Applied Natural Language Processing

Info 256

Lecture 24: Information extraction (Nov. 20, 2023)

David Bamman, UC Berkeley

Project presentations next Monday!

- Create a single slide (pdf) representing your work and discuss it in 3 minutes.
- Your slide should outline your research goals, any data you've used, models, results, and analysis with enough detail for others to provide feedback on your project from that slide alone.
- Submit to bCourses by noon on Monday 11/27!

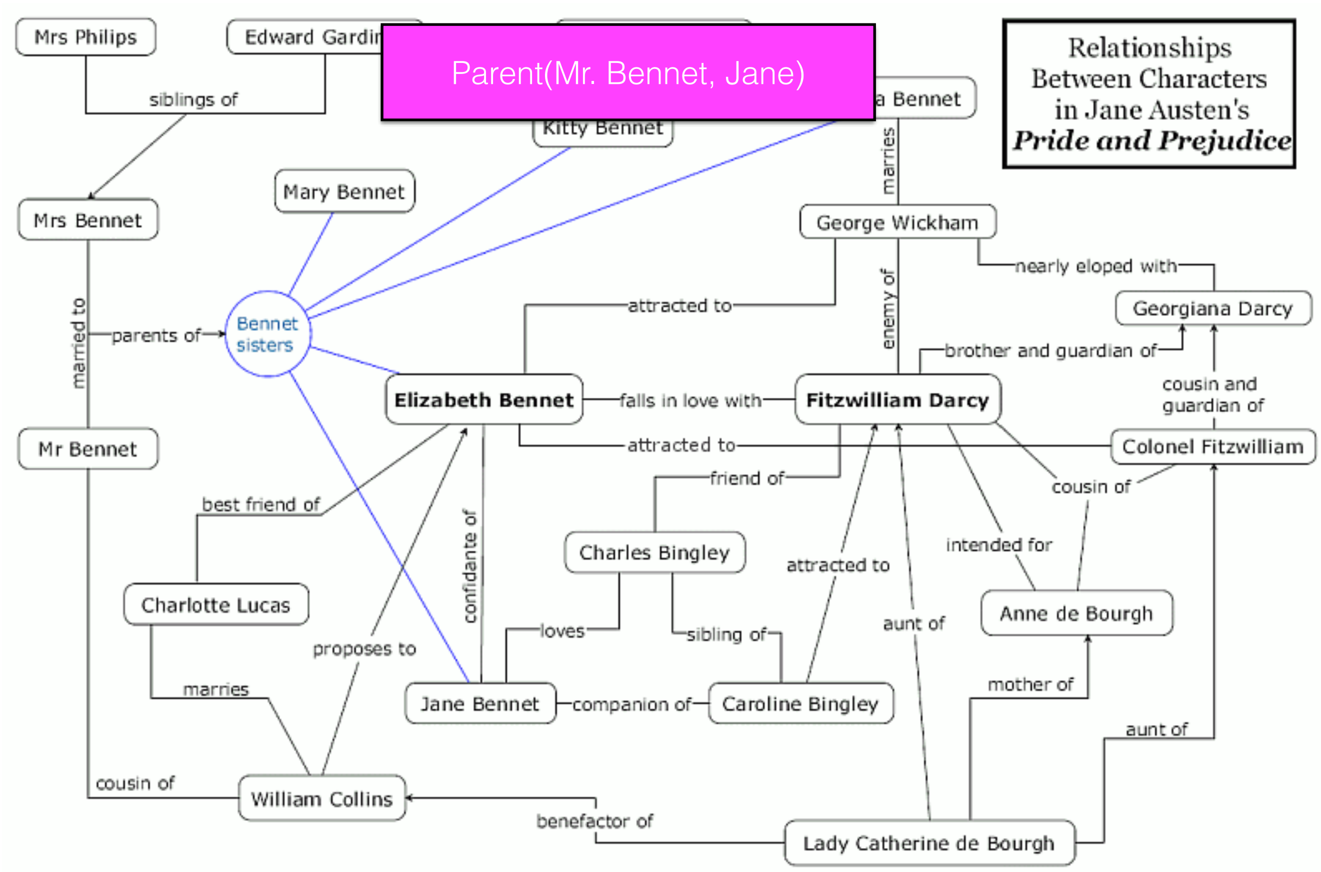
Investigating(SEC, Tesla)

LEARN TO INVEST
Don’t know where to start? Sign up for the Investing Basics newsletter.
Ad

SEC Probing Tesla CEO Musk's Tweets: Reports

By Deborah DSouza | Updated August 9, 2018 — 4:47 AM EDT

On Tuesday, Tesla Inc. (TSLA) CEO Elon Musk made the dramatic announcement that he was considering taking Tesla private for $420 a share on Twitter. In an email sent to Tesla employees posted on the company's official blog, Musk explained that he is mulling taking the firm private to protect it from short sellers and wild swings in stock prices. However, the email didn't provide any details regarding financing. (See also: What if Tesla Goes Private?)



Parent(Mr. Bennet, Jane)

https://en.wikipedia.org/wiki/Pride_and_Prejudice

Information extraction

- Named entity recognition
- Entity linking
- Relation extraction

Named entity recognition

[tim cook]$_{PER}$ is the ceo of [apple]$_{ORG}$

- Identifying spans of text that correspond to typed entities

Relation extraction

The Big Sleep is a 1946 film noir directed by Howard Hawks,[2][3] the first film version of Raymond Chandler's 1939 novel of the same name. The film stars Humphrey Bogart as private detective Philip Marlowe and Lauren Bacall as Vivian Rutledge in a story about the "process of a criminal investigation, not its results."[4] William Faulkner, Leigh Brackett and Jules Furthman co-wrote the screenplay.

subject	*predicate*	*object*
The Big Sleep	directed_by	Howard Hawks
The Big Sleep	stars	Humphrey Bogart
The Big Sleep	stars	Lauren Bacall
The Big Sleep	screenplay_by	William Faulkner
The Big Sleep	screenplay_by	Leigh Brackett
The Big Sleep	screenplay_by	Jules Furthman

Relation extraction

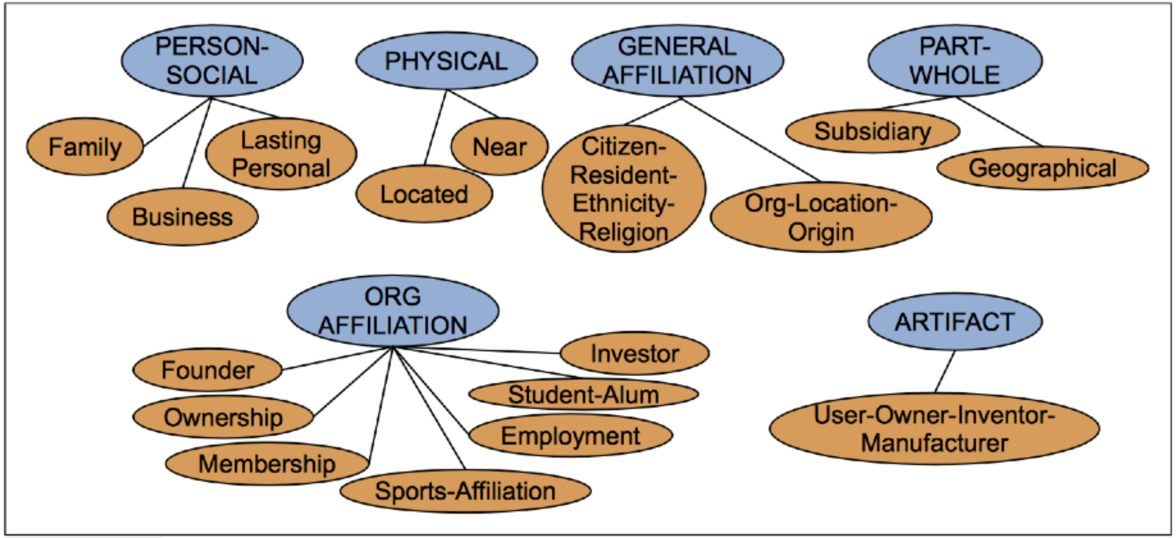


Figure 17.9 The 17 relations used in the ACE relation extraction task.

ACE relations, SLP3

Relation extraction

Entity	Relation	Entity
Injury	disrupts	Physiological Function
Bodily Location	location-of	Biologic Function
Anatomical Structure	part-of	Organism
Pharmacologic Substance	causes	Pathological Function
Pharmacologic Substance	treats	Pathologic Function

Unified Medical Language System (UMLS), SLP3

Regular expressions

- Regular expressions are precise ways of extracting high-precision relations
    - “NP_1 is a film directed by NP_2” → directed_by(NP_1, NP_2)
    - “NP_1 was the director of NP_2”→ directed_by(NP_2, NP_1)

Hearst patterns

pattern	*sentence*
NP {, NP}* {,} (and|or) other NP_H	temples, treasuries, and other important civic buildings
NP_H such as {NP,}* {(or|and)} NP	red algae such as Gelidium
such NP_H as {NP,}* {(or|and)} NP	such authors as Herrick, Goldsmith, and Shakespeare
NP_H {,} including {NP,}* {(or|and)} NP	common-law countries, including Canada and England
NP_H {,} especially {NP}* {(or|and)} NP	European countries, especially France, England, and Spain

Hearst 1992; SLP3

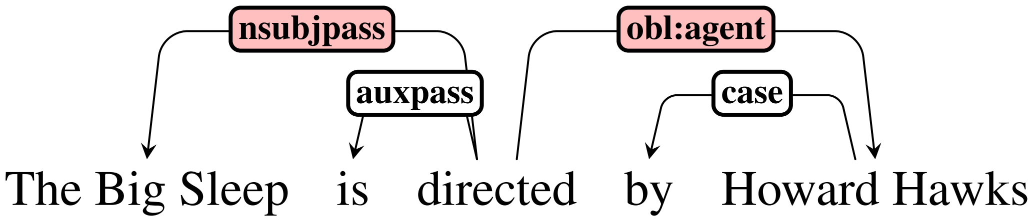
Supervised relation extraction

[The Big Sleep]$_{m1}$ is a 1946 film noir directed by [Howard Hawks]$_{m2}$, the first film version of Raymond Chandler's 1939 novel of the same name.

feature(m1, m2)
headwords of m1, m2
bag of words in m1, m2
bag of words between m1, m2
named entity types of m1, m2
syntactic path between m1, m2

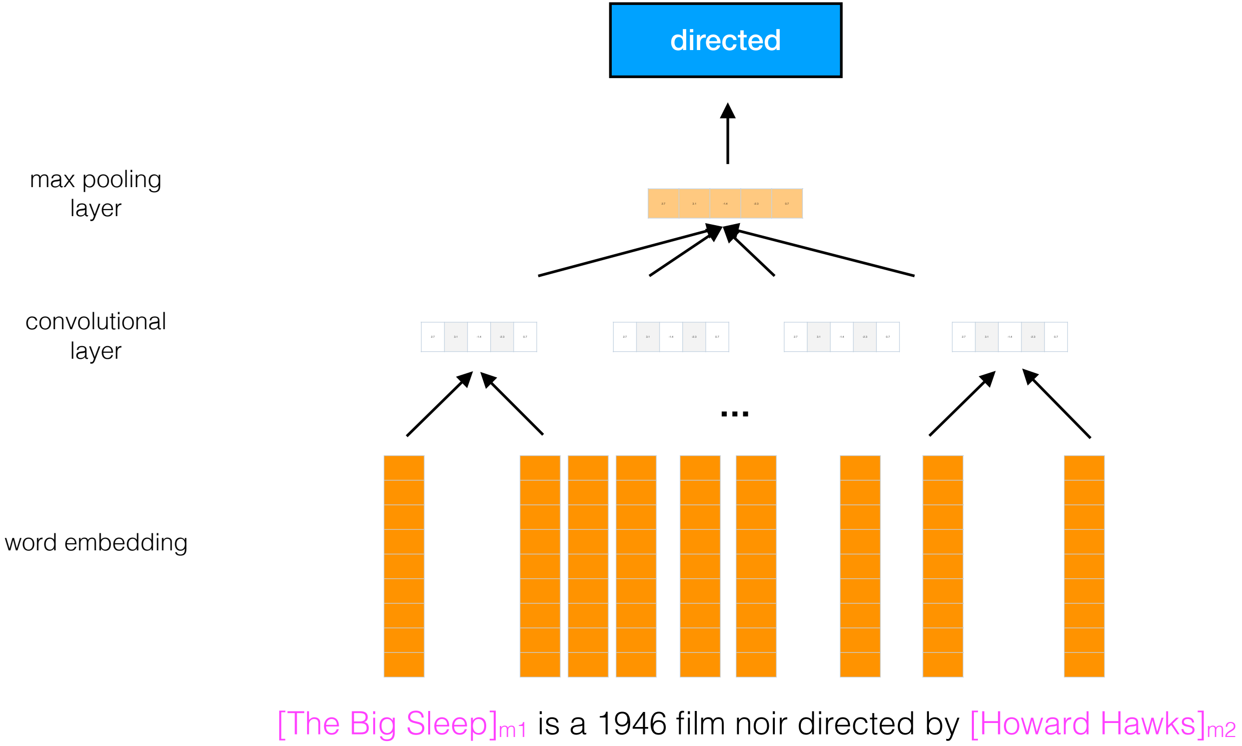
Supervised relation extraction

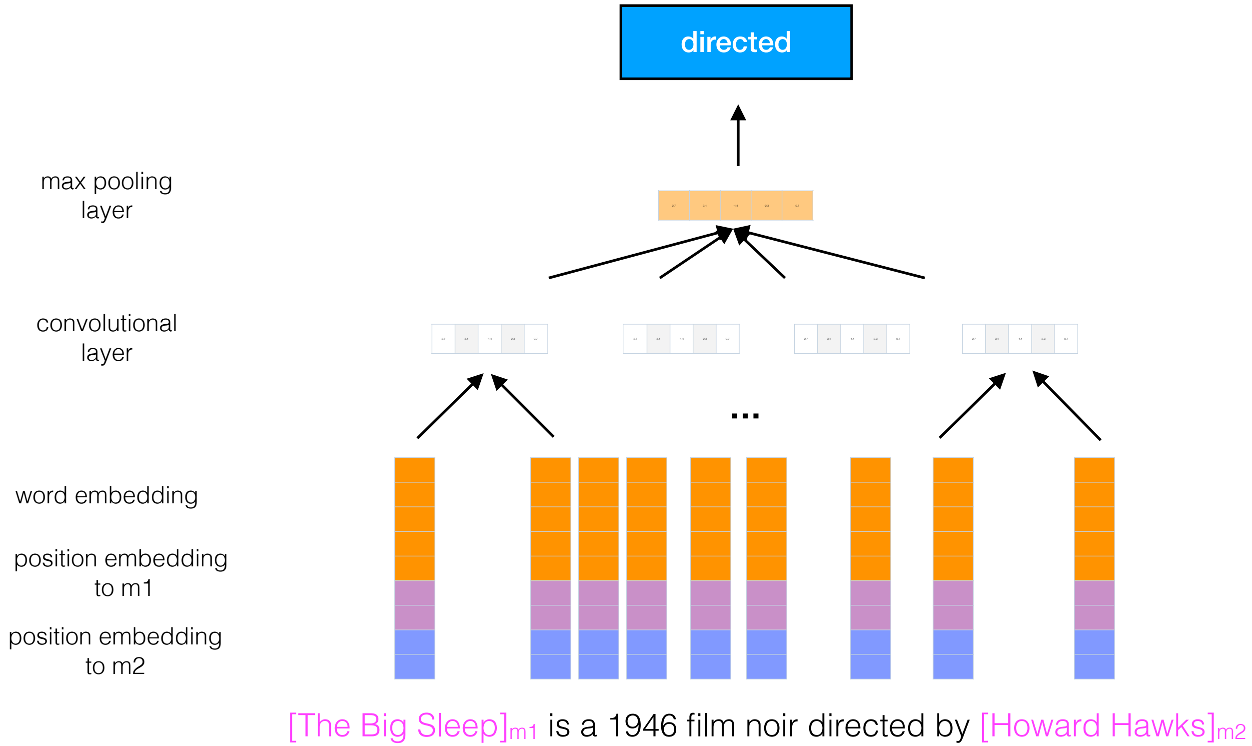
[The Big Sleep]$_{m1}$ is a 1946 film noir directed by [Howard Hawks]$_{m2}$, the first film version of Raymond Chandler's 1939 novel of the same name.



[The Big Sleep]$_{m1}$ ←*nsubjpass* directed→*obl:agent* [Howard Hawks]$_{m2}$,

m1←*nsubjpass* ← directed→*obl:agent* → m2


directed
max pooling layer
convolutional layer
...
word embedding
[The Big Sleep]m1 is a 1946 film noir directed by [Howard Hawks]m2


directed
max pooling layer
convolutional layer
...
word embedding
position embedding to m1
position embedding to m2
[The Big Sleep]m1 is a 1946 film noir directed by [Howard Hawks]m2

Distant supervision

- It’s uncommon to have labeled data in the form of <sentence, relation> pairs

sentence	*relations*
[The Big Sleep]$_{m1}$ is a 1946 film noir directed by [Howard Hawks]$_{m2}$, the first film version of Raymond Chandler's 1939 novel of the same name.	directed_by(The Big Sleep, Howard Hawks)

Distant supervision

- More common to have knowledge base data about entities and their relations that's separate from text.
- We know the text likely expresses the relations somewhere, but not exactly where.

Wikipedia Infoboxes

The Big Sleep is a 1946 film noir directed by Howard Hawks,[2][3] the first film version of Raymond Chandler's 1939 novel of the same name. The film stars Humphrey Bogart as private detective Philip Marlowe and Lauren Bacall as Vivian Rutledge in a story about the "process of a criminal investigation, not its results."[4] William Faulkner, Leigh Brackett and Jules Furthman co-wrote the screenplay.

A remake starring Robert Mitchum as Philip Marlowe was released in 1978. This was the second film in three years featuring Mitchum as Marlowe. The remake was arguably more faithful to the novel, possibly due to fewer restrictions in 1978 on what could be portrayed on screen, however, it was far less successful than the original 1946 version. In 1997, the U.S. Library of Congress deemed the film "culturally, historically, or aesthetically significant," and added it to the National Film Registry.

The Big Sleep

Theatrical release lobby card

Directed by	Howard Hawks
Produced by	Howard Hawks
Screenplay by	William Faulkner Leigh Brackett Jules Furthman
Based on	*The Big Sleep* by Raymond Chandler
Starring	Humphrey Bogart Lauren Bacall
Music by	Max Steiner
Cinematography	Sidney Hickox
Edited by	Christian Nyby
Distributed by	Warner Bros.
Release date	August 23, 1946 (United States)
Running time	114 minutes (released cut) 116 minutes (re-released original cut)

Relation name	Size	Example
/people/person/nationality	281,107	John Dugard, South Africa
/location/location/contains	253,223	Belgium, Nijlen
/people/person/profession	208,888	Dusa McDuff, Mathematician
/people/person/place_of_birth	105,799	Edwin Hubble, Marshfield
/dining/restaurant/cuisine	86,213	MacAyo's Mexican Kitchen, Mexican
/business/business_chain/location	66,529	Apple Inc., Apple Inc., South Park, NC
/biology/organism_classification_rank	42,806	Scorpaeniformes, Order
/film/film/genre	40,658	Where the Sidewalk Ends, Film noir
/film/film/language	31,103	Enter the Phoenix, Cantonese
/biology/organism_higher_classification	30,052	Calopteryx, Calopterygidae
/film/film/country	27,217	Turtle Diary, United States
/film/writer/film	23,856	Irving Shulman, Rebel Without a Cause
/film/director/film	23,539	Michael Mann, Collateral
/film/producer/film	22,079	Diane Eskenazi, Aladdin
/people/deceased_person/place_of_death	18,814	John W. Kern, Asheville
/music/artist/origin	18,619	The Octopus Project, Austin
/people/person/religion	17,582	Joseph Chartrand, Catholicism
/book/author/works_written	17,278	Paul Auster, Travels in the Scriptorium
/soccer/football_position/players	17,244	Midfielder, Chen Tao
/people/deceased_person/cause_of_death	16,709	Richard Daintree, Tuberculosis
/book/book/genre	16,431	Pony Soldiers, Science fiction
/film/film/music	14,070	Stavisky, Stephen Sondheim
/business/company/industry	13,805	ATS Medical, Health care

Table 2: The 23 largest Freebase relations we use, with their size and an instance of each relation.

Mintz et al. 2009

Distant supervision

mayor(Maynard Jackson, Atlanta)

Elected mayor of Atlanta in 1973, Maynard Jackson…
Atlanta's airport will be renamed to honor Maynard Jackson, the city's first Black mayor
Born in Dallas, Texas in 1938, Maynard Holbrook Jackson, Jr. moved to Atlanta when he was 8.

Eisenstein 2018

Distant supervision

- For feature-based models, we can represent the tuple <m1, m2> by aggregating together the representations from all the sentences they appear in

Distant supervision

[The Big Sleep]$_{m1}$ is a 1946 film noir directed by [Howard Hawks]$_{m2}$, the first film version of Raymond Chandler's 1939 novel of the same name.

[Howard Hawks]$_{m2}$ directed the [The Big Sleep]$_{m1}$

feature(m1, m2)	value (e.g., normalized over all sentences)
“directed” between m1, m2	0.37
“by” between m1, m2	0.42
m1←*nsubjpass* ← directed→*obl:agent* → m2	0.13
m2←*nsubj* ← directed→*obj* → m2	0.08

Distant supervision

- Discovering Hearst patterns from distant supervision using WordNet (Snow et al. 2005)

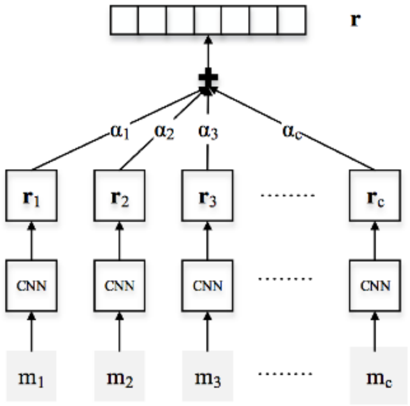
pattern	*sentence*
NP_H like NP	Many hormones like leptin...
NP_H called NP	a markup language called XHTML
NP is a NP_H	Ruby is a programming language...
NP, a NP_H	IBM, a company with a long...

Multiple Instance Learning

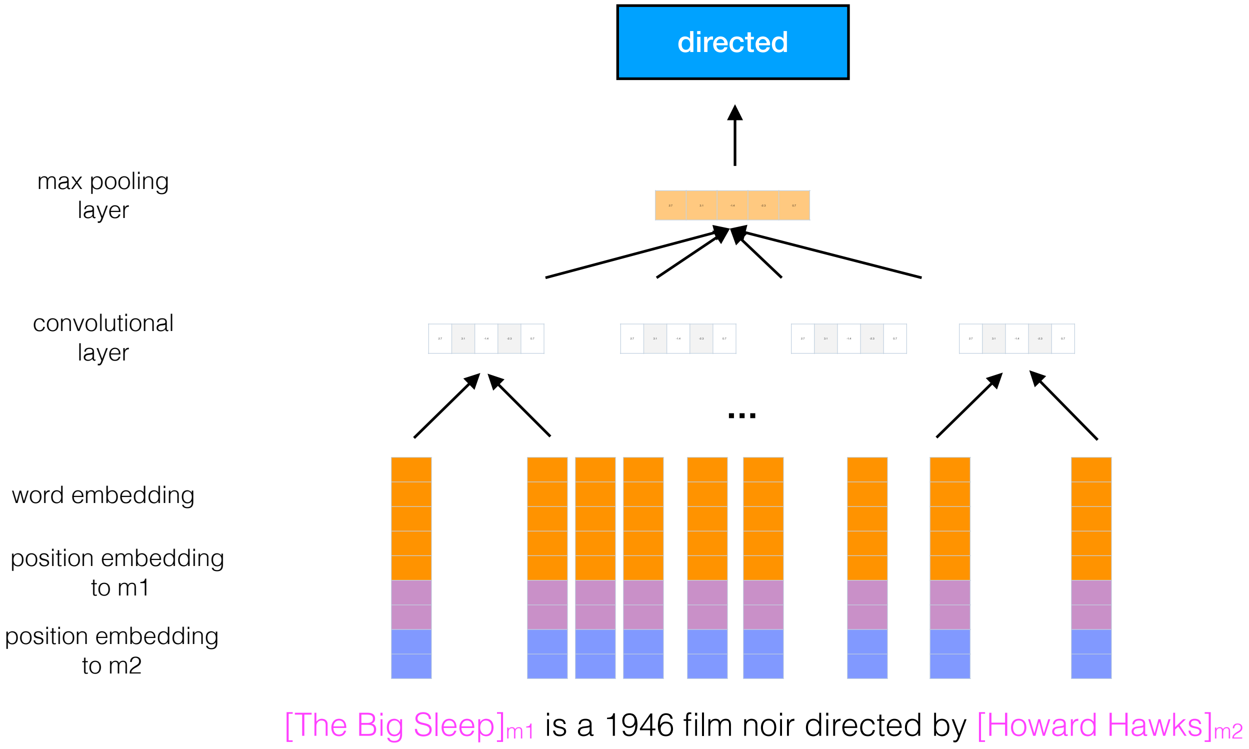
- Labels are assigned to a set of sentences, each containing the pair of entities m1 and m2; not all of those sentences express the relation between m1 and m2.

Attention

- Let's incorporate structure (and parameters) into a network that captures which sentences in the input we should be attending to (and which we can ignore).

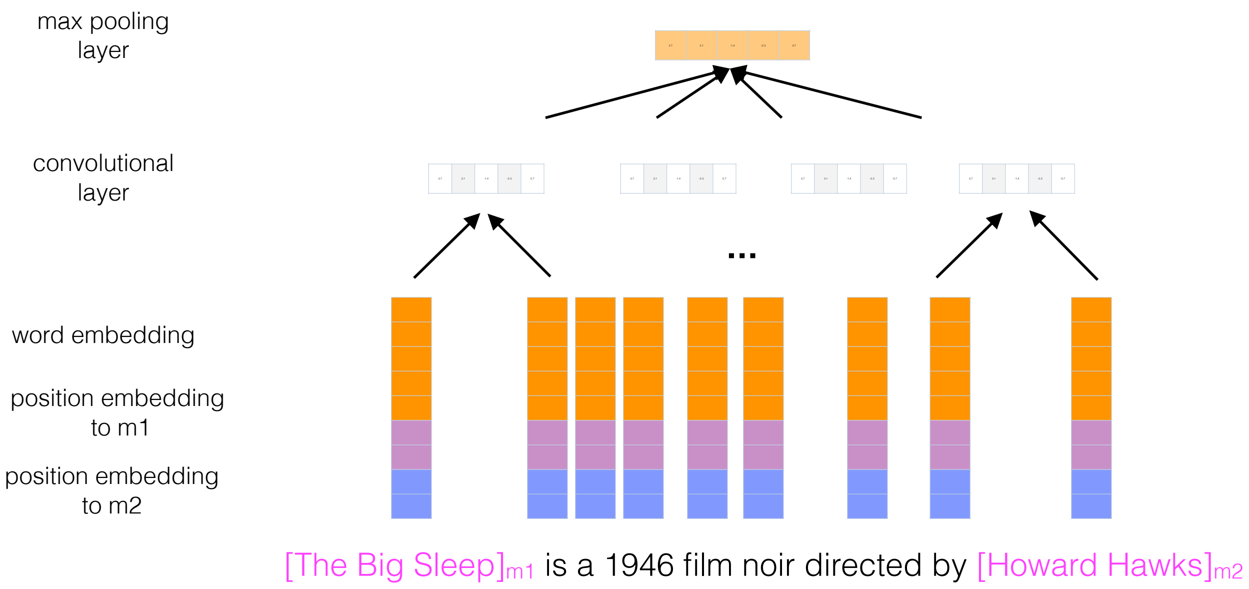


Lin et al (2016), "Neural Relation Extraction with Selective Attention over Instances" (ACL)

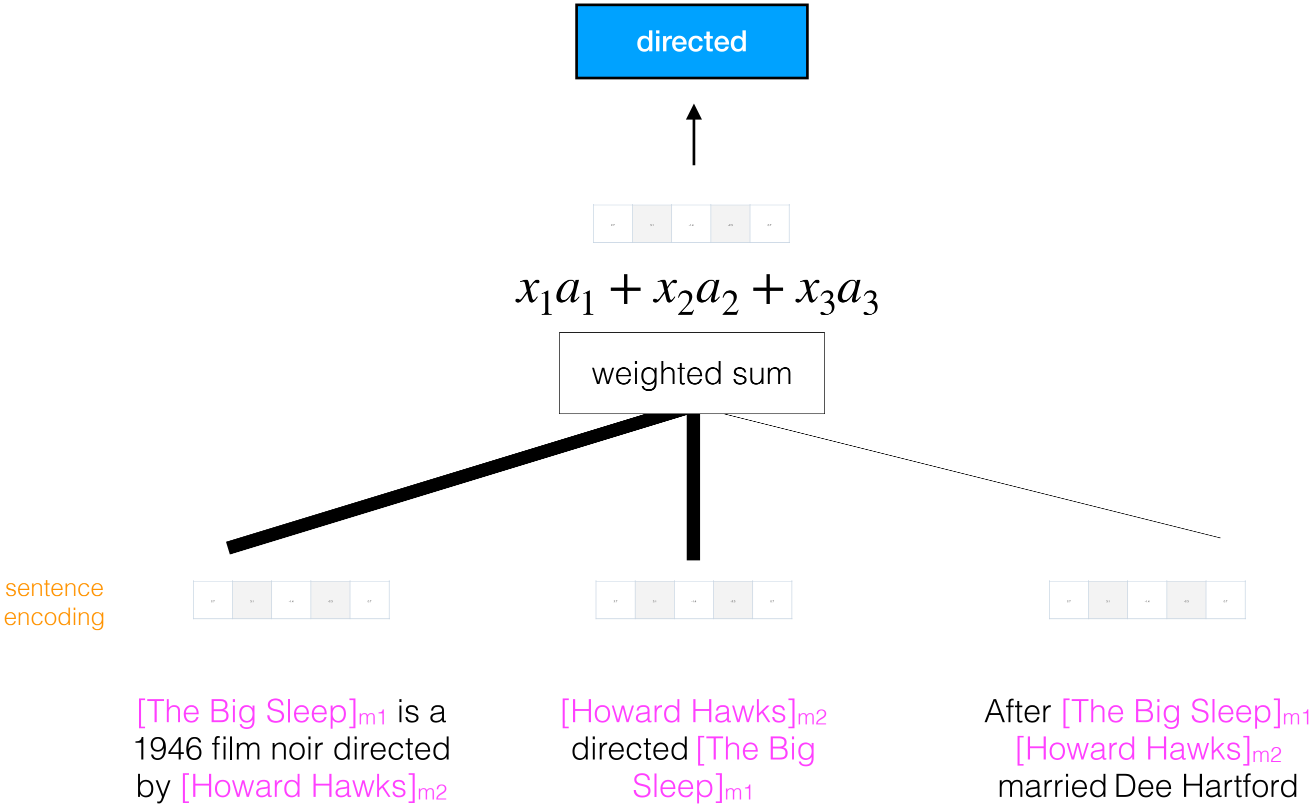


Lin et al (2016), “Neural Relation Extraction with Selective Attention over Instances” (ACL)

Now we just have an encoding of a sentence



Lin et al (2016), "Neural Relation Extraction with Selective Attention over Instances" (ACL)


directed
$x_1 a_1 + x_2 a_2 + x_3 a_3$
weighted sum
sentence encoding
[The Big Sleep]m1 is a 1946 film noir directed by [Howard Hawks]m2
[Howard Hawks]m2 directed [The Big Sleep]m1
After [The Big Sleep]m1 [Howard Hawks]m2 married Dee Hartford

Information Extraction

- Named entity recognition
- Entity linking
- Relation extraction
- Template filling
- Event detection
- Event coreference
- Extra-propositional information (veridicality, hedging)

Applied Natural Language Processing

Info 256

Lecture 25: Sequence alignment (Nov. 20, 2023)

David Bamman, UC Berkeley

“Five score years ago, a great American, in whose symbolic shadow we stand today, signed the Emancipation Proclamation.”

“Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.”

Sequence alignment

Wilkerson et al. 2014, "Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach"

ing mothers a in general sec-	ing mothers———- section
tion 7 of the fair labor stan-	7 of the fair labor standards act
dards act——— 29 usc 207 is	of 1938 29 usc 207 is amended
amended by adding at the	by adding at the end the fol-
end the following r 1 an em-	lowing r 1 an employer shall
ployer shall provide—- reason-	provide a a reasonable break
able break time for an em-	time for an employee to ex-
ployee to express breast milk	press breast milk for her nurs-
for her nursing child for 1 year	ing child for 1 year after the
after the childs birth each time	childs birth each time such
such employee has need to ex-	employee has need to express
press the milk the employer	the milk and ———————
shall make reasonable efforts	———————b a place
to provide a place other than	other than a bathroom that
a bathroom that is shielded	is shielded from view and free
from view and free from in-	from intrusion from coworkers
trusion from coworkers and	and the public which may be
the public which may be used	used by an employee to ex-
by an employee to express	press breast milk 2 an em-
breast milk– an employer shall	ployer shall not be required
not be required to compen-	to compensate an employee re-
sate an employee———————	ceiving reasonable break time
——————————— for	under paragraph 1 for any
any work time spent for such	work time spent for such pur-
purpose 2 for purposes of this	pose 3 ———————————
subsection the term employer	———an employer —-that em-
means an employ	ploy

Table 1: A Local Alignment Example

Sequence alignment

Category of Alignment	Proportion of Total	Example
Religious	11%	Father Which art in Heaven hallowed be thy Name
Lyric	10%	Amazing Grace, how sweet the sound, That saved a wretch
Self-Citation	4%	asked: "Did I snore?" "Terribly," he said, "you sounded like a chain saw
Juridical	4%	find this defendant guilty of murder in the first degree
Quotation	6%	Patrick Henry said 'Give me liberty or give me death'
Aphorism/Saying	2%	to make a long story short
Onomatopoeia	2%	Kitty-kitty-kitty, here kitty-kitty-kitty

So et al. 2019, "Race, Writing, and Computation: Racial Difference and the US Novel, 1880-2000," *Cultural Analytics*

Spelling correction

It was the best of times, it was the blurst of times

Levenshtein distance

- For a pair of strings, the minimal number of insert, delete and substitution operations required to transform one into the other.

- Each operation has a cost:

    - insert: 1
    - delete: 1
    - substitution: 2

pints

\+ +

pints

Levenshtein distance: 2

Levenshtein distance

- For a pair of strings, the minimal number of insert, delete and substitution operations required to transform one into the other.

- Each operation has a cost:

    - insert: 1
    - delete: 1
    - substitution: 2

~~p~~in

-

in

Levenshtein distance: 1

Levenshtein distance

- For a pair of strings, the minimal number of insert, delete and substitution operations required to transform one into the other.

- Each operation has a cost:
    - insert: 1
    - delete: 1
    - substitution: 2

pin

pan

Levenshtein distance: 2

Levenshtein distance

- For a pair of strings, the minimal number of insert, delete and substitution operations required to transform one into the other.
- Each operation has a cost:
    - insert: 1
    - delete: 1
    - substitution: 2

Shaxper

Shakespeare

scost = 0 if t[i] = s[i], otherwise 2

$$d_{i,j} = \min($$

insert $\quad d_{i-1,j} + 1$

delete $\quad d_{i,j-1} + 1$

substitute $\quad d_{i-1,j-1} + \text{scost}$

$)$

		S	h	a	x	p	e	r
	0	1	2	3	5	6	7	8
S	1							
h	2							
a	3							
k	4							
e	5							
s	6							
p	7							
e	8							
a	9							
r	10							
e	11							

scost = 0 if t[i] = s[i], otherwise 2

$$d_{i,j} = \min($$

insert $d_{i-1,j} + 1$

delete $d_{i,j-1} + 1$

substitute $d_{i-1,j-1} + \text{scost}$

$)$

		S	h	a	x	p	e	r
	0	1	2	3	5	6	7	8
S	1	0						
h	2		0					
a	3			0				
k	4				2			
e	5							
s	6							
p	7							
e	8							
a	9							
r	10							
e	11							

scost = 0 if t[i] = s[i], otherwise 2

$$d_{i,j} = \min($$

insert $d_{i-1,j} + 1$

delete $d_{i,j-1} + 1$

substitute $d_{i-1,j-1} + \text{scost}$

$)$

		S	h	a	x	p	e	r
	0	1	2	3	5	6	7	8
S	1	0						
h	2		0					
a	3			0				
k	4				2			
e	5							
s	6							
p	7							
e	8							
a	9							
r	10							
e	11							

scost = 0 if t[i] = s[i], otherwise 2

$$d_{i,j} = \min($$

insert $\quad d_{i-1,j} + 1$

delete $\quad d_{i,j-1} + 1$

substitute $\quad d_{i-1,j-1} + \text{scost}$

$)$

		S	h	a	x	p	e	r
	0	1	2	3	5	6	7	8
S	1	0						
h	2		0					
a	3			0				
k	4				2	3	4	5
e	5							
s	6							
p	7							
e	8							
a	9							
r	10							
e	11							

scost = 0 if t[i] = s[i], otherwise 2

$$d_{i,j} = \min($$

insert $d_{i-1,j} + 1$

delete $d_{i,j-1} + 1$

substitute $d_{i-1,j-1} + \text{scost}$

$)$

		S	h	a	x	p	e	r
	0	1	2	3	5	6	7	8
S	1	0						
h	2		0					
a	3			0				
k	4				2	3	4	5
e	5							
s	6							
p	7							
e	8							
a	9							
r	10							
e	11							

scost = 0 if t[i] = s[i], otherwise 2

$$d_{i,j} = \min($$

insert $d_{i-1,j} + 1$

delete $d_{i,j-1} + 1$

substitute $d_{i-1,j-1} + \text{scost}$

$)$

		S	h	a	x	p	e	r
	0	1	2	3	5	6	7	8
S	1	0						
h	2		0					
a	3			0				
k	4				2	3	4	5
e	5							
s	6							
p	7							
e	8							
a	9							
r	10							
e	11							

scost = 0 if t[i] = s[i], otherwise 2

$$d_{i,j} = \min($$

insert $d_{i-1,j} + 1$

delete $d_{i,j-1} + 1$

substitute $d_{i-1,j-1} + \text{scost}$

$)$

		S	h	a	x	p	e	r
	0	1	2	3	5	6	7	8
S	1	0	1	2	3	4	5	6
h	2	1	0	1	2	3	4	5
a	3							
k	4							
e	5							
s	6							
p	7							
e	8							
a	9							
r	10							
e	11							

scost = 0 if t[i] = s[i], otherwise 2

$$d_{i,j} = \min($$

insert $d_{i-1,j} + 1$

delete $d_{i,j-1} + 1$

substitute $d_{i-1,j-1} + \text{scost}$

)

		S	h	a	x	p	e	r
	0	1	2	3	5	6	7	8
S	1	0	1	2	3	4	5	6
h	2	1	0	1	2	3	4	5
a	3	2	1	0	1	2	3	4
k	4	3	2	1	2	3	4	5
e	5	4	3	2	3	4	5	6
s	6	5	4	3	4	5	6	**7**
p	7	6	5	4	5	4	5	6
e	8	7	6	5	6	5	4	5
a	9	8	7	6	7	6	5	6
r	10	9	8	7	8	7	6	5
e	11	10	9	8	9	8	7	6

Identical S (do nothing)

Identical h (do nothing)

Identical a (do nothing)

Insert k

Insert e

Replace x with s

Identical p (do nothing)

Identical e (do nothing)

Insert a

Identical r (do nothing)

Insert e

		S	h	a	x	p	e	r
	0	1	2	3	5	6	7	8
S	1	0	1	2	3	4	5	6
h	2	1	0	1	2	3	4	5
a	3	2	1	0	1	2	3	4
k	4	3	2	1	2	3	4	5
e	5	4	3	2	3	4	5	6
s	6	5	4	3	4	5	6	7
p	7	6	5	4	5	4	5	6
e	8	7	6	5	6	5	4	5
a	9	8	7	6	7	6	5	6
r	10	9	8	7	8	7	6	5
e	11	10	9	8	9	8	7	6

Sequence alignment

- Levenstein gives us the minimal number of operations required to transform one string into another.
- But what if we want to find the best alignment between a pair of strings?

I should'a quit you, a long time ago
I should'a quit you, baby, long time ago
I should'a quit you, and went on to Mexico

If I ha'da followed my first mind
If I ha'da followed my first mind
I'd'a been gone since my second time

Killing Floor, Howlin' Wolf

I should have quit you a long time ago
Ooh-whoa, yeah, yeah, long time ago
I wouldn't be here, my children
Down on this killin' floor
I should have listened, baby, a-to my second mind
Oh, I should have listened, baby, to my second mind

Lemon Song, Led Zeppelin

I
should
a
quit
you
a
long
time
ago
...
If
I
had
a
followed
my
first
mind
I'd
a
been
gone
since
my
second
time

I
should
have
quit
you
a
long
time
ago
...
I
should
have
listened
baby
a-to
my
second
mind

sscore = 1 if t[i] = s[i], otherwise -1
d (gap penalty) = -1

$$d_{i,j} = \max($$

insert $d_{i-1,j} + d$

delete $d_{i,j-1} + d$

substitute $d_{i-1,j-1} + \text{sscore}$

$)$

		It	was	the	blurst	of	times
	0	-1	-2	-3	-4	-5	-6
It	-1	1	0	-1	-2	-3	-4
was	-2	0	2	1	0	-1	-2
the	-3	-1	1	3	2	1	0
worst	-4	-2	0	2	2	1	0
of	-5	-3	-1	1	1	3	2
times	-6	-4	-2	0	0	2	4

sscore = 1 if t[i] = s[i], otherwise -1
d (gap penalty) = -1

$$d_{i,j} = \max($$

insert $d_{i-1,j} + d$

delete $d_{i,j-1} + d$

substitute $d_{i-1,j-1} + \text{sscore}$

$)$

		It	was	the	blurst	of	times
	0	-1	-2	-3	-4	-5	-6
It	-1	1	0	-1	-2	-3	-4
was	-2	0	2	1	0	-1	-2
the	-3	-1	1	3	2	1	0
worst	-4	-2	0	2	2	1	0
of	-5	-3	-1	1	1	3	2
times	-6	-4	-2	0	0	2	4

Needleman-Wunsch

- We can think about this as a generalization of Levenshtein distance, with the ability to specify costs for individual matches/mismatches (Levenshtein “substitutions”)

	A	G	T	C
A	1	-1	-2	-3
G	-1	1	-2	-3
T	-2	-2	1	-3
C	-3	-3	-3	1

	a	an	dog	times
a	1	0	-1	-1
an	0	1	-1	-1
dog	-1	-1	1	-1
times	-1	-1	-1	1

Needleman-Wunsch

- Needleman-Wunsch is a global alignment algorithm, finding the optimal alignment for the entirety of string a and the entirety of string b.

- Many applications in text as data involve finding smaller regions of similarity within two strings.

As I walk through the valley of the shadow of death
I take a look at my life and realize there's nothin' left
'Cause I've been blastin' and laughin' so long that
Even my momma thinks that my mind is gone

But I ain't never crossed a man that didn't deserve it
Me be treated like a punk, you know that's unheard of
You better watch how you talkin' and where you walkin'
Or you and your homies might be lined in chalk

I really hate to trip, but I gotta loc
As they croak, I see myself in the pistol smoke
Fool, I'm the kinda G the little homies wanna be like
On my knees in the night, sayin' prayers in the streetlight

Been spendin' most their lives
Livin' in a gangsta's paradise
Been spendin' most their lives
Livin' in a gangsta's paradise
Keep spendin' most our lives
Livin' in a gangsta's paradise
Keep spendin' most our lives
Livin' in a gangsta's paradise

Look at the situation they got me facing
I can't live a normal life, I was raised by the street
So I gotta be down with the hood team
Too much television watchin', got me chasing dreams

Been spending most their lives
Living in a pastime paradise
They've been spending most their lives
Living in a pastime paradise

They've been wasting most their time
Glorifying days long gone behind
They've been wasting most their days
In rememberance of ignorance oldest praise

Tell me who of them will come to be
How many of them are you and me

Dissipation
Race relations
Consolation
Segregation
Dispensation
Isolation
Exploitation
Mutilation
Mutations
Miscreation
Confirmation to the evils of the world

Been spending most their lives
Living in a future paradise
They've been spending most their lives
Living in a future paradise
They've been looking in their minds
For the days that sorrow's gone from time
They keep telling of the day
When the Savior of love will come to stay

Smith-Waterman

- Smith-Waterman alignment addresses this by focusing on local alignment.
- Two main differences from Needleman-Wunsch:
  - No negative scores. This enables two regions to be similar even if they are preceded by long regions that are not the same (which would otherwise have a strong negative score under NW).
  - Traceback starts with the highest score in matrix, not the bottom/rightmost corner, enabling two similar regions to be found *anywhere* in the strings.

sscore = 1 if t[i] = s[i], otherwise -1
d (gap penalty) = -1

$$d_{i,j} = \max($$

insert $d_{i-1,j} + d$

delete $d_{i,j-1} + d$

substitute $d_{i-1,j-1} + \text{sscore}$

0

$)$

		It	was	the	blurst	of	times
	0	0	0	0	0	0	0
It	0	1	0	0	0	0	0
was	0	0	2	1	0	0	0
the	0	0	1	3	2	1	0
worst	0	0	0	2	2	1	0
of	0	0	0	1	1	3	2
times	0	0	0	0	0	2	4

Traceback:

- Find the highest score in the matrix
- Follow the source of each decision to find the best alignment
- End when a 0 score is encountered.

		It	was	the	blurst	of	times
	0	0	0	0	0	0	0
It	0	1	0	0	0	0	0
was	0	0	2	1	0	0	0
the	0	0	1	3	2	1	0
worst	0	0	0	2	2	1	0
of	0	0	0	1	1	3	2
times	0	0	0	0	0	2	4

		It	was	the	blurst	of	times
	0	0	0	0	0	0	0
It	0	1	0	0	0	0	0
was	0	0	2	1	0	0	0
the	0	0	1	3	2	1	0
worst	0	0	0	2	2	1	0
of	0	0	0	1	1	3	2
times	0	0	0	0	0	2	4
it	0	1	0	0	0	1	3
was	0	0	2	1	0	0	2
the	0	0	1	3	2	1	1
age	0	0	0	2	1	0	0

		It	was	the	blurst	of	times
	0	0	0	0	0	0	0
It	0	1	0	0	0	0	0
was	0	0	2	1	0	0	0
the	0	0	1	3	2	1	0
worst	0	0	0	2	2	1	0
of	0	0	0	1	1	3	2
times	0	0	0	0	0	2	4
it	0	1	0	0	0	1	3
was	0	0	2	1	0	0	2
the	0	0	1	3	2	1	1
age	0	0	0	2	1	0	0

Traceback:

- Find the highest score in the matrix
- Follow the source of each decision to find the best alignment
- End when a 0 score is encountered.

		It	was	the	blurst	of	times
	0	0	0	0	0	0	0
It	0	1	0	0	0	0	0
was	0	0	2	1	0	0	0
the	0	0	1	3	2	1	0
worst	0	0	0	2	2	1	0
of	0	0	0	1	1	3	2
times	0	0	0	0	0	2	4
it	0	1	0	0	0	1	3
was	0	0	2	1	0	0	2
the	0	0	1	3	2	1	1
age	0	0	0	2	1	0	0

Activity

`16.sequence_alignment/Smith-Waterman Alignment`

- Run Smith-Waterman alignment on song lyrics to identify instances of text reuse within them.