

# CERES: DISTANTLY SUPERVISED RELATION EXTRACTION FROM THE SEMI-STRUCTURED WEB

Colin Lockard, Xin Luna Dong, Arash Einolghozati, Prashant Shiralkar



Find Movies, TV shows, Celebrities and more...

All

Movies, TV & Showtimes

Cebs, Events & Photos

News & Community

Watchlist

Enjoy unlimited streaming on Prime Video  
Start your 30-day free trial. Plans starting at \$8.99/mo



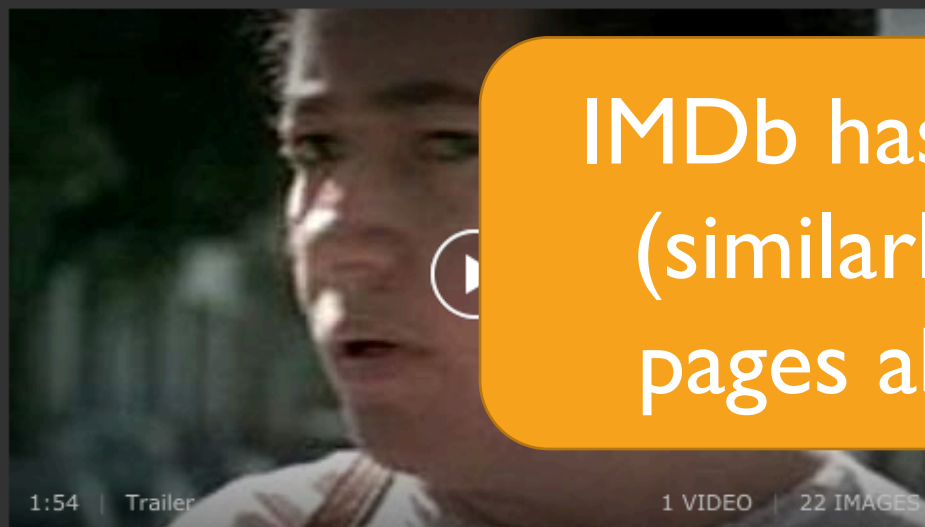
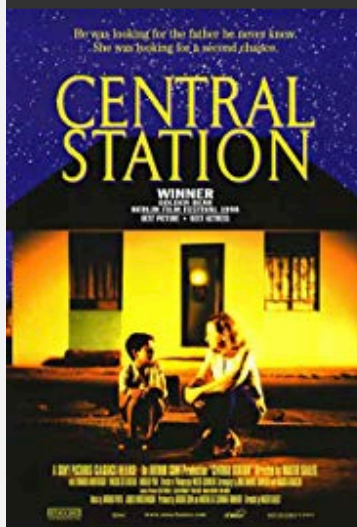
FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE

# + Central Station (1998)

★ 8.0/10  
31,566

★ Rate This

Central do Brasil (*original title*)  
R | 1h 53min | Drama | 20 November 1998 (USA)



IMDb has thousands of (similarly formatted) pages about movies.



On Disc  
at Amazon

An emotive journey of a former school teacher, who writes letters for illiterate people, and a young boy, whose mother has just died, as they search for the father he never knew.

**Director:** [Walter Salles](#)

**Writers:** [Marcos Bernstein](#), [João Emanuel Carneiro](#) | [1 more credit](#) »

**Stars:** [Fernanda Montenegro](#), [Vinícius de Oliveira](#), [Marília Pêra](#) | [See full cast & crew](#) »

# Twisted (Short film)

Nigerian films

D" Trailer



Year of production: 2014

Running Time: 2:12 mins

Written by: Daniel Ademinokan

Produced by: Daniel Ademinokan

Directed by: Daniel Ademinokan

Starring: [Stella Damasus](#) Rob Byrnes, Matt Meinsen and David Ademinokan

Not on IMDb!

Bollywood films

# Made In China (2019)

Not on IMDb!

**Banner:** Maddock Films  
**Director:** [Mihil Musale](#)  
**Producer:** [Dinesh Vijan](#)  
**Star:** [Rajkummar Rao](#) , [Mouni Roy](#), [Boman Irani](#) ...see full cast & crew





# Abegweit

Serge Morin  
1998 | 1 h 11 min

AVAILABLE

Canadian films

Not on IMDb!

- SEARCH
- FILMS
- DOCUMENTARY
- ANIMATION
- INTERACTIVE
- EDUCATION
- SIGN IN
- f
- 🐦
- v
- YouTube

### ▼ CREDITS

<b>DIRECTOR</b> Serge Morin	<b>SCRIPT</b> Serge Morin	<b>PRODUCER</b> Pierre Bernier Diane Poitras	<b>CAMERA</b> Marc Paulin
<b>SOUND</b> Georges Hannan	<b>EDITING</b> Fernand Bélanger	<b>RE-RECORDING</b> Serge Boivin Jean Paul Vialard	<b>SOUND EDITING</b> Fernand Bélanger Claude Langlois
<b>NARRATOR</b> Alex Madsen	<b>MUSIC</b> Richard Gibson	<b>PARTICIPATION</b> Francine Blais Peter Briden Ralph Murray Guy Cormier Jim Feltham Kim Gallant	

# SEMI-STRUCTURED WEB EXTRACTION

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE ▾ | SHARE

+ **Do the Right Thing** (1989) ★ 7.9/10 | ☆ Rate This

R | 2h | Comedy, Drama | 21 July 1989 (USA)

Watch Now  
From \$2.99 (SD) on Prime Video

ON DISC

On the hottest day of the year on a street in the Bedford-Stuyvesant section of Brooklyn, everyone's hate and bigotry smolders and boils until it explodes into violence.

Director: Spike Lee  
Writer: Spike Lee  
Stars: Danny Aiello, Ossie Davis, Ruby Dee

## TOPIC ENTITY

“Do the Right Thing”

“Do the Right Thing”, film.release, “21 July 1989”

“Do the Right Thing”, film.genre, “Comedy”

“Do the Right Thing”, film.genre, “Drama”

“Do the Right Thing”, film.runtime, “2h”

“Do the Right Thing”, film.rating, “R”

“Do the Right Thing”, film.director, “Spike Lee”

“Do the Right Thing”, film.writer, “Spike Lee”

“Do the Right Thing”, film.actor, “Danny Aiello”

“Do the Right Thing”, film.actor, “Ossie Davis”

“Do the Right Thing”, film.actor, “Ruby Dee”

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE  SHARE

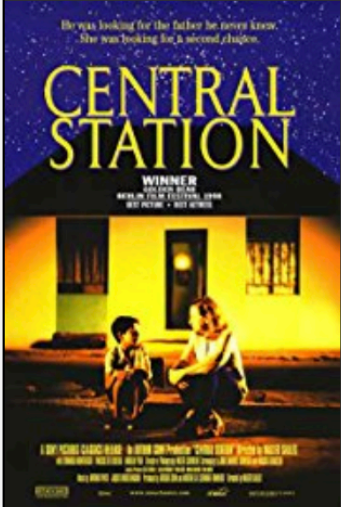
# + Central Station (1998)

★ 8.0 /10  
31,520

☆ Rate This

Central do Brasil (original title)

R | 1h 53min | Drama | 20 November 1998 (USA)



1:54 | Trailer

1 VIDEO | 22 IMAGES



On Disc  
at Amazon

An emotive journey of a former school teacher, who writes letters for illiterate people, and a young boy, whose mother has just died, as they search for the father he never knew.

**Director:** [Walter Salles](#)

**Writers:** [Marcos Bernstein](#), [João Emanuel Carneiro](#) | [1 more credit](#) »

**Stars:** [Fernanda Montenegro](#), [Vinícius de Oliveira](#), [Marília Pêra](#) | [See full cast & crew](#) »

80

Metascore  
From [metacritic.com](#)

Reviews  
261 user | 73 critic

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE  SHARE

# + Do the Right Thing (1989)

★ 7.9 /10  
69,694

☆ Rate This

R | 2h | Comedy, Drama | 21 July 1989 (USA)



2:09 | Trailer

3 VIDEOS | 49 IMAGES



Watch Now

From \$2.99 (SD) on Prime Video



ON DISC

On the hottest day of the year on a street in the Bedford-Stuyvesant section of Brooklyn, everyone's hate and bigotry smolders and builds until it explodes into violence.

**Director:** [Spike Lee](#)

**Writer:** [Spike Lee](#)

**Stars:** [Danny Aiello](#), [Ossie Davis](#), [Ruby Dee](#) | [See full cast & crew](#) »

91

Metascore  
From [metacritic.com](#)

Reviews  
436 user | 110 critic

Popularity  
2,391 (▲ 446)



FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE ▾ | SHARE

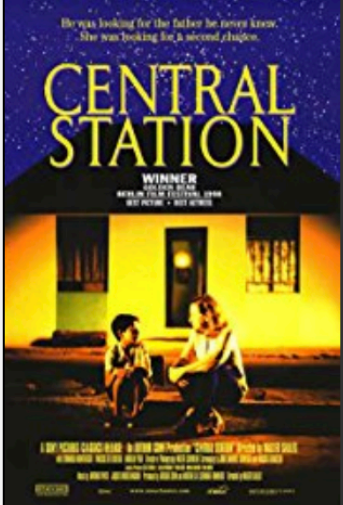
# + Central Station (1998)

Central do Brasil (original title)

R | 1h 53min | Drama | 20 November 1998 (USA)

★ 8.0<sub>31,520</sub>

☆ Rate This



1:54 | Trailer

1 VIDEO | 22 IMAGES



On Disc  
at Amazon

An emotive journey of a former school teacher, who writes letters for illiterate people, and a young boy, whose mother has just died, as they search for the father he never knew.

**Director:** [Walter Salles](#)

**Writers:** [Marcos Bernstein](#), [João Emanuel Carneiro](#) | [1 more credit](#) »

**Stars:** [Fernanda Montenegro](#), [Vinícius de Oliveira](#), [Marília Pêra](#) | [See full cast & crew](#) »

80

Metascore  
From [metacritic.com](#)

Reviews  
261 user | 73 critic

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE ▾ | SHARE

# + Star Wars: The Last Jedi (2017)

Star Wars: Episode VIII - The Last Jedi (original title)

PG-13 | 2h 32min | [Action](#), [Adventure](#), [Fantasy](#) | [15 December 2017 \(USA\)](#)

★ 7.3<sub>404,499</sub>

☆ Rate This



Rey develops her newly discovered abilities with the guidance of Luke Skywalker, who is unsettled by the strength of her powers. Meanwhile, the Resistance prepares for battle with the First Order.

**Director:** [Rian Johnson](#)

**Writers:** [Rian Johnson](#), [George Lucas](#) (based on characters created by)

**Stars:** [Daisy Ridley](#), [John Boyega](#), [Mark Hamill](#) | [See full cast & crew](#) »

85

Metascore  
From [metacritic.com](#)

Reviews  
5,463 user | 645 critic



Popularity  
84 (↓ 3)

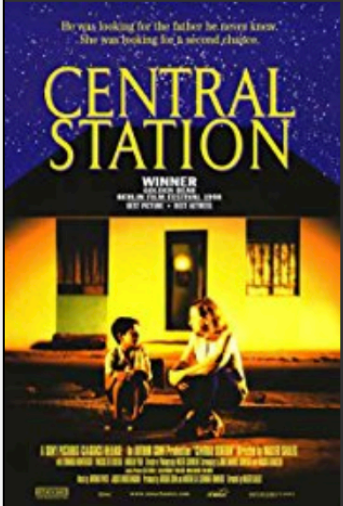
[prime](#) Watch Now



Page formats can differ slightly

# + Central Station (1998)

Central do Brasil (original title)  
R | 1h 53min | Drama | 20 November 1998 ( )



1:54 | Trailer | 1 VIDEO | 22 IMAGES

On Disc at Amazon

An emotive journey of a former school teacher, who writes letters for illiterate people, and a young boy, whose mother has just died, as they search for the father he never knew.

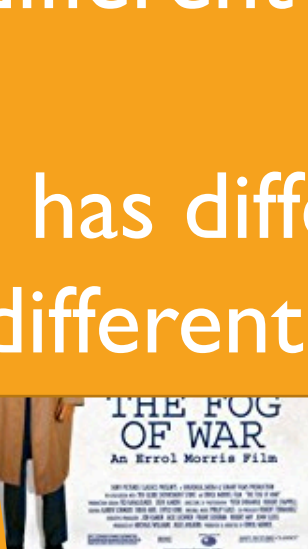
**Director:** [Walter Salles](#)  
**Writers:** [Marcos Bernstein](#), [João Emanuel Carneiro](#) | [1 more credit](#) »  
**Stars:** [Fernanda Montenegro](#), [Vinícius de Oliveira](#), [Marília Pêra](#) | [See full cast & crew](#) »

**80** Metascore | Reviews  
From [metacritic.com](#) | [261 user](#) | [73 critic](#)

Same predicate has different location on different pages.  
Same location has different predicates on different pages

# The Fog of War

An Errol Morris Film



2:09 | Trailer | 2 VIDEOS | 11 IMAGES

Watch Now  
From \$2.99 (SD) on Prime Video

ON DISC

The story of America as seen through the eyes of the former Secretary of Defense under President John F. Kennedy and President Lyndon Johnson, Robert McNamara.

**Director:** [Errol Morris](#)  
**Stars:** [Robert McNamara](#), [John F. Kennedy](#), [Fidel Castro](#) | [See full cast & crew](#) »

**87** Metascore | Reviews  
From [metacritic.com](#) | [155 user](#) | [141 critic](#)

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE ▾ | SHARE

# + Central Station (1998)

Central do Brasil (original title)  
R | 1h 53min | Drama | 20 November 1998 (USA)

★ 8.0 /10  
31,520

☆ Rate This

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE ▾ | SHARE

# + The Fog of War: Eleven Lessons from the Life of Robert S. McNamara (2003)

PG-13 | 1h 47min | Documentary, Biography, History | 5 March 2004 (USA)

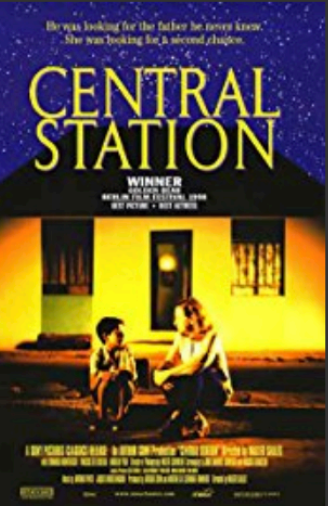
★ 8.2 /10  
20,953

☆ Rate This

Traditional approach: Wrapper Induction

Learn rules based on manually annotated pages.

Labor intensive: Need annotations for every site



 On Disc at Amazon

An emotive journey of a for young boy, whose mother h

**Director:** [Walter Salles](#)

**Writers:** [Marcos Bernstein,](#)

**Stars:** [Fernanda Montenegro,](#)

 **Metascore** | Reviews  
From [metacritic.com](#) | 261 user | 73 critic

2 VIDEOS | 11 IMAGES



ON DISC

ry of Defense under  
Namara.

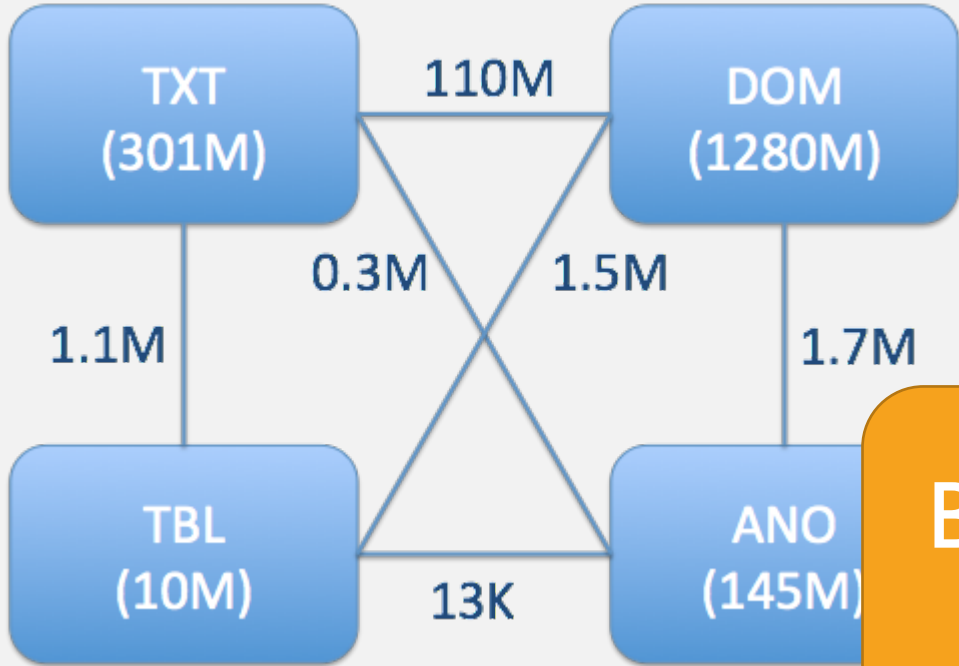
cast & crew »

 **Metascore** | Reviews  
From [metacritic.com](#) | 155 user | 141 critic

# BIG PROMISE FROM SEMI-STRUCTURED DATA

❑ Knowledge Vault @ Google showed big potential from distantly supervised DOM-tree extraction [Dong et al., KDD'14][Dong et al., VLDB'14]

Accu	Accu (conf $\geq .7$ )
0.36	0.52



Accu	Accu (conf $\geq .7$ )
0.43	0.63
0.09	0.62

Can't find new entities

But accuracy is still low

Can we automatically and accurately  
extract from semi-structured pages?



## PROBLEM DEFINITION

- Input:
  - Pages from a semi-structured website
  - Seed KB (and ontology)
- Output:
  - Newly extracted triples corresponding to the given ontology
  - Subject and object are strings
    - We do not address entity linkage or knowledge fusion

# DISTANT SUPERVISION

film.release\_year

Central  
Station

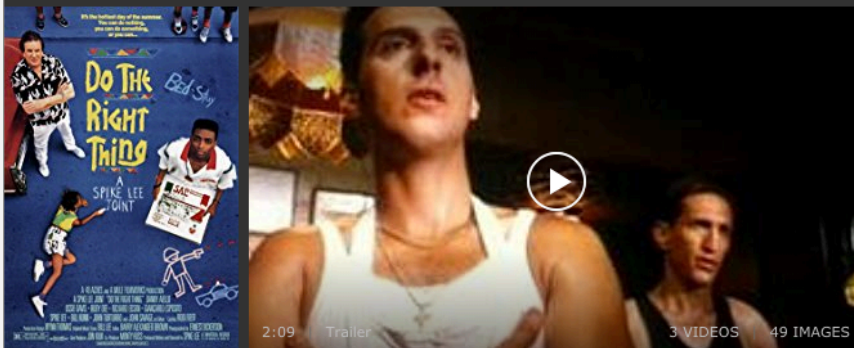
1998



- Assume joint mentions of subjects and objects indicate relation
- Use KB to automatically annotate data
- Use this as training data to learn classifier

# + Do the Right Thing (1989) ★ 7.9/10 70,044 ☆ Rate This

R | 2h | Comedy, Drama | 21 July 1989 (USA)



Watch Now  
From \$2.99 (SD) on Prime Video

ON DISC

On the hottest day of the year on a street in the Bedford-Stuyvesant section of Brooklyn, everyone's hate and bigotry smolders and builds until it explodes into violence.

**Director:** Spike Lee

**Writer:** Spike Lee

**Stars:** Danny Aiello, Ossie Davis, Ruby Dee | [See full cast & crew >](#)

**91** Metascore  
From [metacritic.com](#) | **435** user | **110** critic | Popularity  
1,038 (+141)

## Director's Trademarks: Your Guide to Spike Lee Joins



Get a closer look at the cinematic stylings of *BlacKkKlansman* director Spike Lee to discover what *Do the Right Thing*, *Malcolm X*, and *Inside Man* have in common.

Free unlimited data from Sprint in 2018 with AutoPay on select devices  
  
[Learn more](#)  
Pen sold separately. Roll over for offer details.  
[ad feedback](#)

## IMDb Picks: August



Are you ready for *The Happytime Murders*? We are! See which other movies and TV shows we're excited about.

[Visit IMDb Picks >](#)

## Related News

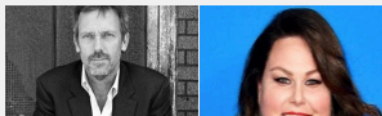
['BlacKkKlansman' Is Spike Lee's Smartest, Sharpest and Most Incendiary Film Since 'Do the Right Thing'](#)  
15 August 2018 | [TVovermind.com](#)

[The Academy's "Separate But Equal" Award for "Popular" Films Is an Insult to Movie Lovers Everywhere](#)  
15 August 2018 | [TVovermind.com](#)

[Posterized: Spike Lee](#)  
12 August 2018 | [FilmExperience](#)

[See all related articles >](#)

## Around The Web Powered by ZergNet



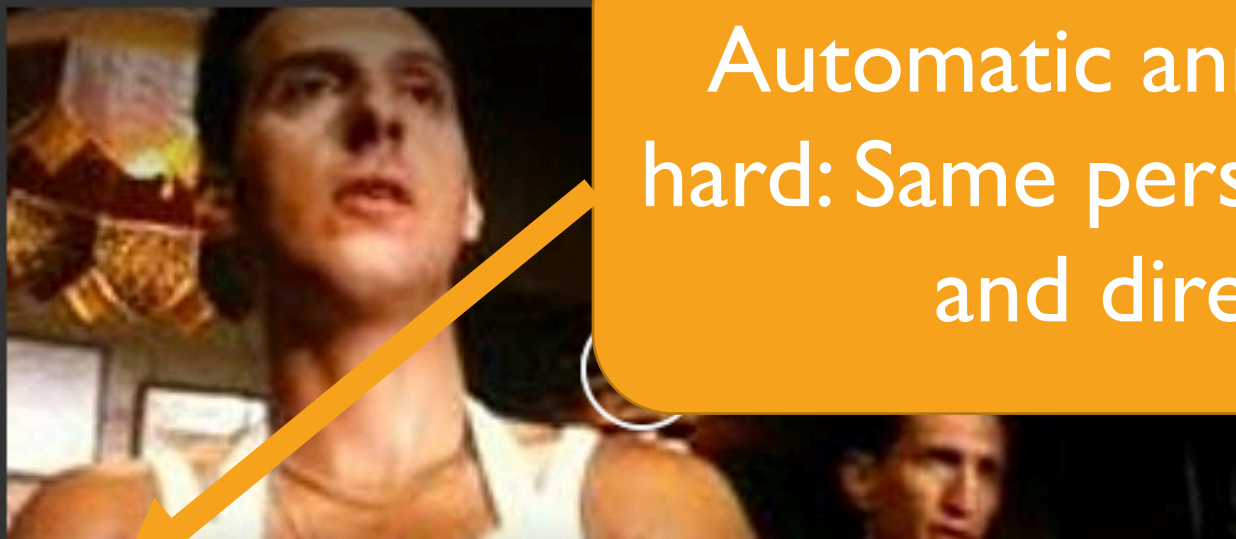
Problem:  
Webpages may mention thousands of entities, creating millions of pairs:  
1. Computationally complex.  
2. Spurious matches very likely

# + Do the Right Thing (1989)

R | 2h | Comedy, Drama | 21 July 1989 (USA)

★ 7.9/10  
70,044

☆ Rate This



Automatic annotation is hard: Same person is writer and director

**Director:** [Spike Lee](#)

**Writer:** [Spike Lee](#)

**Stars:** [Danny Aiello](#), [Ossie Davis](#), [Ruby Dee](#) | [See full cast & crew](#) »

On the hottest day of the year on a street in the Bedford-Stuyvesant section of Brooklyn, everyone's hate and bigotry smolders and builds until it explodes into violence.

**Director:** [Spike Lee](#)

**Writer:** [Spike Lee](#)

**Stars:** [Danny Aiello](#), [Ossie Davis](#), [Ruby Dee](#) | [See full cast & crew](#) »





# Do the Right Thing (1989)

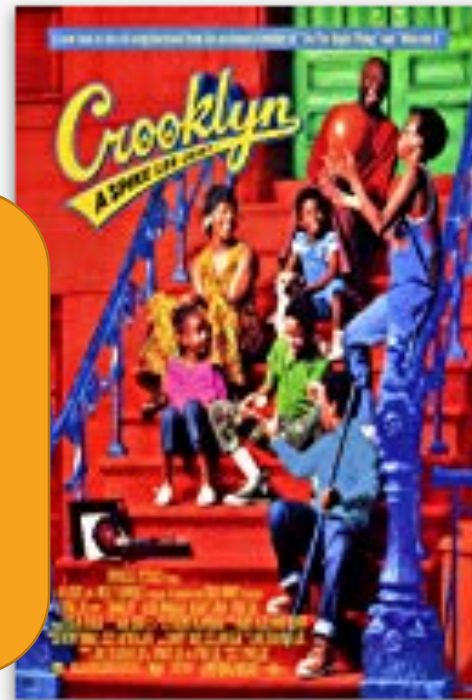
R | 2h | Comedy, Drama | 21 July 1989 (USA)

★ 7.9 /10  
70,044

☆ Rate This

## More Like This

[Learn more](#)



### Crooklyn (1994)

**PG-13** Comedy | Drama

★★★★★☆☆☆☆ 6.9/10

Spike Lee's vibrant semi-autobiographical portrait of a school teacher, her stubborn jazz musician husband and their five kids living in Brooklyn in 1973.

Add to Watchlist

Next »

**Director:** Spike Lee  
**Stars:** Alfre Woodard, Delroy Lindo, ...

Same person also mentioned in recommendation

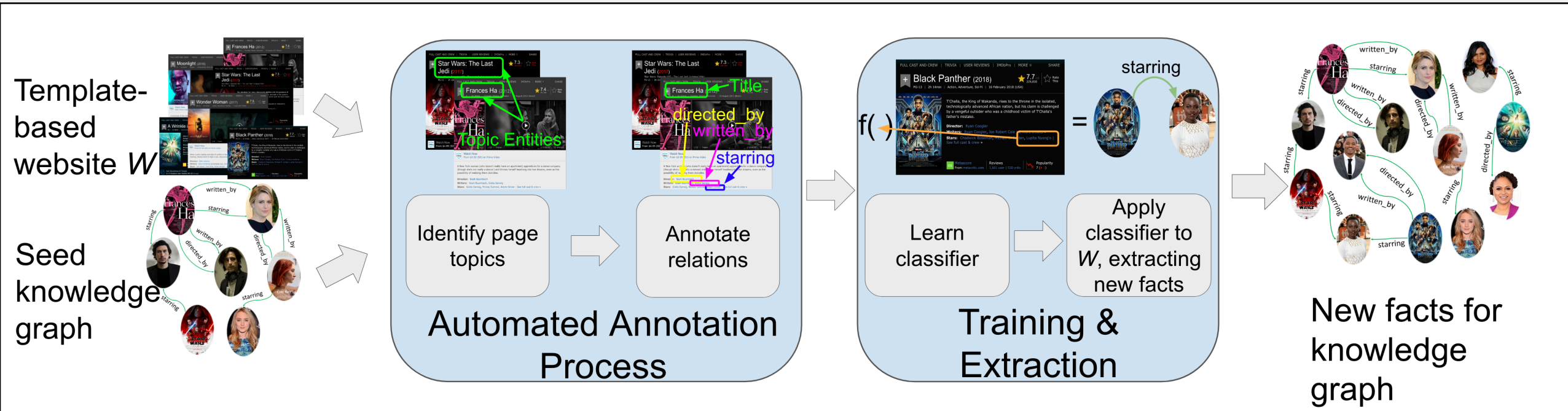
◀ Prev 6 Next 6 ▶

On the ho  
everyone

Director  
Writer:

Stars: Danny Aiello, Ossie Davis, Ruby Dee | [See full cast & crew](#) »

# CERES OVERVIEW

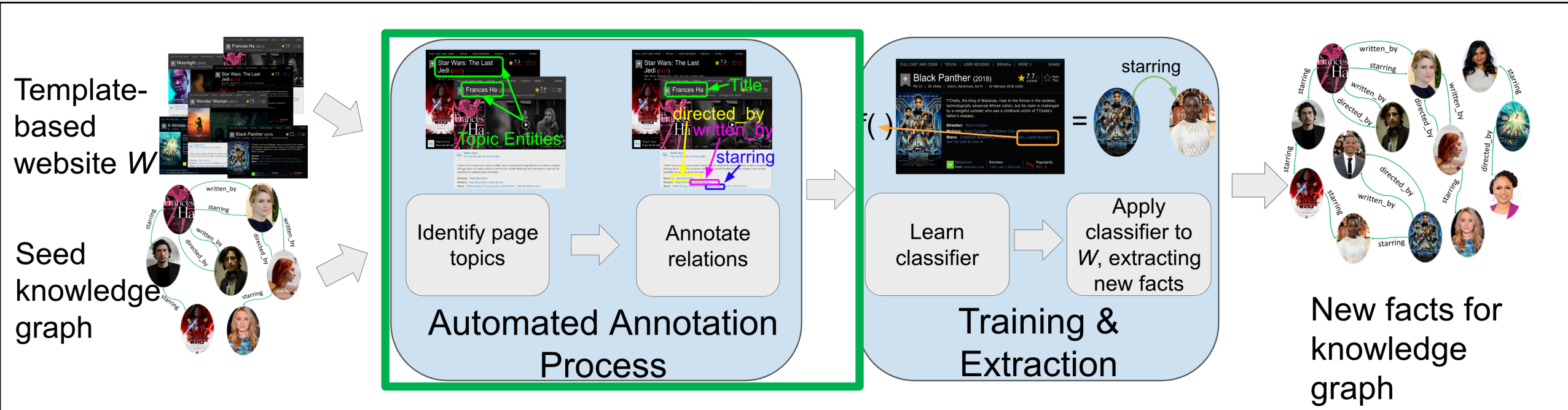


Contribution 1:  
CERES Annotator

Contribution 2:  
CERES Extractor

	No manual annotations	Finds new entities	Accurate
Supervised Wrapper Induction	X	✓	✓
Knowledge Vault	✓	X	X
CERES	✓	✓	✓

# AUTOMATED ANNOTATION PROCESS



Use both local (on a single page) and global (site-wide patterns) information





**Rita Moreno** (1) Top 5000

Actress Soundtrack

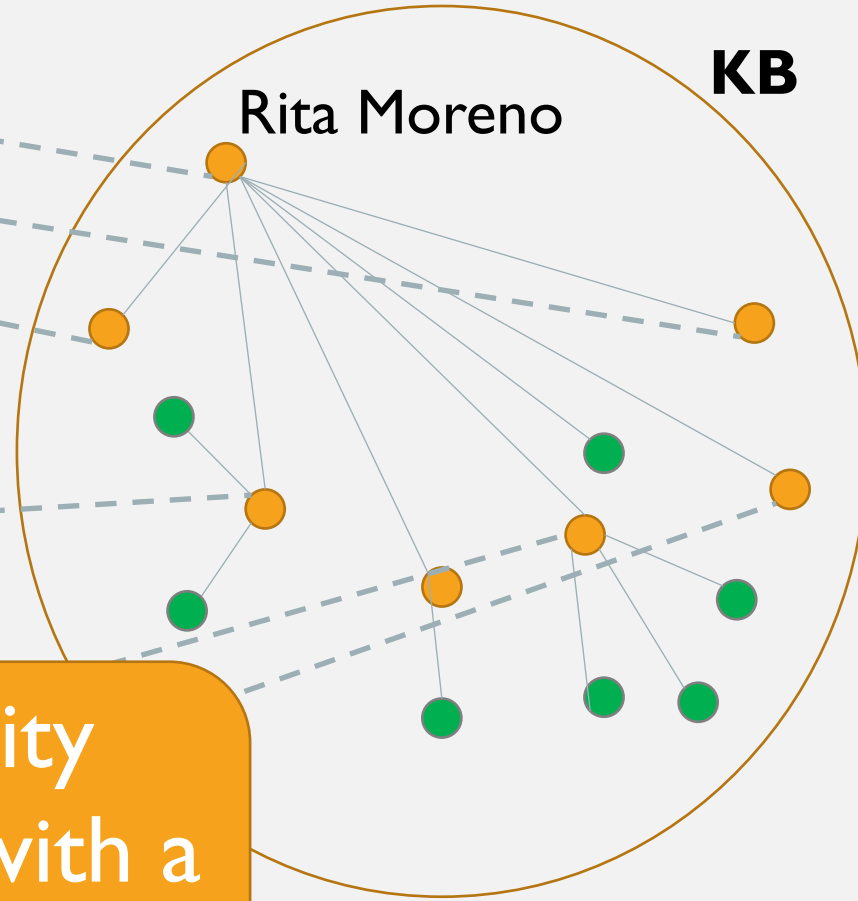
[View Resume](#) | [Official Photos](#) »

Rita Moreno has had a thriving acting career for the better part of six decades. One of the very few performers (and the very first) to win an Oscar, an Emmy, a Tony and a Grammy, she was born Rosita Dolores Alverío in Humacao, Puerto Rico, on December 11, 1931, to seamstress Rosa María (Marcano) and farmer Francisco José "Paco" Alverío. She and ... [See full bio](#) »

**Born:** December 11, 1931 in **Humacao, Puerto Rico**

[More at IMDbPro](#) »

[Contact](#)



## Filmography

Jump to: [Actress](#) | [Soundtrack](#) | [Self](#) | [A](#)

### Actress (155 credits)

#### Nina's World (TV Series)

- Abuelita
- [The Best Ending Ever!](#) (2018) ... Abuelita (voice)
- [Carlos' Winning Shirt](#) (2018) ... Abuelita (voice)
- [Nina Live](#) (2018) ... Abuelita (voice)
- [Nina in Charge](#) (2018) ... Abuelita (voice)

Local: The topic entity should be associated with a large number of entities on the page

Overlapping entities  
Entities in KB not on page



**Rita Moreno** (1) Top 5000

Actress Soundtrack

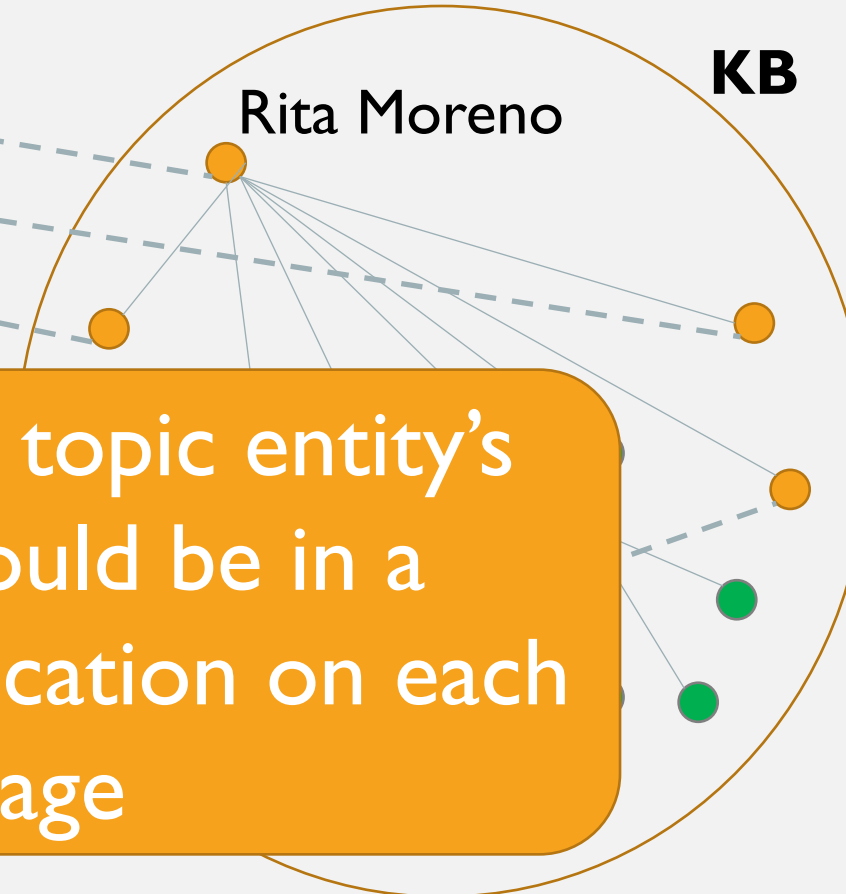
[View Resume](#) | [Official Photos](#) »

Rita Moreno has had a thriving acting career for the better part of six decades. One of the very few performers (and the very first) to win an Oscar, an Emmy, a Tony and a Grammy, she was born Rosita Dolores...  
 Puerto Rico, on December 11, 1931...  
 María (Marcano) and farmer Francisco...  
 She and ... [See full bio](#) »

**Born:** December 11, 1931 in H...

[More at IMDbPro](#) »

**Contact Info:** [View agent, p...](#)



Global: The topic entity's name should be in a consistent location on each page

## Filmography

Jump to: [Actress](#) | [Soundtrack](#) | [Self](#) | [Archive footage](#)

**Actress** (155 credits) Hide

<b>Nina's World</b> (TV Series)	2015-2018
Abuelita	
- <a href="#">The Best Ending Ever!</a> (2018) ... Abuelita (voice)	
- <a href="#">Carlos' Winning Shirt</a> (2018) ... Abuelita (voice)	
<b>Nina Live</b> (2018) ... Abuelita (voice)	
- <a href="#">Nina in Charge</a> (2018) ... Abuelita (voice)	

- Overlapping entities
- Entities in KB not on page

# RELATION ANNOTATION

**Frances Ha** (2012) ★ 7.4/10 56,282 ☆ Rate This

R | 1h 26min | Comedy, Drama, Romance | 23 August 2013 (Brazil)

**Frances Ha**

1:53 | Trailer | 2 VIDEOS | 25 IMAGES

Watch Now  
From \$3.99 (SD) on Prime Video

A New York woman (who doesn't really have an apartment) apprentices for a dance company (though she's not really a dancer) and throws herself headlong into her dreams, even as the possibility of realizing them dwindles.

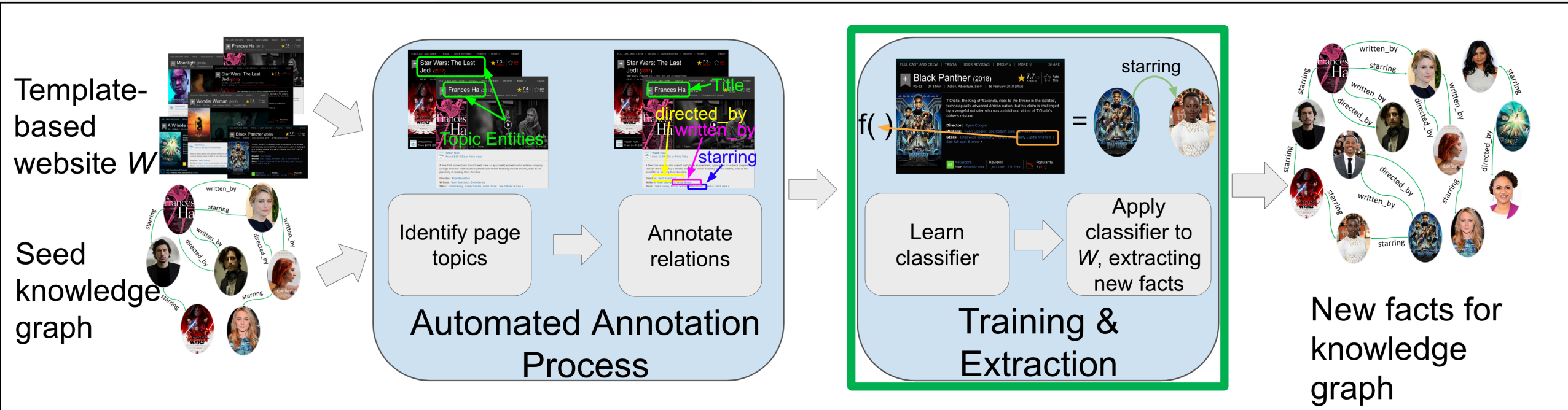
**Director:** Noah Baumbach

**Writers:** Noah Baumbach, Greta Gerwig

**Stars:** Greta Gerwig, Mickey Sumner, Adam Driver [See full cast & crew »](#)

- Annotate known facts found on page
- If ambiguous:
  - Local: Objects of same predicate should be in same section of page
  - Global: Predicates should be in *similar* location on all pages. Cluster all potential mentions of a relation across site, choose most common location.

# TRAINING PROCESS



- Probabilistic classifier
- Robust to noise in training data
  - (compared to wrapper induction)

## MODEL & FEATURES

- Multi-class logistic regression model
  - Input: Featurized DOM node
  - Output: Relation label (or “None”)
- Features based on Vertex (Gulhane et al, ICDE 2011)
  - Tag, ID, Class of ancestors/siblings
  - DOM path to template strings
- Important to limit # of features to prevent overfitting to noise in training data

Experiments show CERES is competitive with state-of-the-art supervised extractors.

# BENCHMARK DATASET

- **Baselines:**
  - CERES-Baseline: Naïve Distant Supervision
  - CERES-Topic: Uses topic identification, but does not resolve ambiguous objects
- SWDE dataset (Hao et al, SIGIR 2011)
- 10 sites in each of 4 domains (Book, Movie, NBA, University)



# SWDE

System	Manual Labels	Movie	NBA Player	University	Book
Hao <i>et al.</i> [19]	yes	0.79	0.82		
XTPath [7]	yes	0.94	<b>0.98</b>		
BigGrams [26]	yes	0.74	0.90		
LODIE-Ideal [15]	no	0.86	0.9		
LODIE-LOD [15]	no	0.76	0.87 <sup>a</sup>	0.91	0.78
RR+WADaR [29]	no	0.73	0.80	0.79	0.70
RR+WADaR 2 [30]	no	0.75	0.91	0.79	0.71
WEIR [4]	no	0.93	0.89	0.97	0.91
Vertex++	yes	0.90	0.97	<b>1.00</b>	0.94
CERES-Baseline	no	NA <sup>b</sup>	0.78	0.72	0.27
CERES-Topic	no	<b>0.99<sup>a</sup></b>	0.97	0.96	0.72
CERES-Full	no	<b>0.99<sup>a</sup></b>	<b>0.98</b>	0.94	0.76

State-of-the-art results in two verticals. (better than supervised systems!)



# SWDE

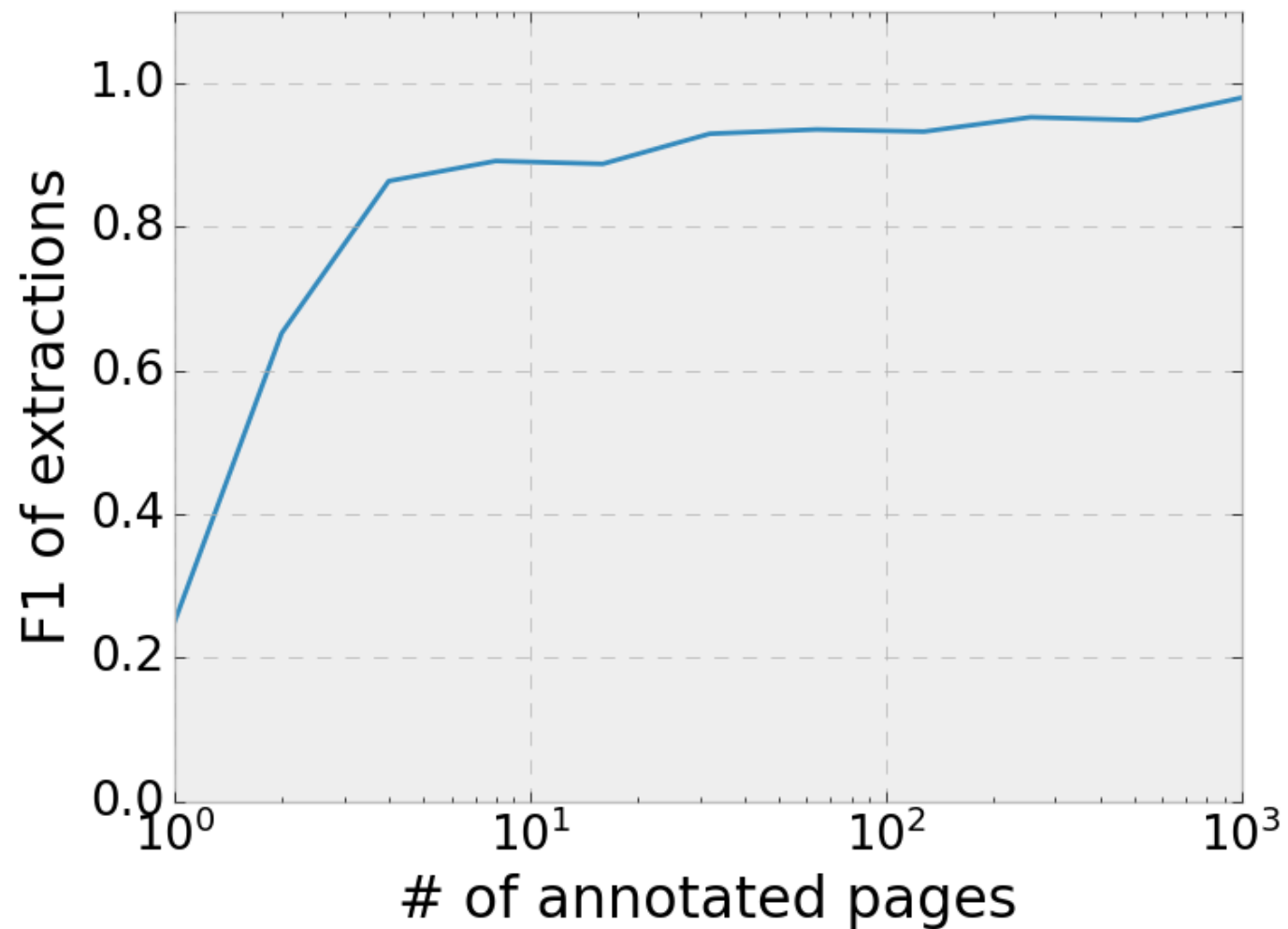
Competitive with top  
unsupervised systems in a  
third

	Manual Labels	Movie	NBA Player	University	Book
	yes	0.79	0.82	0.83	0.86
	yes	0.94	<b>0.98</b>	0.98	<b>0.97</b>
BigGrams [26]	yes	0.74	0.90	0.79	0.78
LODIE-Ideal [15]	no	0.86	0.9	0.96	0.85
LODIE-LOD [15]	no	0.76	0.87 <sup>a</sup>	0.91 <sup>a</sup>	0.78
RR+WADaR [29]	no	0.73	0.80	0.79	0.70
RR+WADaR 2 [30]	no	0.75	0.91	0.79	0.71
WEIR [4]	no	0.93	0.89	0.97	0.91
Vertex++	yes	0.90	0.97	<b>1.00</b>	0.94
CERES-Baseline	no	NA <sup>b</sup>	0.78	0.72	0.27
CERES-Topic	no	<b>0.99<sup>a</sup></b>	0.97	0.96	0.72
CERES-Full	no	<b>0.99<sup>a</sup></b>	<b>0.98</b>	0.94	0.76

# SWDE

System	Manual Labels	Movie	NBA Player	University	Book
Hao <i>et al.</i> [19]	Weak on one vertical. Why?			0.83	0.86
XTPath [7]				0.98	<b>0.97</b>
BigGrams [26]				0.79	0.78
LODIE-Ideal [15]				0.96	0.85
LODIE-LOD [15]		no	0.70	0.87	0.91 <sup>a</sup>
RR+WADaR [29]	Only a few overlapping entities between websites and seed KB			0.79	0.70
RR+WADaR 2 [30]				0.79	0.71
WEIR [4]				0.97	0.91
Vertex++				<b>1.00</b>	0.94
CERES-Baseline	no	NA <sup>b</sup>	0.78	0.72	0.27
CERES-Topic	no	<b>0.99</b> <sup>a</sup>	0.97	0.96	0.72
CERES-Full	no	<b>0.99</b> <sup>a</sup>	<b>0.98</b>	0.94	0.76

# ABOUT 10 ANNOTATED PAGES ARE NEEDED TO LEARN

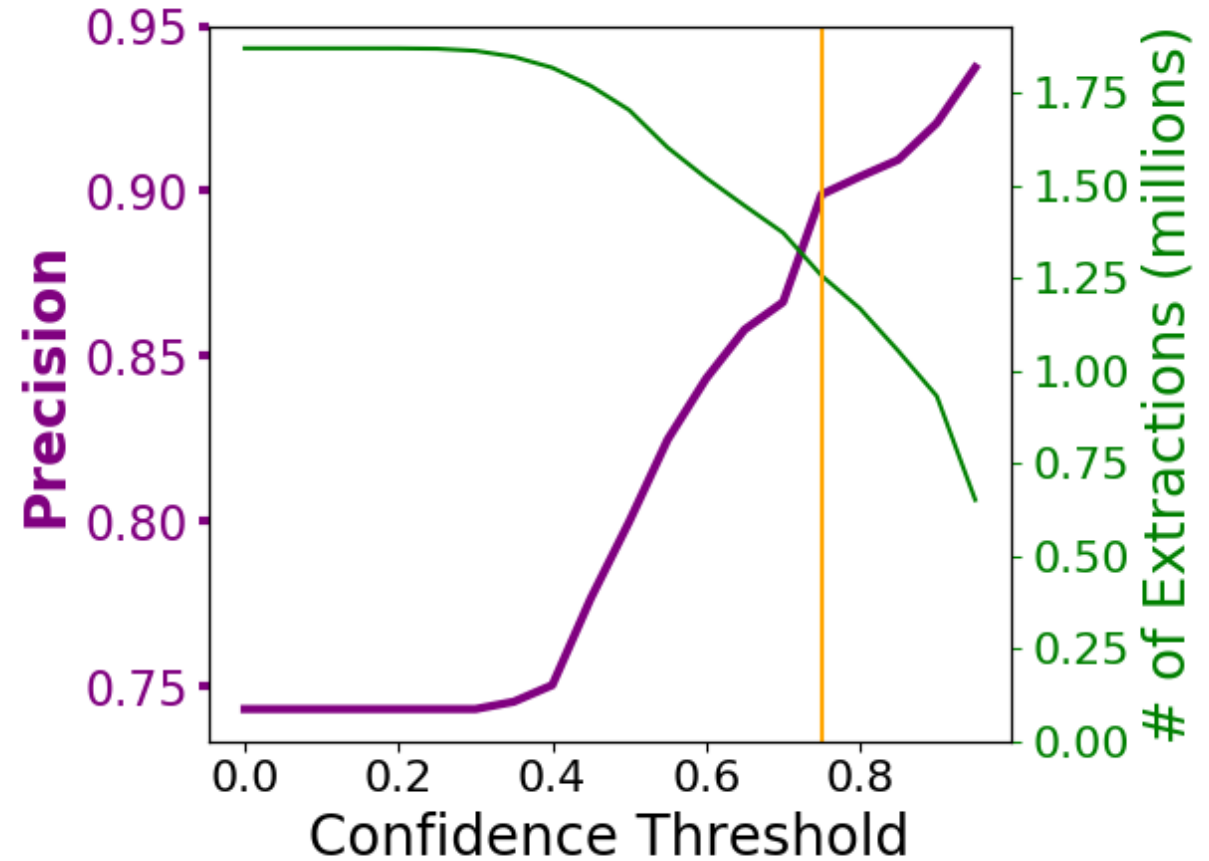


## COMMONCRAWL MOVIE WEBSITES

- 33 websites, 400,000 pages (not all semi-structured)
- Mostly long-tail (foreign, documentary, animated)
- 7 languages

# COMMONCRAWL MOVIE WEBSITES

- 90% precision (compared to ~63% with Knowledge Vault)
- 1.25 million extractions
- Extracted 2.6 new entities for every annotated entity



## TAKEAWAYS

- Automatic extraction from semi-structured pages is practical
  - If you have existing KB you can:
    - Identify new entities
    - Extract new facts with high-precision
- Probabilistic classifiers are effective at semi-structured extraction
  - Allow for precision/recall tradeoff

# FUTURE WORK

## 기술 사양

항목

패스  
너사  
스타일

부분 나사산

축  
시  
템

패스  
산  
표  
준

**ASTM  
A574, IFI-  
115**

12  
플  
지  
스  
크  
류  
타입

나사  
길  
이

**1-1 / 4 "**

기본  
소재

**스틸**

헤드  
높이

**5/16 "**

I don't speak Korean, but I can tell this is a key-value pair. Can a computer?

