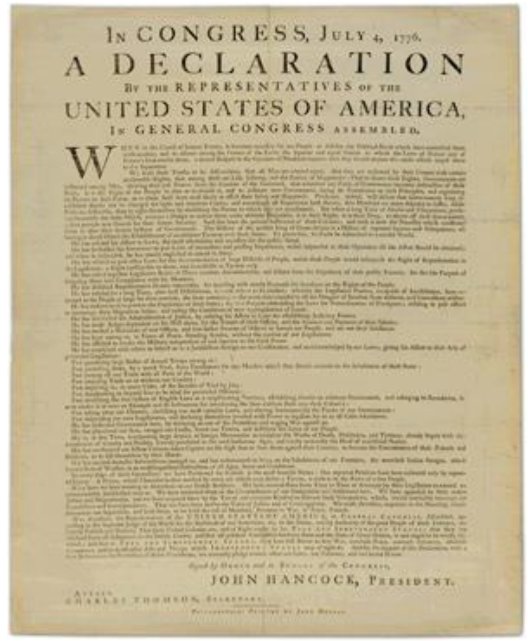
- Document representation
- Extracting features for processing

Problem



- A document is a long variable-length sequence of words and other symbols
  - Not directly amenable to mathematical analysis

Problem


IN CONGRESS, JULY 4, 1776.

A DECLARATION

BY THE REPRESENTATIVES OF THE

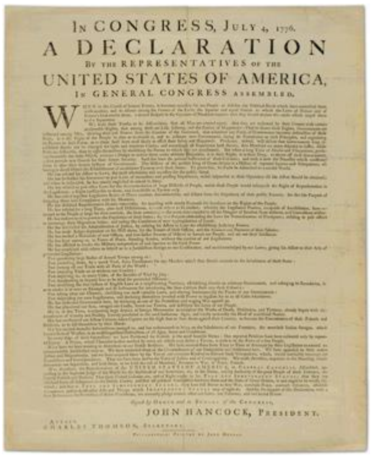
UNITED STATES OF AMERICA,

IN GENERAL CONGRESS ASSEMBLED.

JOHN HANCOCK, PRESIDENT.

- It must be converted to a numerical representation
  - An embedding
- Objective: Convert variable length text document into a fixed-length real-valued vector
  - A *meaningful* representation that makes both intuitive and arithmetic sense

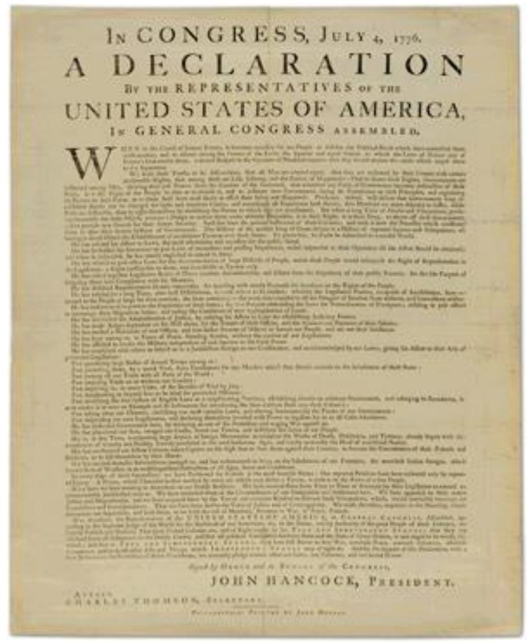
The bag-of-words representation for documents



[4	1	1	1	1	1	1	2	4	1	3	1]
a	an	be	free	govern	hear	in	justice	liberty	many	oppressed	self

- A *word-count* vector
  – Maintains counts of words
  – Ignores *order* in which words occur

Problem

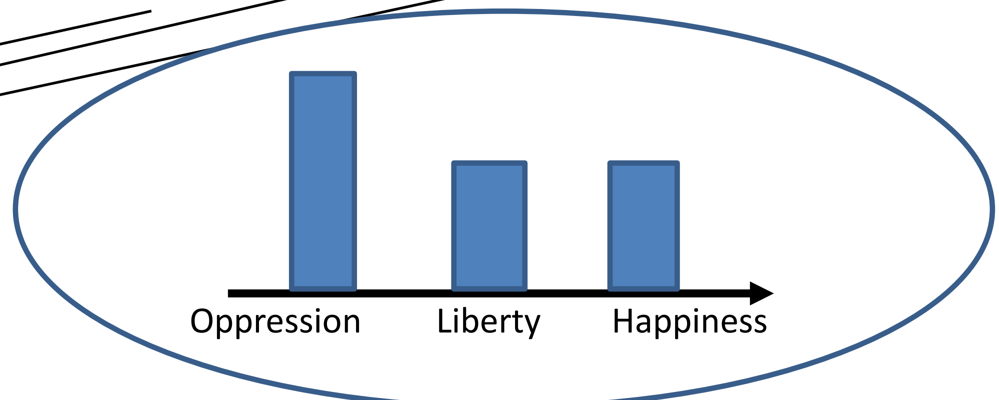

IN CONGRESS, JULY 4, 1776.
A DECLARATION
BY THE REPRESENTATIVES OF THE
UNITED STATES OF AMERICA,
IN GENERAL CONGRESS ASSEMBLED.

JOHN HANCOCK, PRESIDENT.

- Word embeddings: **neural network or other techniques convert sentences to vector representation**
  - BERT (RoBERTa) or ELMO for word embeddings
  - Both are context dependent
  - Older: Word2Vec ion that makes both intuitive and arithmetic sense

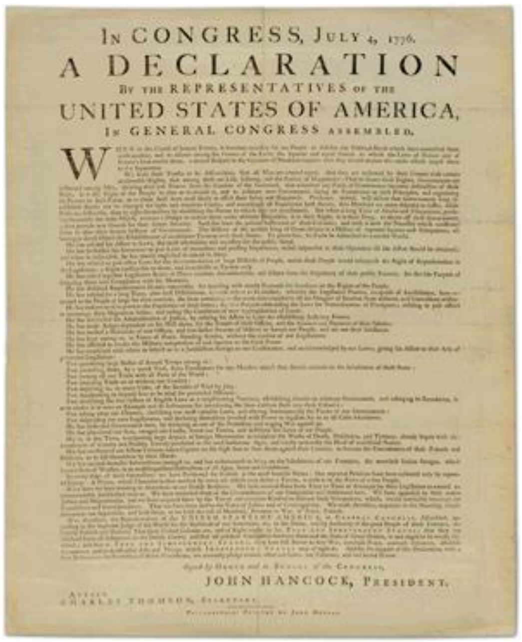
The Topic Model representation

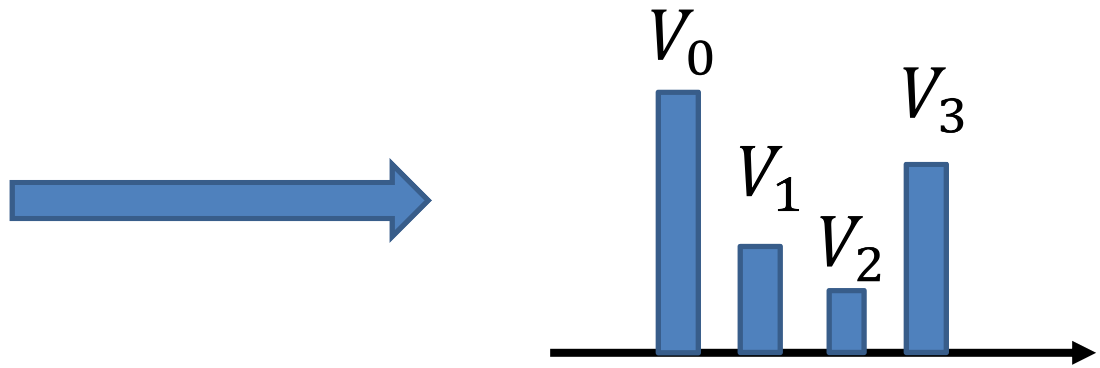
Oppression Liberty Happiness



- Every document combines a number of "topics"

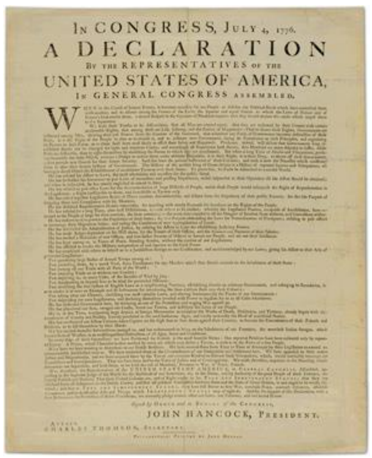
Vector space representation of documents

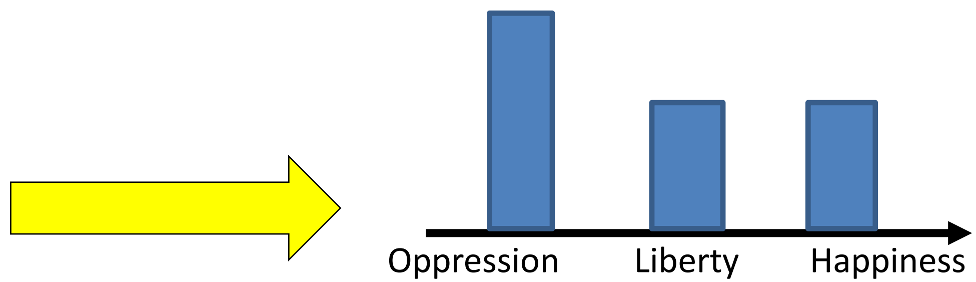

IN CONGRESS, JULY 4, 1776.
A DECLARATION
BY THE REPRESENTATIVES OF THE
UNITED STATES OF AMERICA,
IN GENERAL CONGRESS ASSEMBLED.
JOHN HANCOCK, PRESIDENT.



- Vector space representation: “Amount of each topic” present in the document

Vector space representation as topics

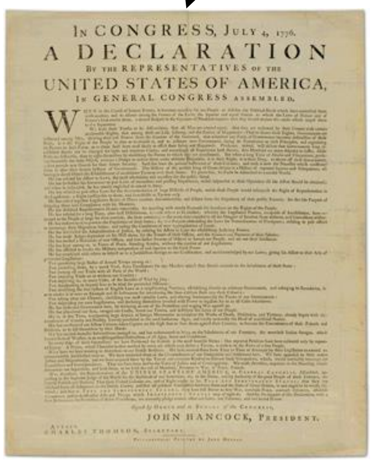




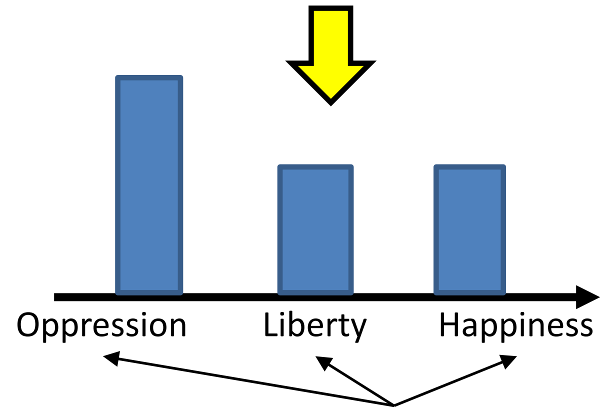
- Example: 3-D vector of topics can be used as a mathematical representation of the document

Generic Challenges

How do you represent the document


IN CONGRESS, JULY 4, 1776.
A DECLARATION
BY THE REPRESENTATIVES OF THE
UNITED STATES OF AMERICA,
IN GENERAL CONGRESS ASSEMBLED.
JOHN HANCOCK, PRESIDENT.

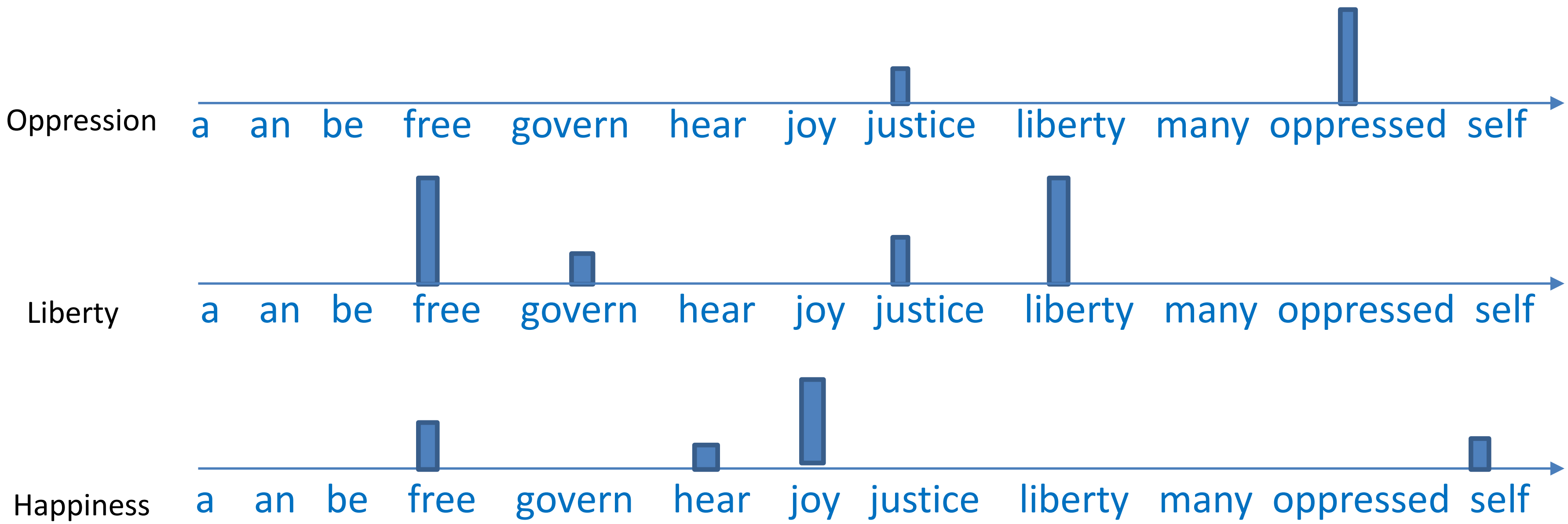
How do you compute these proportions



What *are* these topics?
How do you represent them?

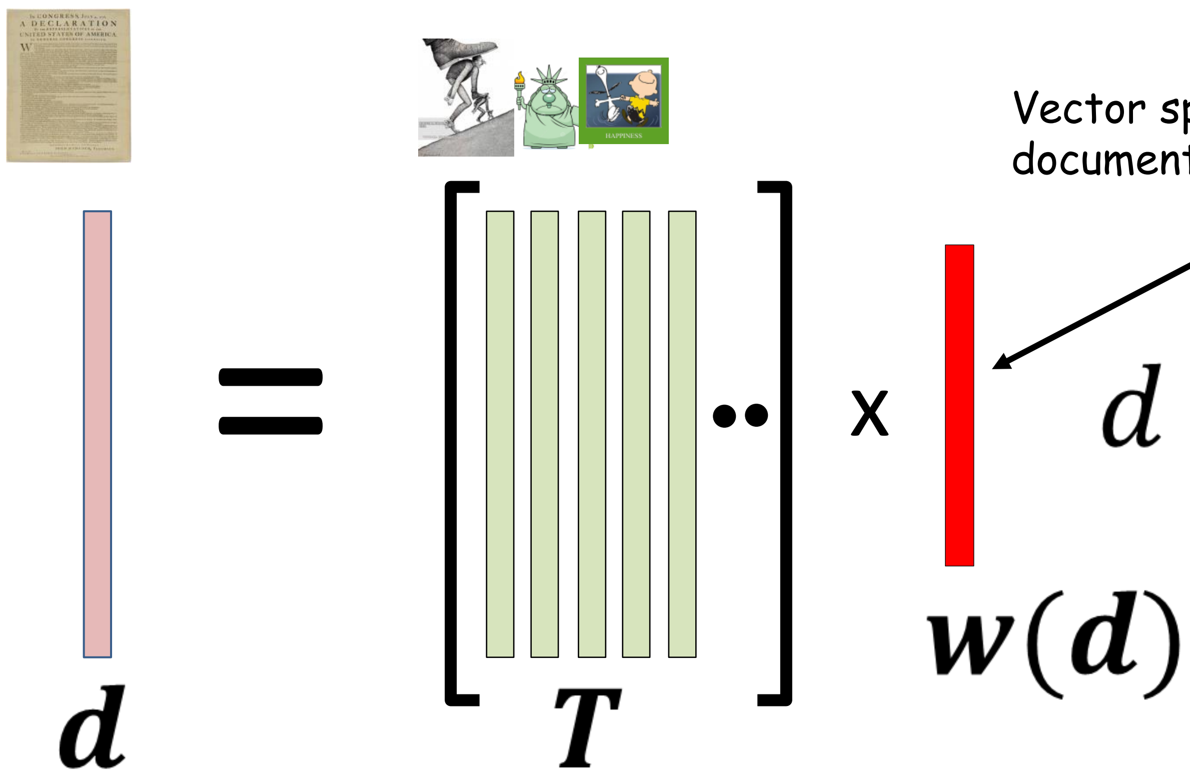
Addressed by topic modeling techniques

Topics



- "Topics" are also represented by vectors of words
  - Representing words and the frequency with which they occur when that topic is discussed

Simple vector space model

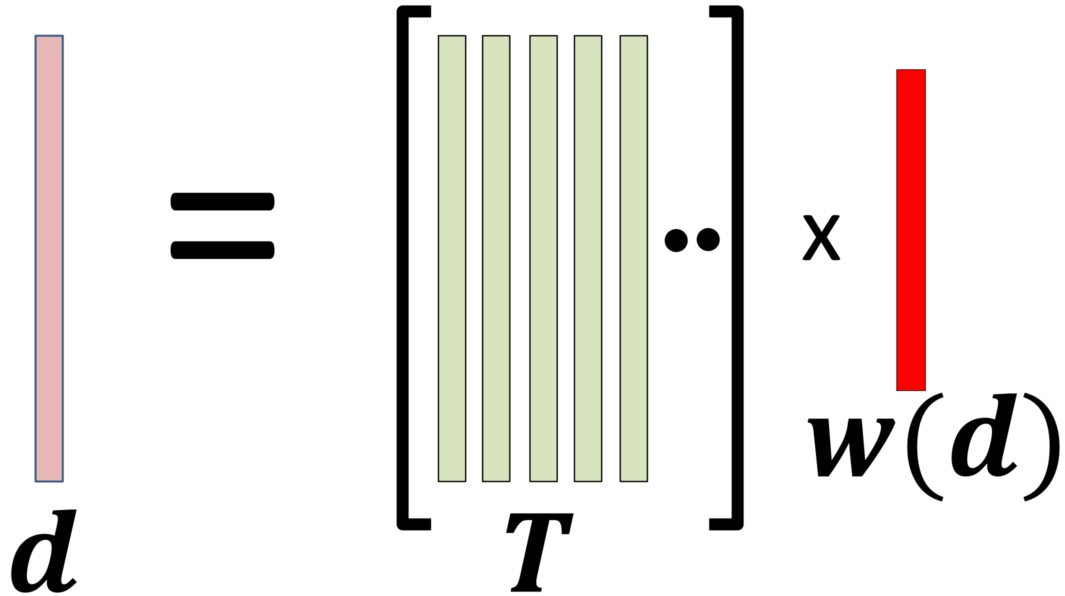


Vector space representation of document (aka what we want)

$$d = Tw(d)$$

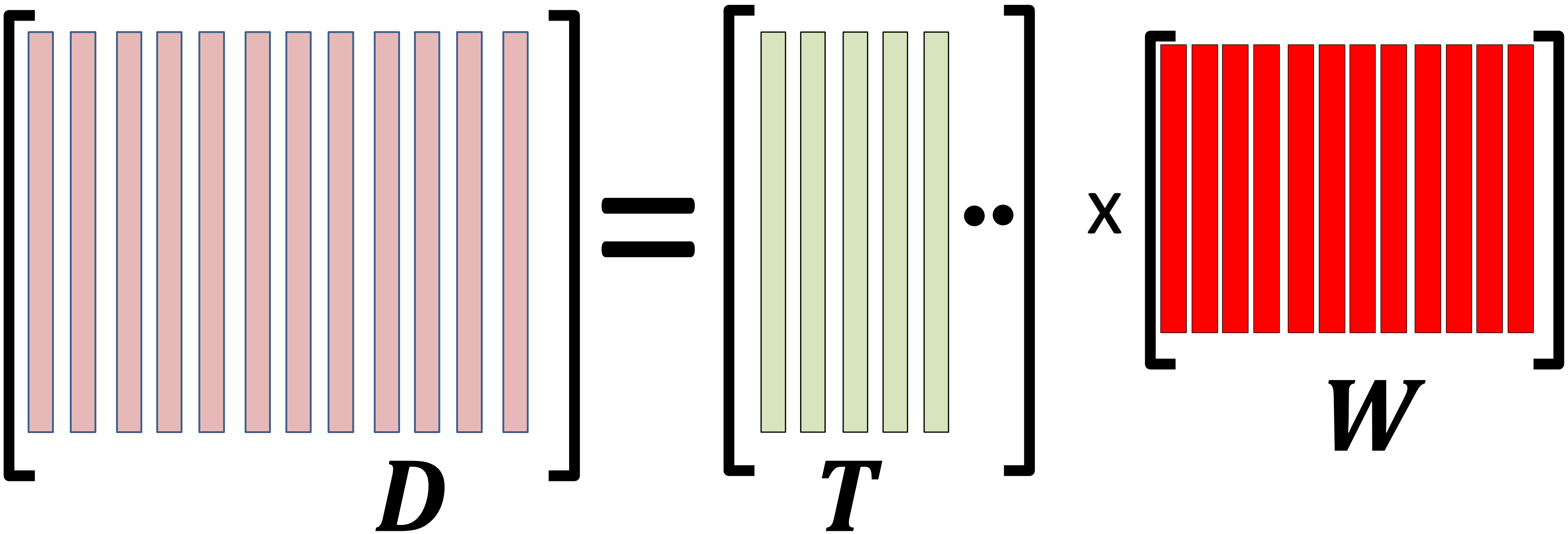
- Document vector is a weighted combination of topics
  - Objective: For each document ***d*** find a nice vector-space representation ***w(d)***

Challenge



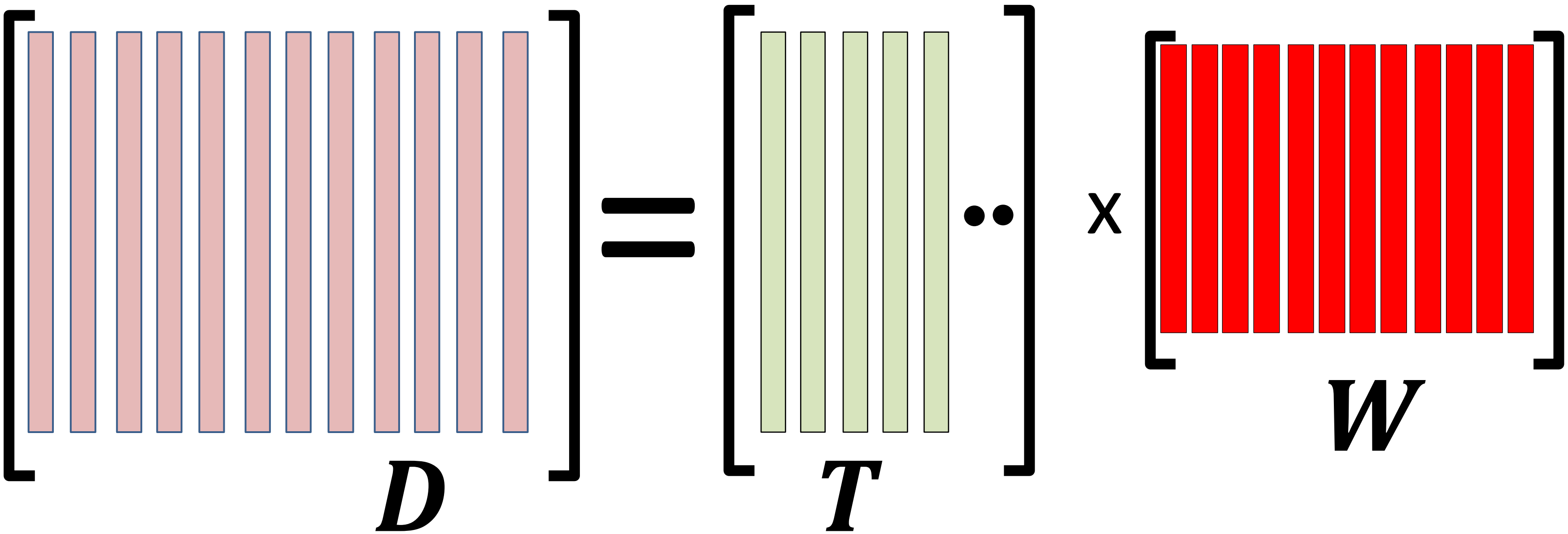
- That was an easy example
- In the real world, the list of topics (and their word representation) is unknown (and potentially unlimited)
  - Even the topics must be *inferred* from analysis of data
  - Basically, T must be *learned* from data

Updated challenge: discovering topics



- Given a collection **D** of documents
- Find set **T** of topics, such that
- Every document in **D** is **well approximated** as a weighted combination of topics
  - In the equation each column **d** of **D** has a corresponding column **w** in **W** which gives the weights to combine topics

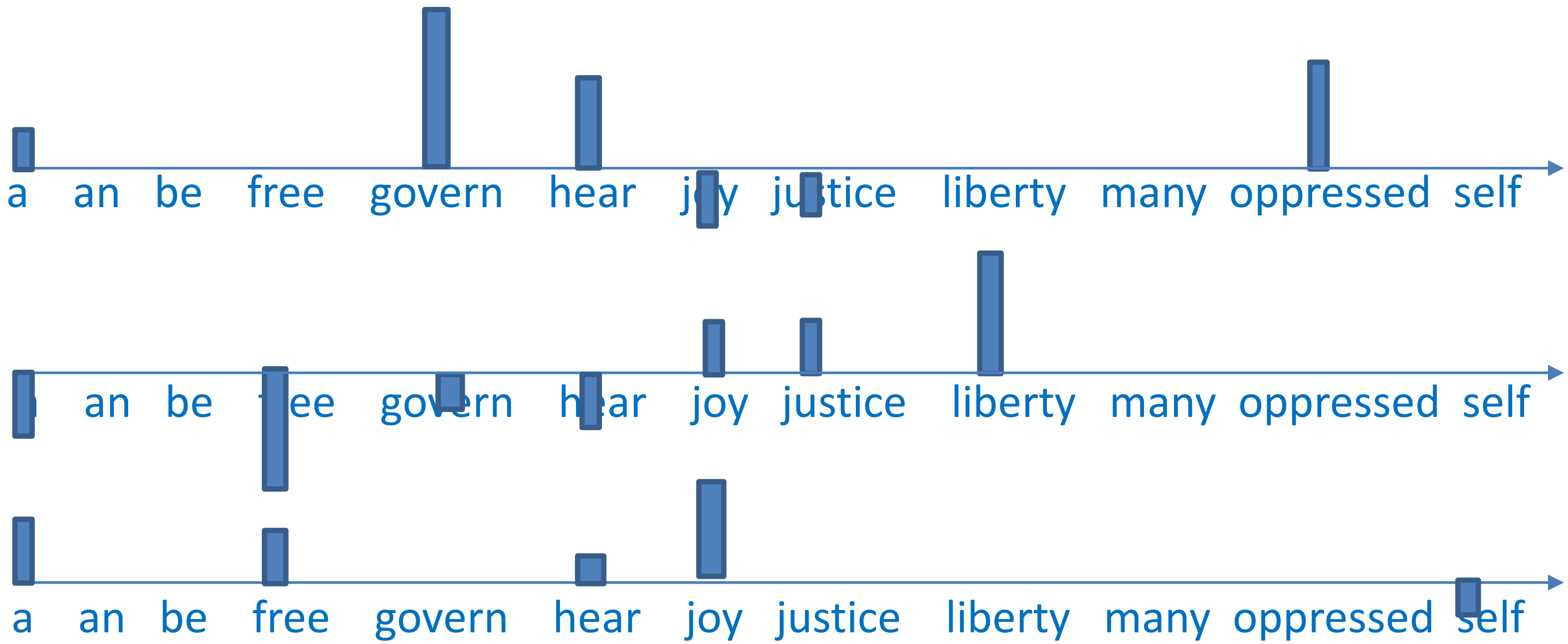
Updated challenge



$$D = TW$$

- **Given D, find T and W**
- This is like PCA
  - The original method to find "topics" automatically *did* use PCA
  - Called ***Latent Semantic Analysis***

Topics



- Unfortunately this ignores that D, T and W are non-negative
  - "Learned" topics will have both positive and negative components..........what do negative words mean?

- So we use ***Latent Dirichlet Allocation***

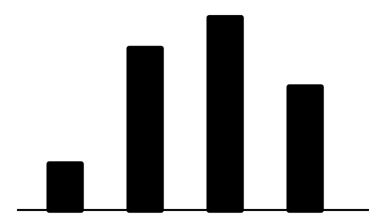
Example from David Blei

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Four "topics" learned from a collection of AP articles.

Analysis of composition of an article about a donation made to a school

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.



Topic distribution for article

Detailed info in probabilistic topic models papers by David Blei (many papers)

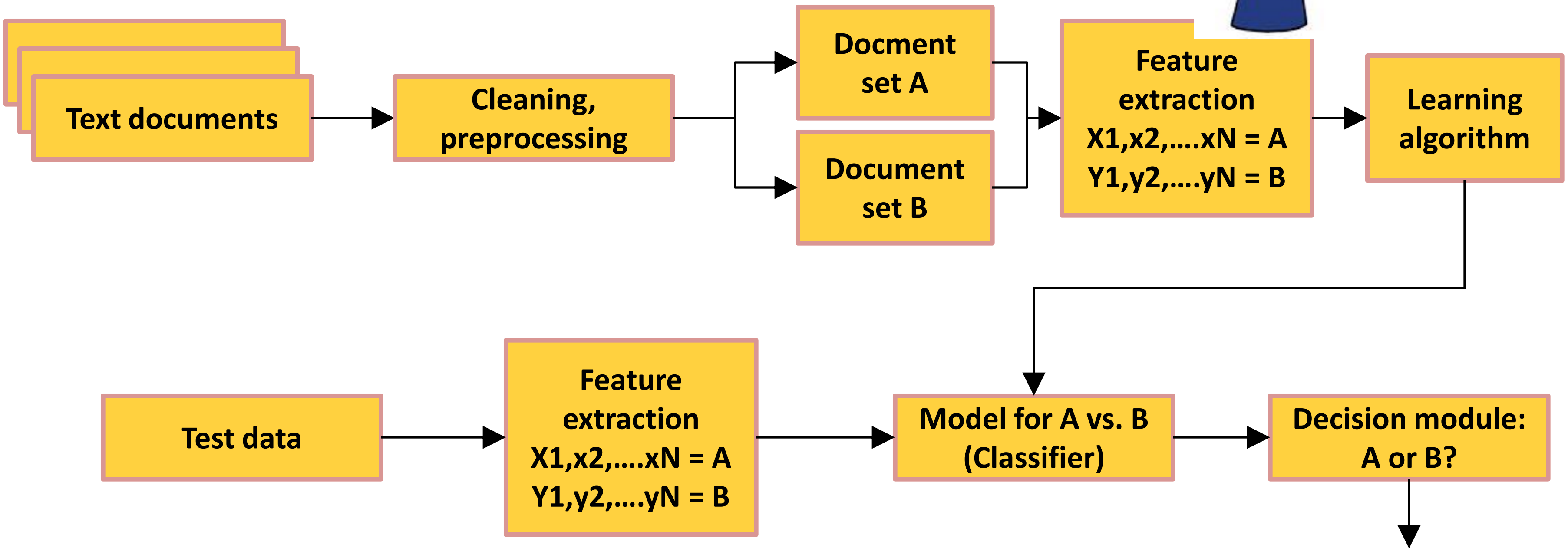
Topic detection using neural nets

- Known topic vectors and corresponding document vectors are given to neural nets
  - Neural nets learn to classify them

Examples of uses of text processing

- **Information extraction**
- **Information retrieval**
- **Detection**
  - misinformation and disinformation: tampering, topic or context misrepresentation, fake text, forgery, plagiarism, fake accounts, authorship issues, Text bots / Cyborgs / Viral bots in social media (twitter example)
- **Tracking**
  - Author profiling, Tracking perpetrators in human trafficking, stolen credentials, tracking origins of email etc.
- **Recovering** encrypted information
  - Cryptography and steganography: uncovering hidden information in text

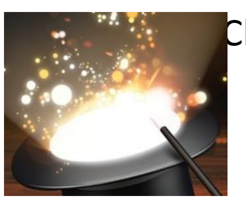
Processing text for classific n



E

Example: Authorship attribution

- Classifiers are trained to recognize patterns that are characteristic of authors
  - Examples and counterexamples of author's work are used to train each classifier
  - Classifier does not directly learn from text
    - learns from *features* (or attributes) extracted from text
    - **The magic is in the features!**
    - Given a new document, the classifier **identifies which of the authors it is trained on** wrote it



Stylometry

- The science of measuring literary style
  - Writing style is very individual and influenced by:
    - Age, gender, education level, native language, personality, emotional state, mental health, region, deception, ideology, political conviction, religious beliefs ....
- Stylometry is done by capturing an author’s **distinguishing styles, e.g.**
  - Usage patterns of rare (and noticeable) words
  - Usage patterns of frequent words (e.g. "to", "with“, “in” etc.)?
    - "It's much harder for someone to imitate my frequency pattern of 'but' and 'in’.” John Burrows - English professor of the University of Newcastle

Stylometry

Quantifying writing style: ***The magic is in the features! These are also called STYLE MARKERS***

- **Lexical and character features:** Consider text as a sequence of entities or tokens; text is represented as a vector of counts of these entities or their combinations
- **Vocabulary distributions**
- **Token and Type Lengths**
  - Token : **All words**
  - Type : **Unique words**
    - For the sentence "I cannot **bear** to see a **bear**"
    - **7 tokens, 6 (context-free) types**
- **Syllable Count in Tokens**
- **Syllable Count in Types**
- Words, characters, punctuation marks, spaces
- **Word frequencies**
  - **Frequency of Most Frequent Words**

Stylometry

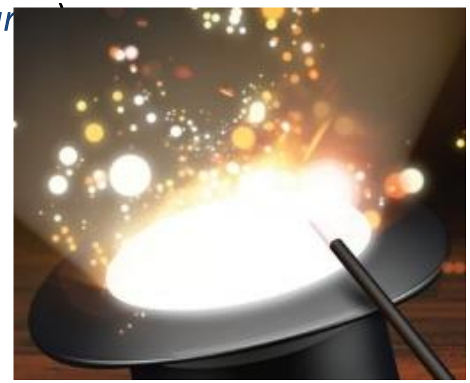
Quantifying writing style (contd)

- **Complexity measures** (**e.g**. average sentence length, average syllables per word, average word length, character frequencies, word frequencies, vocabulary richness functions, Sentence Lengths in absolute)
- **Idiosyncrasies** (a **distinctive writing feature**)
  - "errors" *identified by* spelling/grammar checker *: e.g. spelling errors, neologisms, unexpected syntax*

- **Syntactic and semantic features:** Derived through deeper linguistic analysis
  - **Morphology** (grammatical prefixes and suffixes, e.g. *–ing, re-*).
  - **Distribution of parts of speech**
  - **Function word usage**
  - **Function words**
  - **Content words**
    - E.g. character n-grams:
  - **Unstable words** (features that might be replaced in a rewrite, e.g. *huge:lar*

- **Application specific features**
  - Can be defined only in certain text domains

- 1000 different measures were estimated by Rudman (1998)



Stylometry

- Learning algorithms
  - k-NN (k Nearest Neighbor)
  - Bayesian analysis (Naïve Bayes)
  - SVM (Support Vector Machines)
  - Markovian Models
  - Neural Networks
  - Decision Trees
  - Etc..
- Lessons
  - Content words and character n-grams work well
  - For unedited texts, idiosyncrasies are best
  - For some languages like Arabic, morphology is needed
  - Etc…

More applications

- Some more examples
  - Stylometry and authorship attribution
  - Author profiling
  - Investigating human trafficking
  - Investigating stolen credentials
  - Detecting fake accounts
  - Forensic document analysis
  - Uncovering hidden information in text

Author profiling

Useful when prior examples of authorship are not available to learn classifiers from

- Identifying personal traits
  - Gender
  - Age
  - Personality traits
  - Native language
  - Etc..

What are the Distinguishing Features?

- Fiction
  - Male: *a, the, as*
  - Female: *she, for, with, not*
- Non-Fiction
  - Male: *that, one, of*
  - Female: *she, for, with, and, in*
- *How do we determine the above?*
  - *Learning based feature reduction*
    - Apply a learning algorithm
    - Eliminate features with low weights
    - Learn again

J. W. Pennebaker. The Secret Life of Pronouns: What Our Words Say about Us. Bloomsbury USA, 2013.

- **Males** use more Informational features
  - Determiners
  - Adjectives
  - of modifiers (e.g. pot of gold)
- **Females** use more Involvedness features
  - Pronouns
  - for and with
  - Negation
  - Present tense

Gender: female/male?

- My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

[examples: Moshe Koppel]

Gender: female/male?

Female

- *My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .*

Male

- *The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.*

+ PRONOUNS -

Gender: female/male?

Female

- *My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .*

Male

- *The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.*

+ PRESENT TENSE -

Gender: female/male?

Female

- *My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .*

Male

- *The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.*

+ NEGATION -

Gender Guesser

Black Hat Briefings, USA 2006

Gender Guesser

The words you use can disclose identifying features. This tool attempts to determine an author's gender based on the words used.

Submitted text is evaluated based on two types of writing: formal and informal. Formal writing includes fiction and non-fiction stories, articles, and news reports. Informal writing includes blog and chat-room text. (Email can be formal, informal, or some combination.) You should view the results based on the appropriate type of writing.

Analyze

Type or paste a writing sample for gender analysis. Then click on "Analyze" to see the results. For best performance, use at least 300 words -- more words is generally more accurate.

```
In 2003, a team of researchers developed a method to estimate gender from word
usage. Their paper described a Bayesian network where weighted word frequencies
and parts of speech could be used to estimate the gender of an author. Their
approach made a distinction between fiction and non-fiction writing styles.
```

Analyze Clear About

Results

```
Total words: 338
```

```
Genre: Informal
  Female = 268
  Male   = 898
  Difference = 630; 77.01%
  Verdict: MALE
```

```
Genre: Formal
  Female = 475
  Male   = 468
  Difference = -7; 49.62%
  Verdict: Weak FEMALE

Weak emphasis could indicate European.
```

- **Try it at: http://www.hackerfactor.com/GenderGuesser.php**

Other profile parameters

- Age
- Native language
  - **Yesterday we had our second jazz competition. Thank God we weren't competing. We were sooo bad. Like, I was so ashamed, I didn't even want to talk to anyone after. I felt so rotton, and I wanted to cry, but...it's ok.**
    - Teen American Female

- Personality traits
  - Pennebaker data:
    - Students written essays
    - Same students took personality assessment tests
  - Experiment: **Given text, determine if author is**
    - **Open**
    - **Conscientious**
    - **Neurotic**
    - **Extroverted**
    - **Agreeable**

Projects

- Detecting the existence and source of deepfake multimedia
  – Audio
  – Images
  – Video
- Human-guided AI for multimedia generation, indexing and classification
  – Emphasis on new AI architectures