Towards the Exploration of Archives

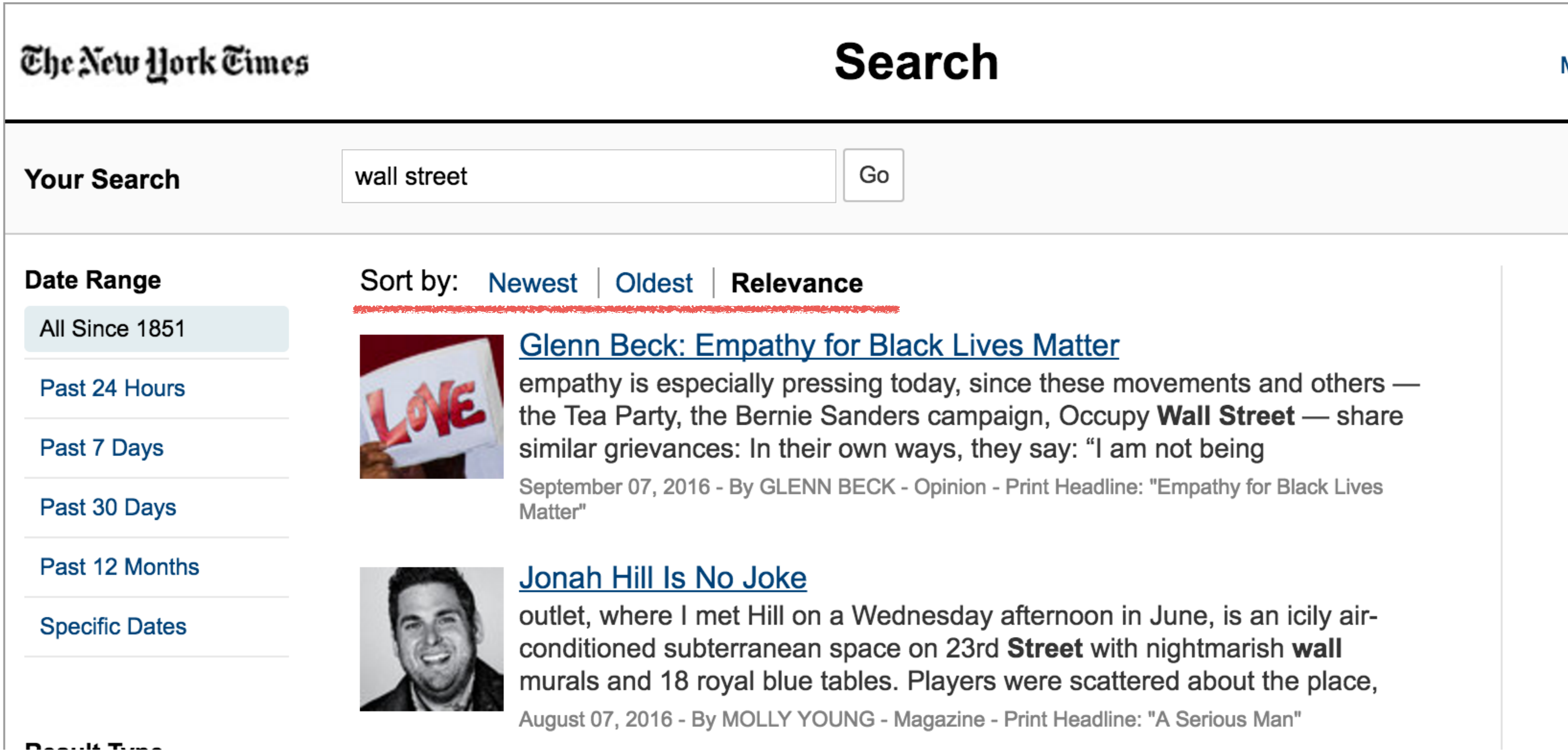
Jaspreet Singh
singh@l3s.de
🐦 @movedthecheese

Forschungszentrum · Research Center L3S

ALEXANDRIA

Exploration?

How do we explore?



The New York Times

Search

Your Search

wall street

Date Range

- All Since 1851
- Past 24 Hours
- Past 7 Days
- Past 30 Days
- Past 12 Months
- Specific Dates

Result Type

Sort by: Newest | Oldest | **Relevance**

Glenn Beck: Empathy for Black Lives Matter

empathy is especially pressing today, since these movements and others — the Tea Party, the Bernie Sanders campaign, Occupy **Wall Street** — share similar grievances: In their own ways, they say: "I am not being

September 07, 2016 - By GLENN BECK - Opinion - Print Headline: "Empathy for Black Lives Matter"

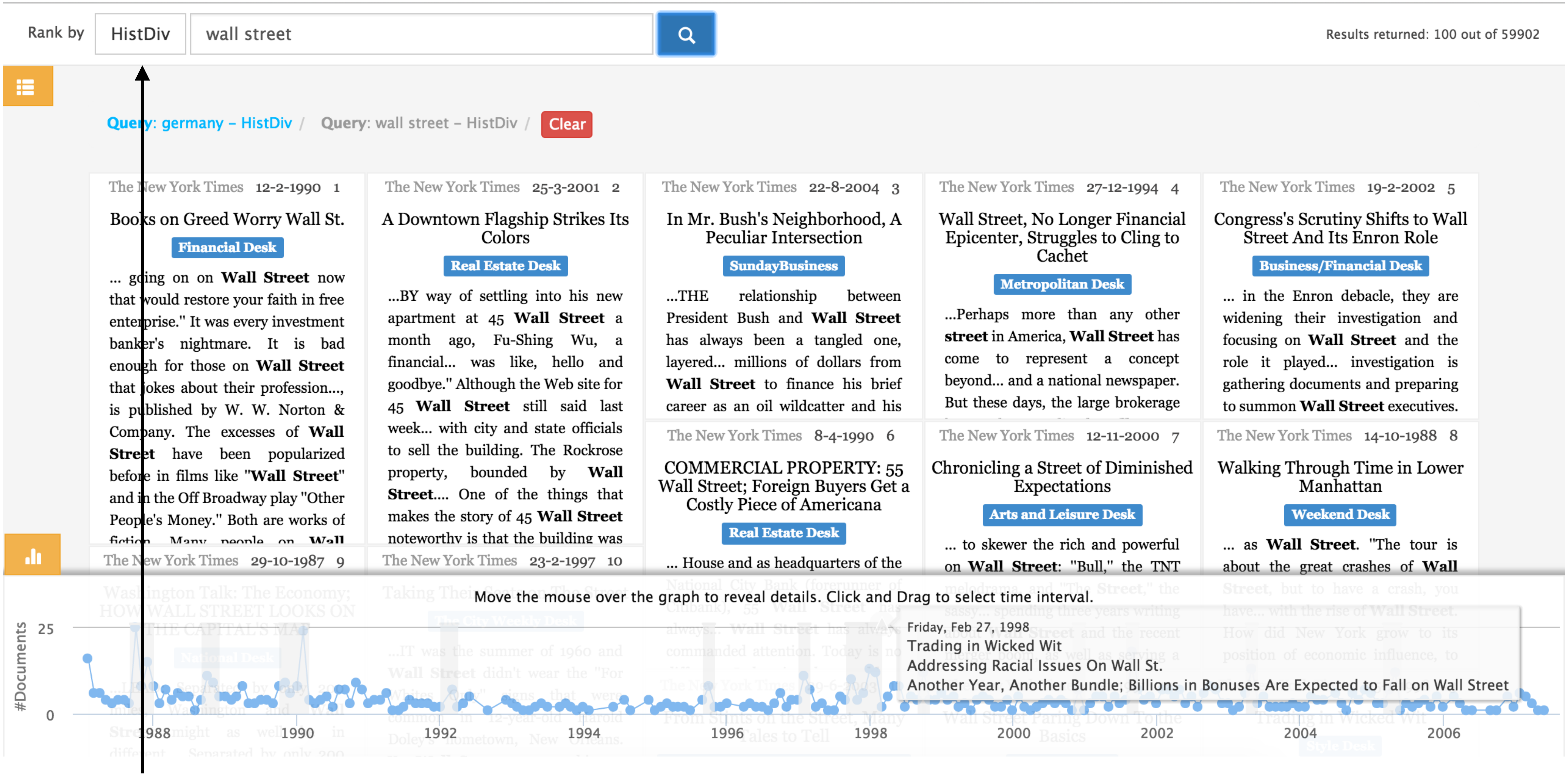
Jonah Hill Is No Joke

outlet, where I met Hill on a Wednesday afternoon in June, is an icily air-conditioned subterranean space on 23rd **Street** with nightmarish **wall** murals and 18 royal blue tables. Players were scattered about the place,

August 07, 2016 - By MOLLY YOUNG - Magazine - Print Headline: "A Serious Man"

Archive Search Engine for **New York Times** (1987 - 2007)

SIGIR'16



Results returned: 100 out of 59902

Query: germany – HistDiv / **Query**: wall street – HistDiv / Clear

The New York Times 12-2-1990 1

Books on Greed Worry Wall St.

Financial Desk

... going on on **Wall Street** now that would restore your faith in free enterprise." It was every investment banker's nightmare. It is bad enough for those on **Wall Street** that jokes about their profession..., is published by W. W. Norton & Company. The excesses of **Wall Street** have been popularized before in films like "**Wall Street**" and in the Off Broadway play "Other People's Money." Both are works of fiction. Many people on **Wall**

The New York Times 25-3-2001 2

A Downtown Flagship Strikes Its Colors

Real Estate Desk

...BY way of settling into his new apartment at 45 **Wall Street** a month ago, Fu-Shing Wu, a financial... was like, hello and goodbye." Although the Web site for 45 **Wall Street** still said last week... with city and state officials to sell the building. The Rockrose property, bounded by **Wall Street**.... One of the things that makes the story of 45 **Wall Street** noteworthy is that the building was

The New York Times 22-8-2004 3

In Mr. Bush's Neighborhood, A Peculiar Intersection

SundayBusiness

...THE relationship between President Bush and **Wall Street** has always been a tangled one, layered... millions of dollars from **Wall Street** to finance his brief career as an oil wildcatter and his

The New York Times 27-12-1994 4

Wall Street, No Longer Financial Epicenter, Struggles to Cling to Cachet

Metropolitan Desk

...Perhaps more than any other **street** in America, **Wall Street** has come to represent a concept beyond... and a national newspaper. But these days, the large brokerage

The New York Times 19-2-2002 5

Congress's Scrutiny Shifts to Wall Street And Its Enron Role

Business/Financial Desk

... in the Enron debacle, they are widening their investigation and focusing on **Wall Street** and the role it played... investigation is gathering documents and preparing to summon **Wall Street** executives.

The New York Times 8-4-1990 6

COMMERCIAL PROPERTY: 55 Wall Street; Foreign Buyers Get a Costly Piece of Americana

Real Estate Desk

... House and as headquarters of the

The New York Times 12-11-2000 7

Chronicling a Street of Diminished Expectations

Arts and Leisure Desk

... to skewer the rich and powerful on **Wall Street**: "Bull," the TNT

The New York Times 14-10-1988 8

Walking Through Time in Lower Manhattan

Weekend Desk

... as **Wall Street**. "The tour is about the great crashes of **Wall**

The New York Times 29-10-1987 9

The New York Times 23-2-1997 10

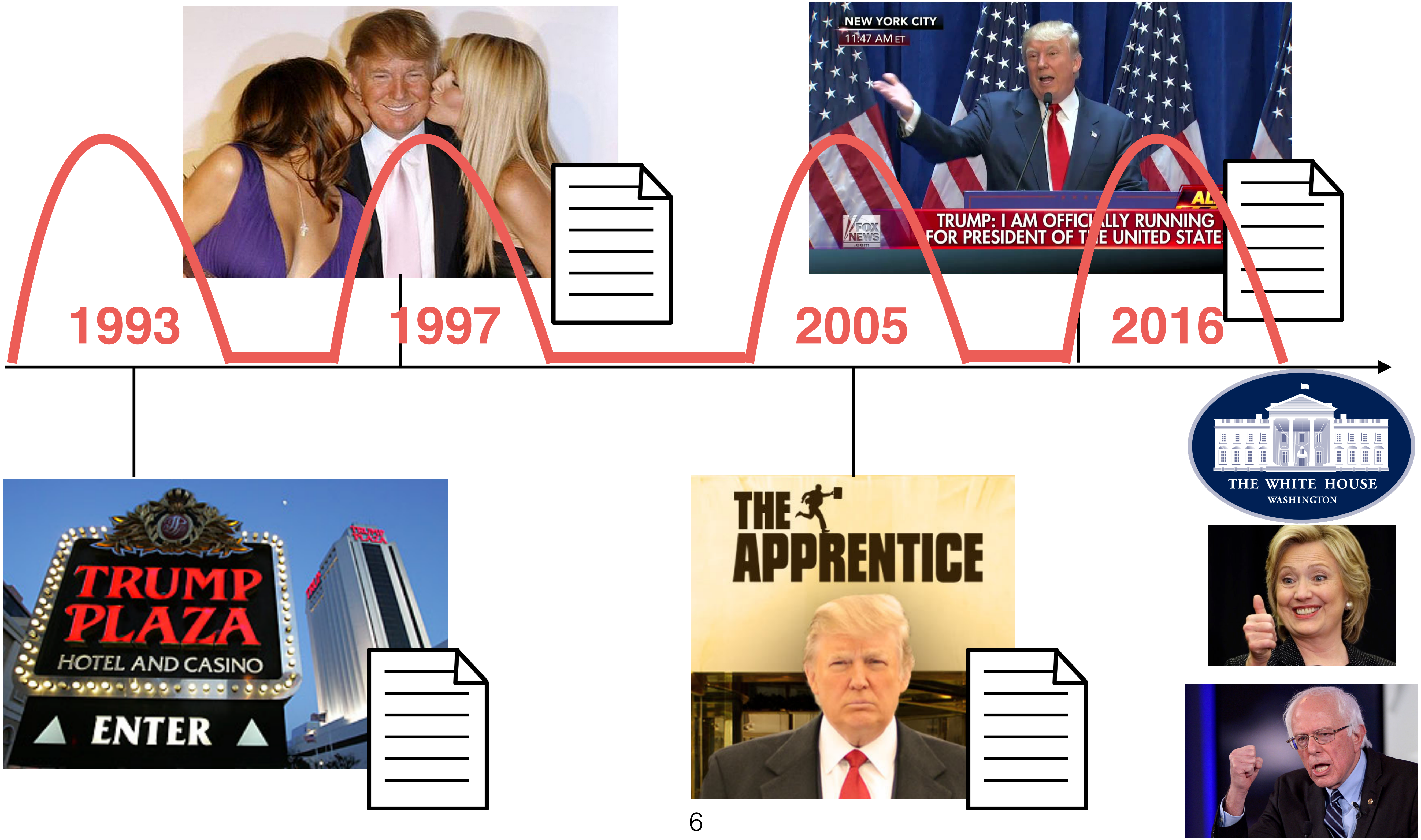
Move the mouse over the graph to reveal details. Click and Drag to select time interval.

Friday, Feb 27, 1998
Trading in Wicked Wit
Addressing Racial Issues On Wall St.
Another Year, Another Bundle; Billions in Bonuses Are Expected to Fall on Wall Street

HistDiv (CHIIR'16)

http://bit.ly/archive-search

History by Diversity(CHIIR'16)



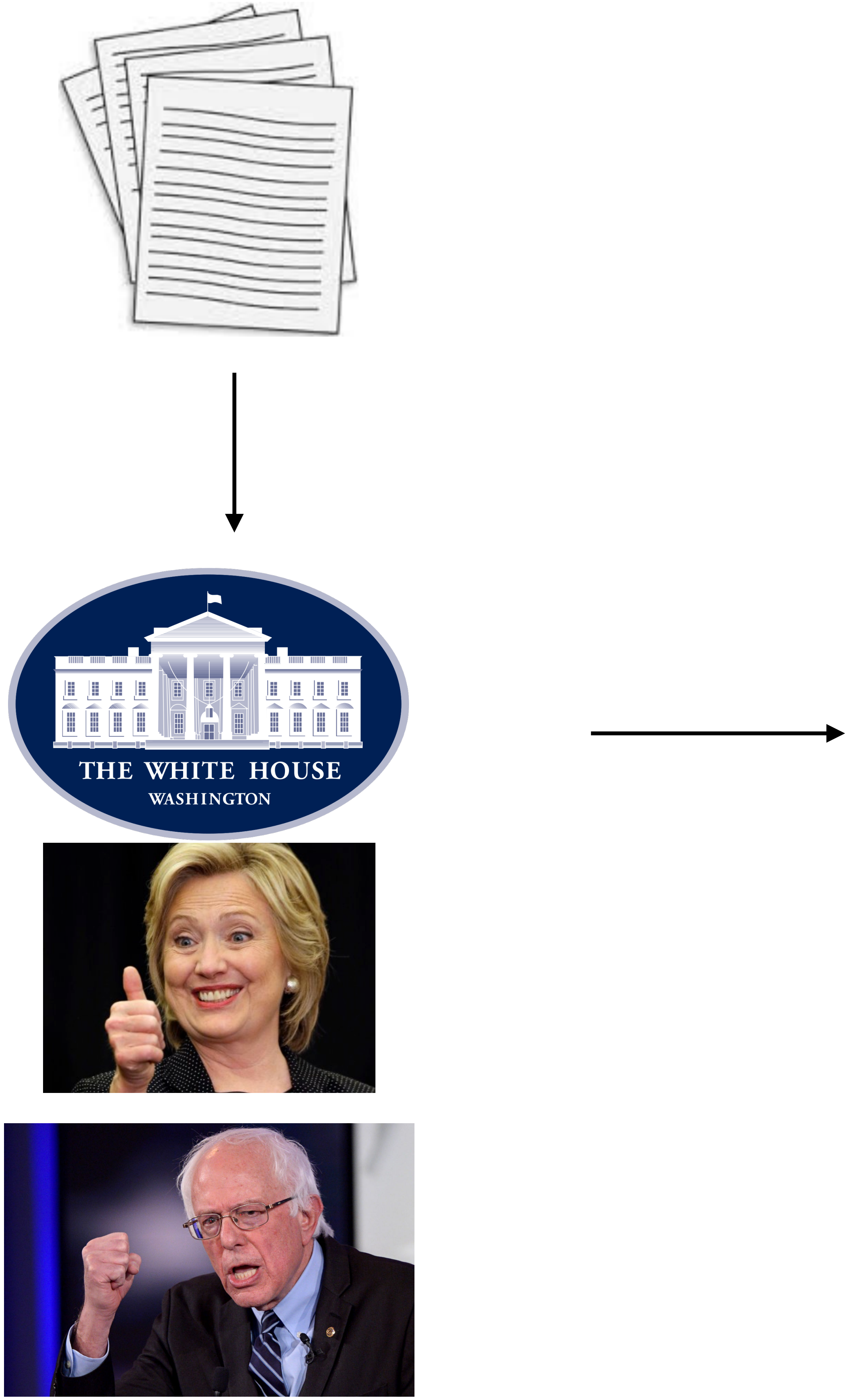
1993

1997

2005

2016

NEW YORK CITY
11:47 AM ET
FOX NEWS .com
TRUMP: I AM OFFICIALLY RUNNING FOR PRESIDENT OF THE UNITED STATES
TRUMP PLAZA HOTEL AND CASINO
ENTER
THE APPRENTICE
THE WHITE HOUSE
WASHINGTON

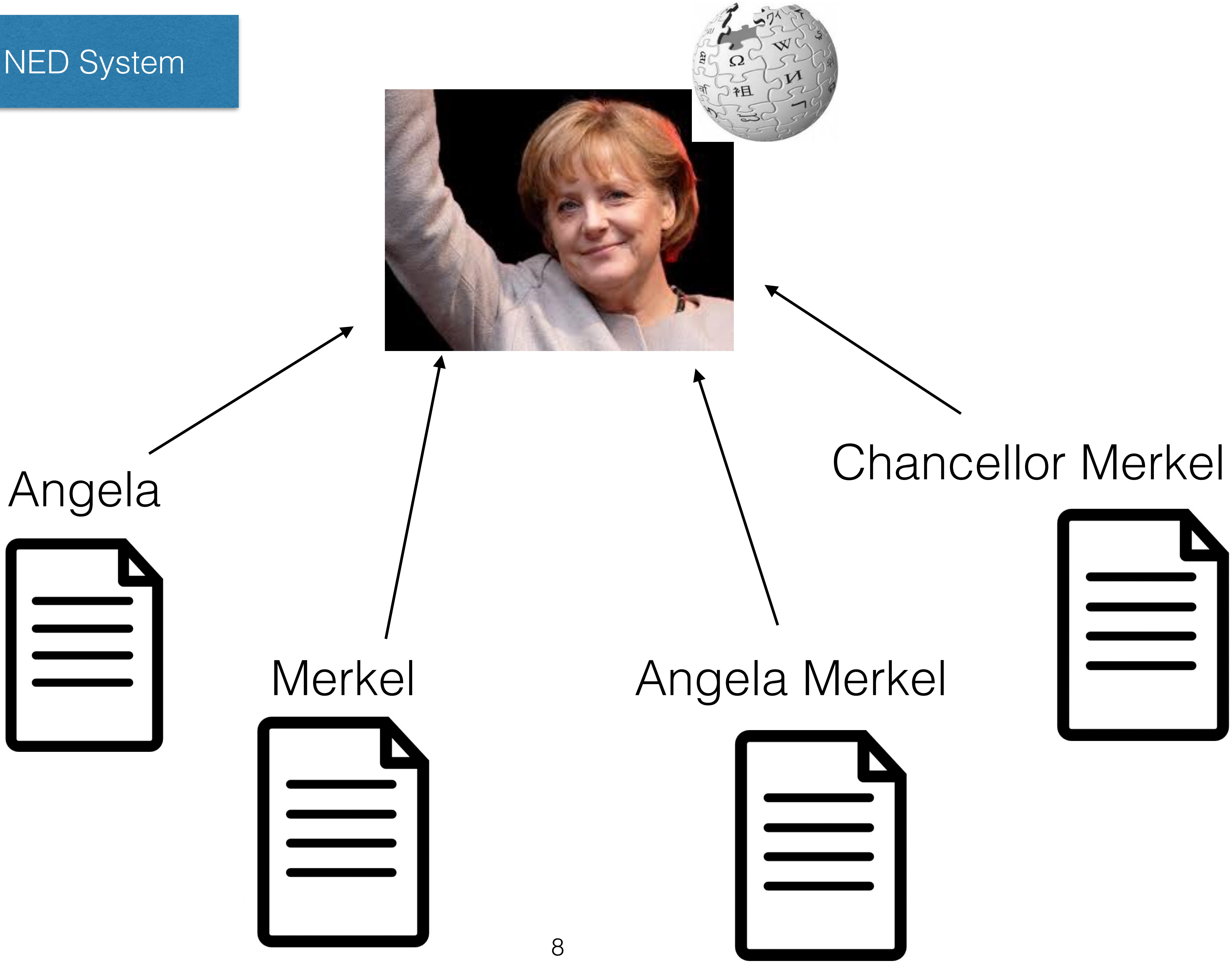


HistDiv

Semantic Search

Text Analysis

NED System



Angela

Merkel

Angela Merkel

Chancellor Merkel

first Labour MP,
Keir Hardie

NED System

Discovering Entities with Just a Little Help from You

Jaspreet Singh, Johannes Hoffart, Avishek Anand

Preprint: l3s.de/~singh

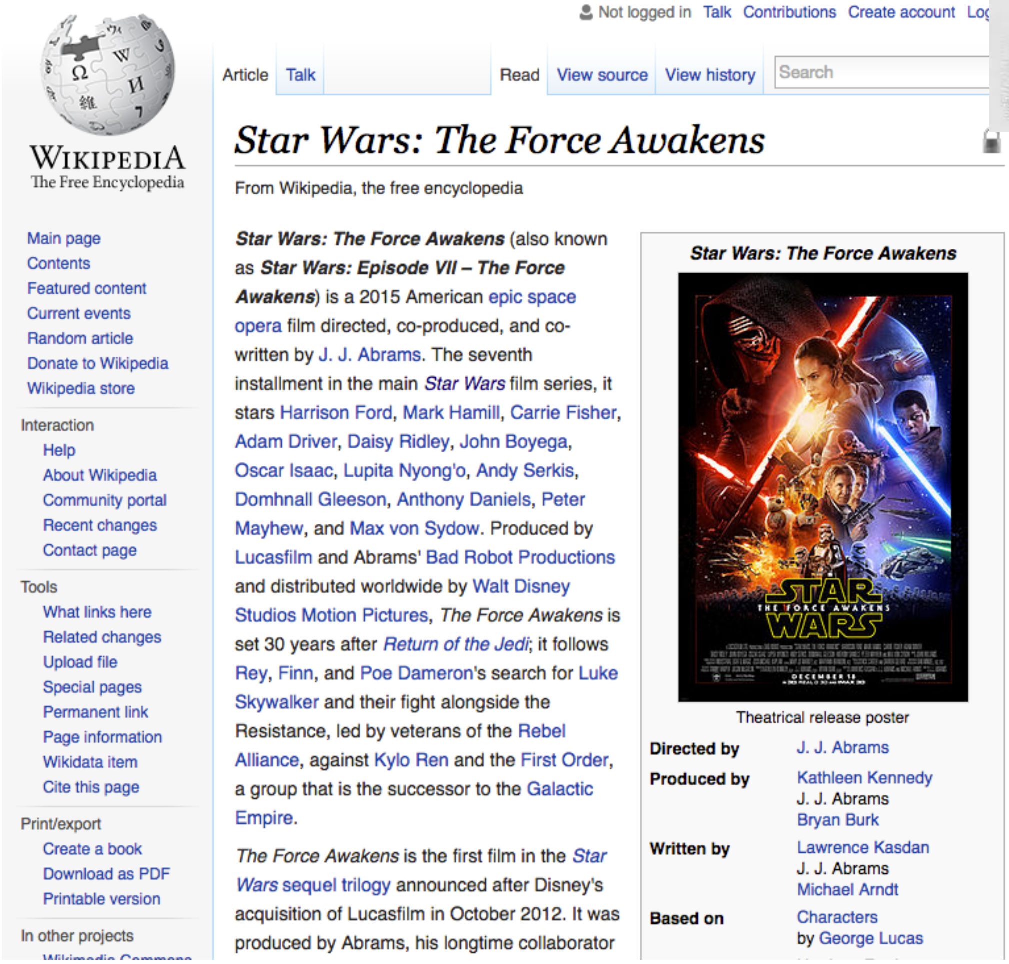
CIKM 25 ANNIVERSARY 2016
Indianapolis
October 24-28, 2016

NED System

Mention to Entity Mapping

Entity Descriptions (Keyphrases)

Game maker Hasbro will include female **Star Wars: The Force Awakens** character ***Rey*** in their Star Wars themed **Monopoly** game



Star Wars: The Force Awakens

From Wikipedia, the free encyclopedia

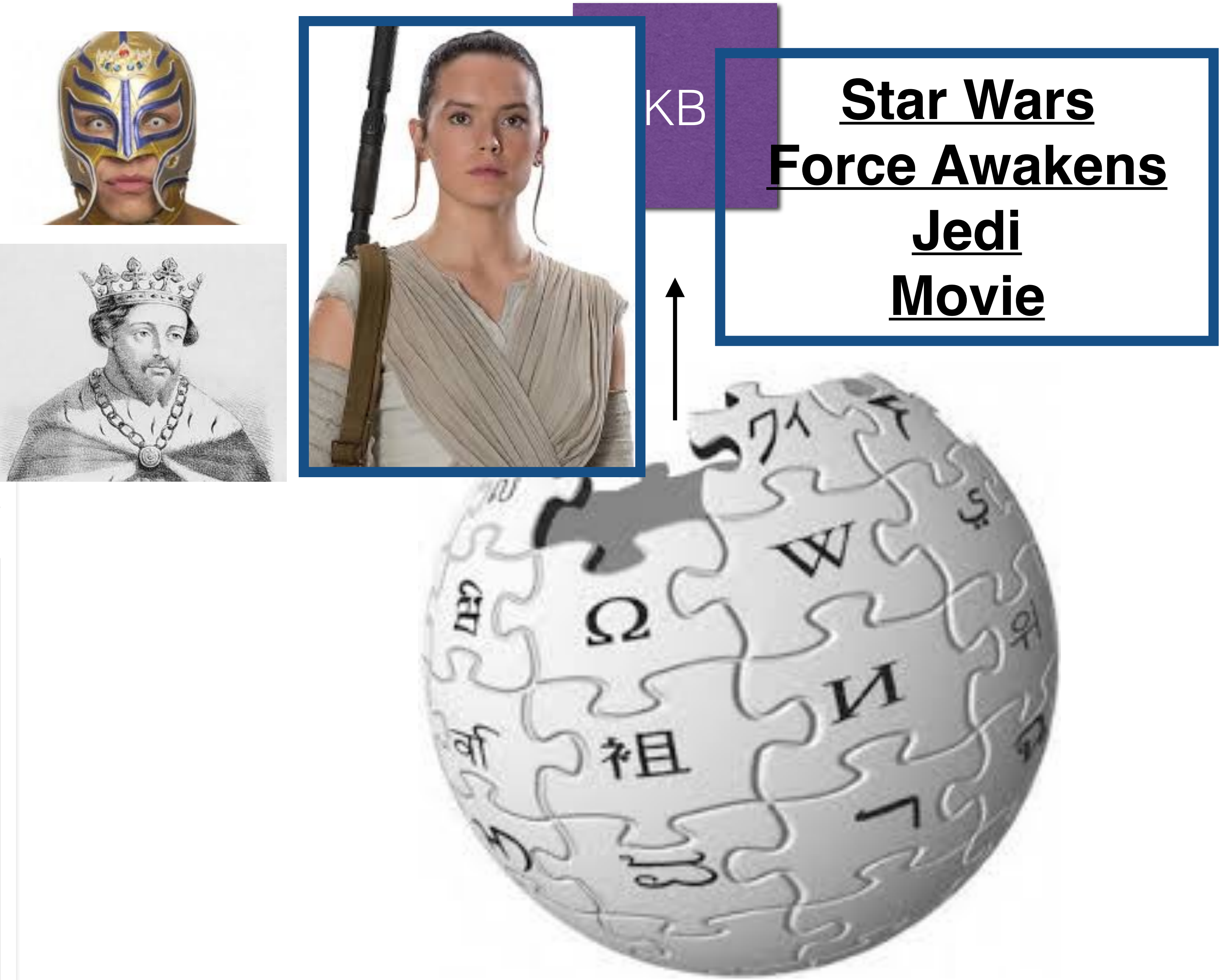
Star Wars: The Force Awakens (also known as ***Star Wars: Episode VII – The Force Awakens***) is a 2015 American epic space opera film directed, co-produced, and co-written by J. J. Abrams. The seventh installment in the main *Star Wars* film series, it stars Harrison Ford, Mark Hamill, Carrie Fisher, Adam Driver, Daisy Ridley, John Boyega, Oscar Isaac, Lupita Nyong'o, Andy Serkis, Domhnall Gleeson, Anthony Daniels, Peter Mayhew, and Max von Sydow. Produced by Lucasfilm and Abrams' Bad Robot Productions and distributed worldwide by Walt Disney Studios Motion Pictures, *The Force Awakens* is set 30 years after *Return of the Jedi*; it follows Rey, Finn, and Poe Dameron's search for Luke Skywalker and their fight alongside the Resistance, led by veterans of the Rebel Alliance, against Kylo Ren and the First Order, a group that is the successor to the Galactic Empire.

The Force Awakens is the first film in the *Star Wars* sequel trilogy announced after Disney's acquisition of Lucasfilm in October 2012. It was produced by Abrams, his longtime collaborator

Star Wars: The Force Awakens

Theatrical release poster

Directed by	J. J. Abrams
Produced by	Kathleen Kennedy J. J. Abrams Bryan Burk
Written by	Lawrence Kasdan J. J. Abrams Michael Arndt
Based on	Characters by George Lucas



Star Wars
Force Awakens
Jedi
Movie

Popular entities?

Mine Wikipedia! (and homepages)

Fully Automatic

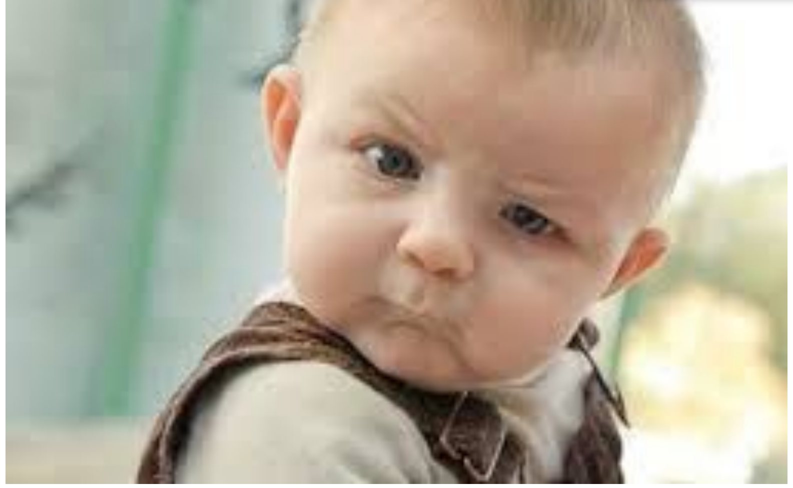
Longtail entities?

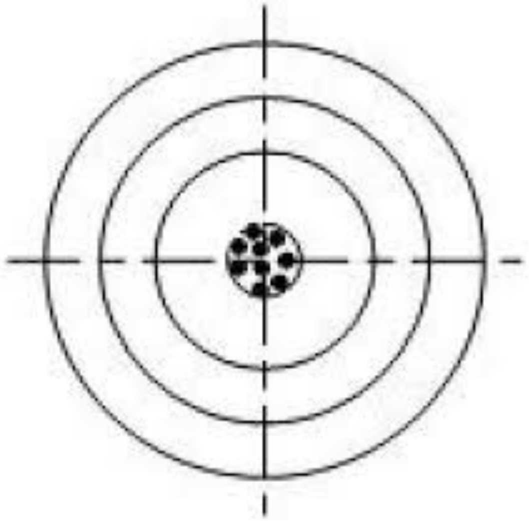
No Wikipedia page or too little context. Maybe the web?

Fully Manual?

Keyphrases
that describe the entity

Human in the loop





NED System

Problem Definition

Document Collection

(ambiguous)
Entity Mention

Star Wars
Force Awakens
Jedi
Movie

Coverage Problem
(over Keyphrases)

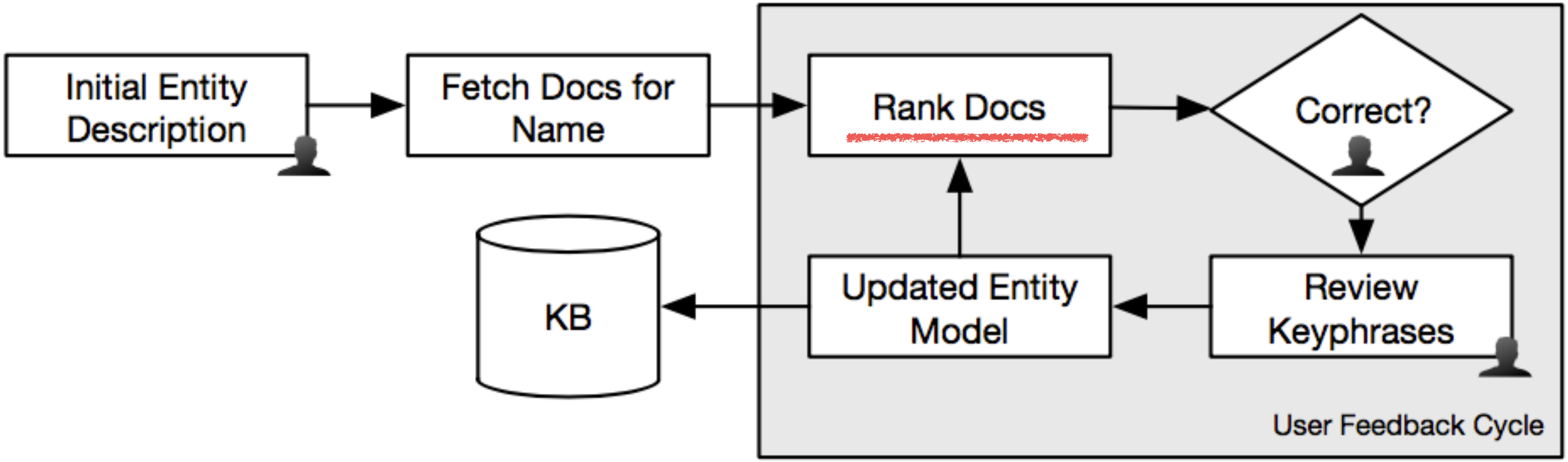
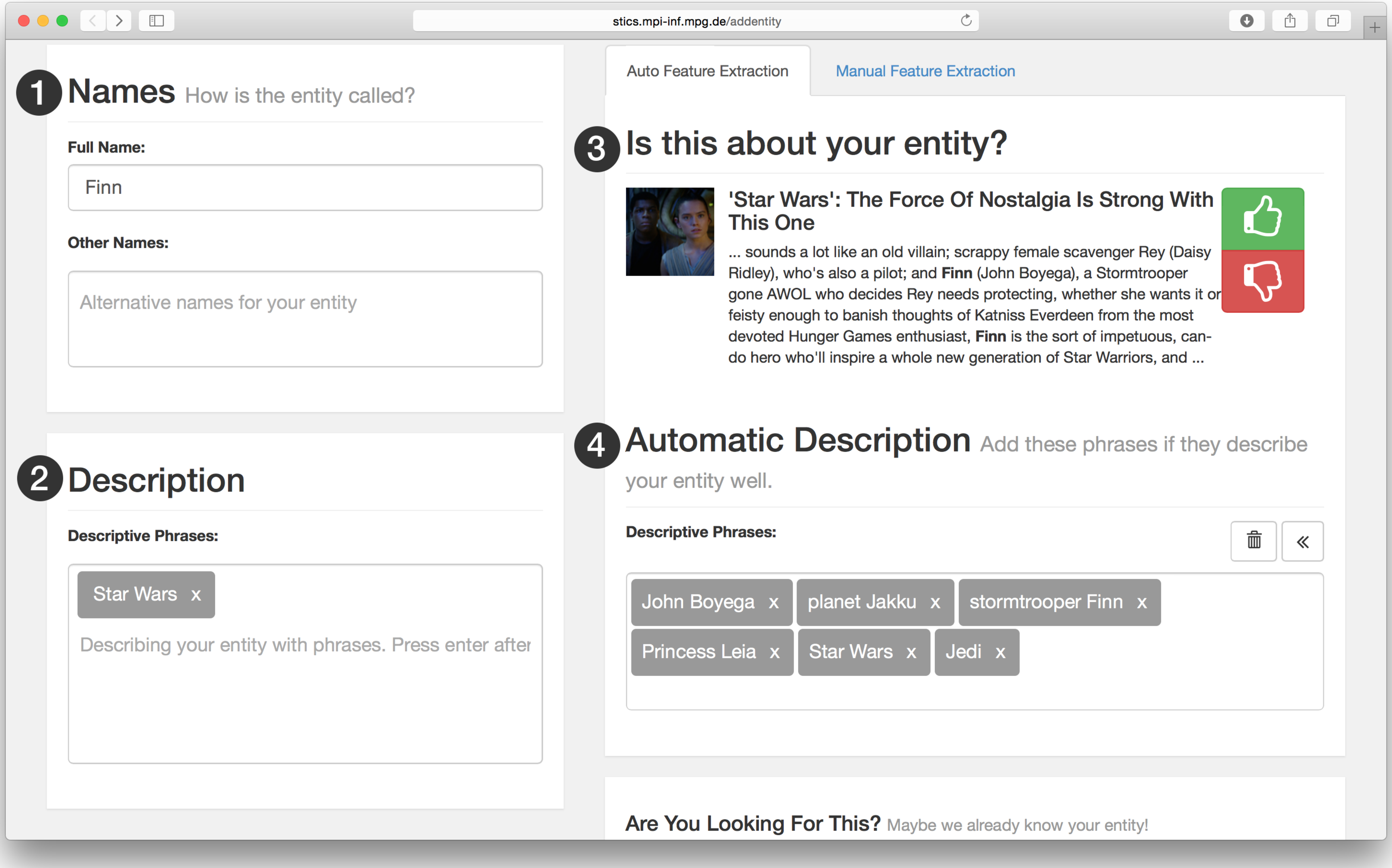


Figure 2: Harvesting Keyphrases with the Help of the User



stics.mpi-inf.mpg.de/addentity

1 Names How is the entity called?

Full Name:

Finn

Other Names:

Alternative names for your entity

2 Description

Descriptive Phrases:

Star Wars x

Describing your entity with phrases. Press enter after

Auto Feature Extraction | Manual Feature Extraction

3 Is this about your entity?

'Star Wars': The Force Of Nostalgia Is Strong With This One

... sounds a lot like an old villain; scrappy female scavenger Rey (Daisy Ridley), who's also a pilot; and **Finn** (John Boyega), a Stormtrooper gone AWOL who decides Rey needs protecting, whether she wants it or feisty enough to banish thoughts of Katniss Everdeen from the most devoted Hunger Games enthusiast, **Finn** is the sort of impetuous, can-do hero who'll inspire a whole new generation of Star Warriors, and ...

4 Automatic Description Add these phrases if they describe your entity well.

Descriptive Phrases:

John Boyega x | planet Jakku x | stormtrooper Finn x | Princess Leia x | Star Wars x | Jedi x

Are You Looking For This? Maybe we already know your entity!

Johannes Hoffart, Dragan Milchevski, Gerhard Weikum, Avishek Anand, and Jaspreet Singh. **The Knowledge Awakens: Keeping Knowledge Bases Fresh with Emerging Entities**. In Proceedings of the 25th International Conference Companion on World Wide Web (WWW '16 Companion).

Approach
(~~Singh-Rank~~)

Query: mention + a few initial keywords

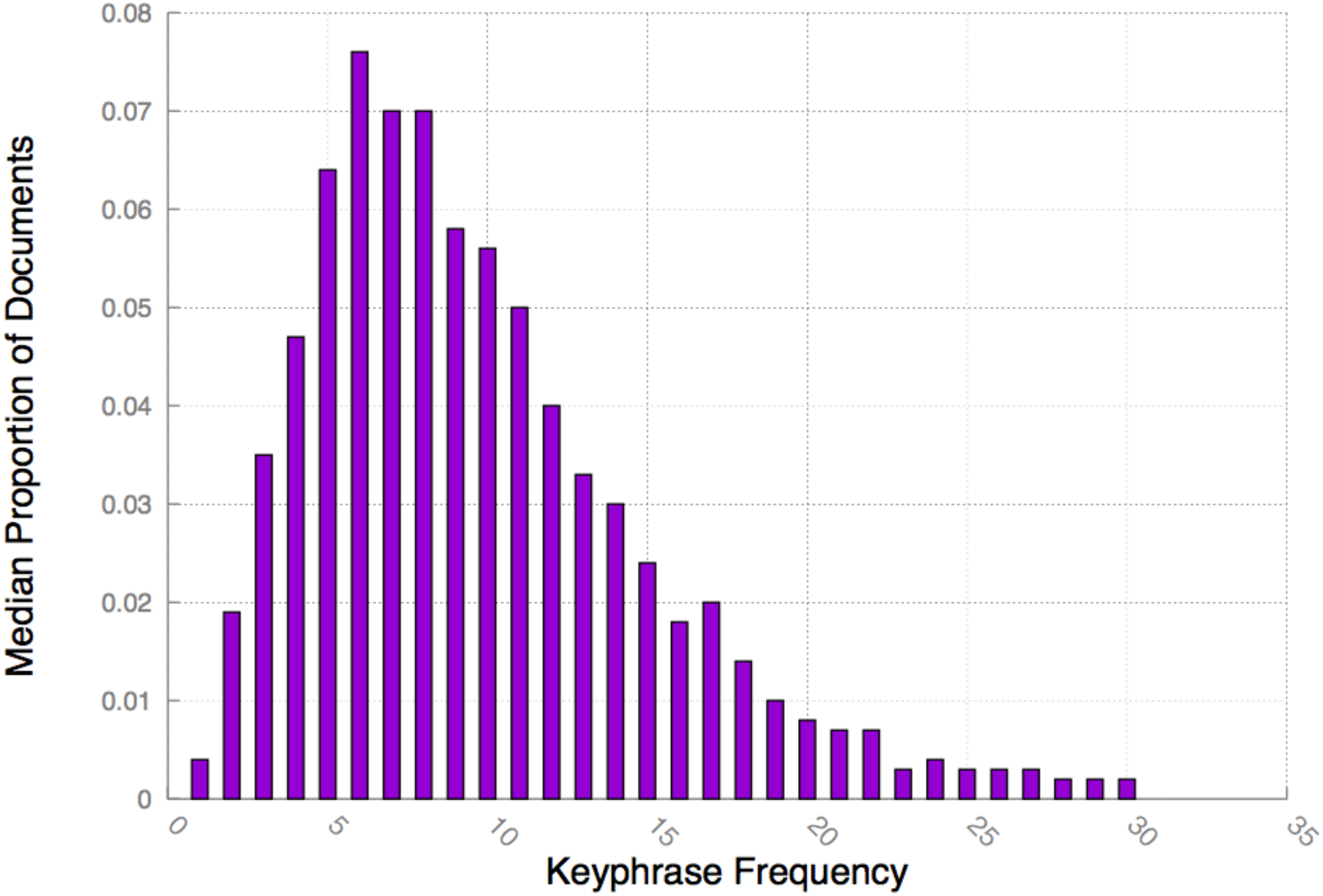
Iterative Ranking Approach

User Feedback

User Engagement

Robustness

User Feedback



Rocchio Algorithm

Expand Query with Keyphrases (positive and negative)

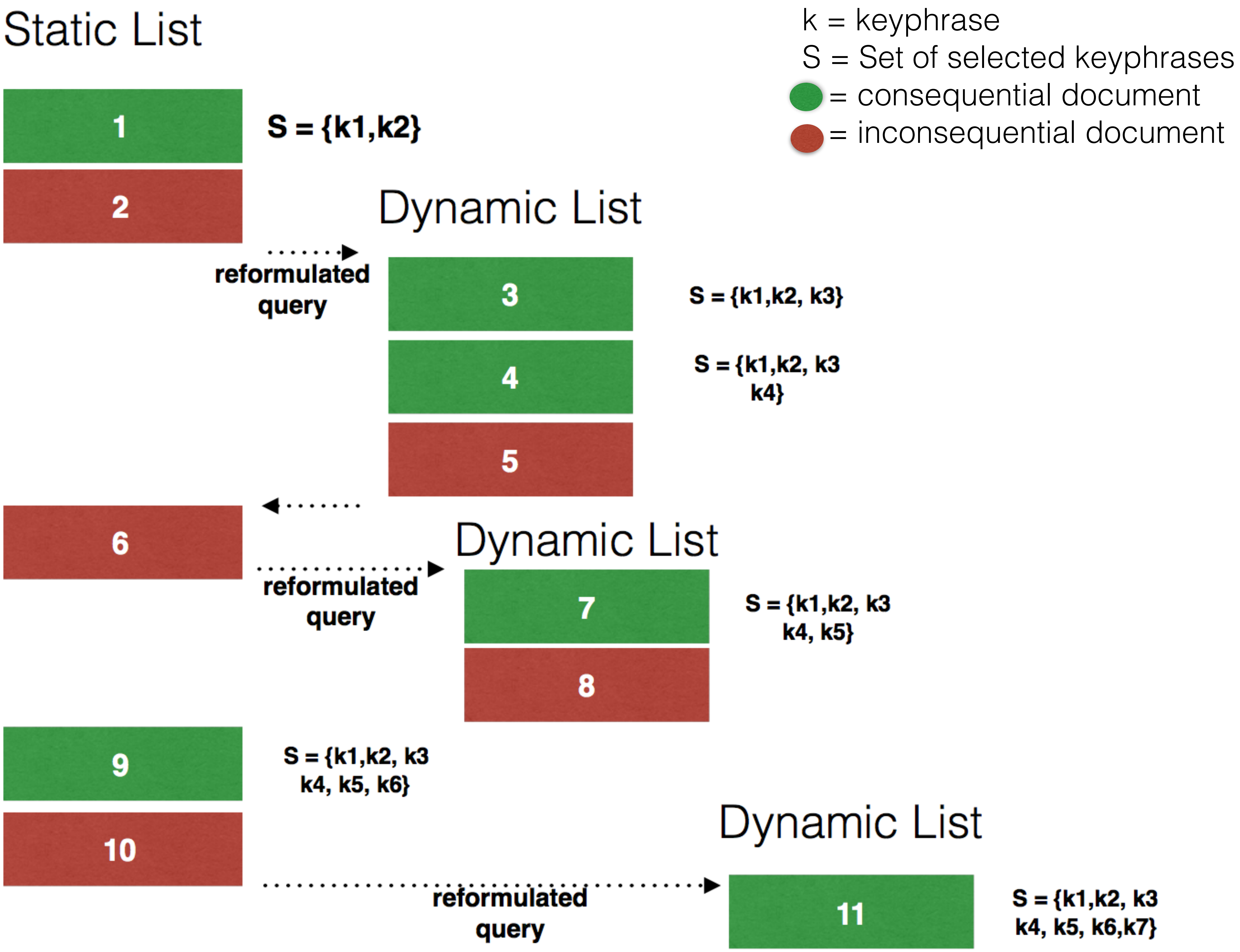
Engagement & Robustness

MJ - golf, basketball

ation due to feedback **cation** to find new keyphrases

- Concept Drift due to diversification
  - Bring it back on track with **Interleaving**

tar Wars - books and movies



Static List

k = keyphrase
S = Set of selected keyphrases
🟢 = consequential document
🔴 = inconsequential document

1

2

S = {k1,k2}

reformulated query

Dynamic List

3

S = {k1,k2, k3}

4

S = {k1,k2, k3 k4}

5

6

reformulated query

Dynamic List

7

S = {k1,k2, k3 k4, k5}

8

9

10

S = {k1,k2, k3 k4, k5, k6}

reformulated query

Dynamic List

11

S = {k1,k2, k3 k4, k5, k6,k7}

Approach Setup

- Static List: LM, Diversified list
- Dynamic List: LM, Diversified list

- Diversification:
  - Keyphrases as aspects
    - Large space, noisy
    - Nigerian people vs People of Nigeria

  - **Entities** as aspects
    - **Joint Disambiguation** is used
    - Smaller space, **canonicalized**
    - Keyphrases occur in the vicinity of entities

Baselines

- **LM** - Language Model
- **LM-Feedback**
- **DivKp** - Keyphrase Diversification
- **DivEnt** - Entity Diversification
- **DivKp-Feedback, DivEnt-Feedback**
- Interleaving Approaches
    - I (static-list, dynamic-list)
    - example: **I (DivEnt, DivEnt-Feedback)**

Measures

- Extrinsic Measure
  - **Disambiguation Accuracy**

- Intrinsic Measures
  - **Coverage** of relevant keyphrases
  - **User Engagement** Index

- **Engagement Index**

A= + - + - + - 🟢

B= + - - - ++ 🔴

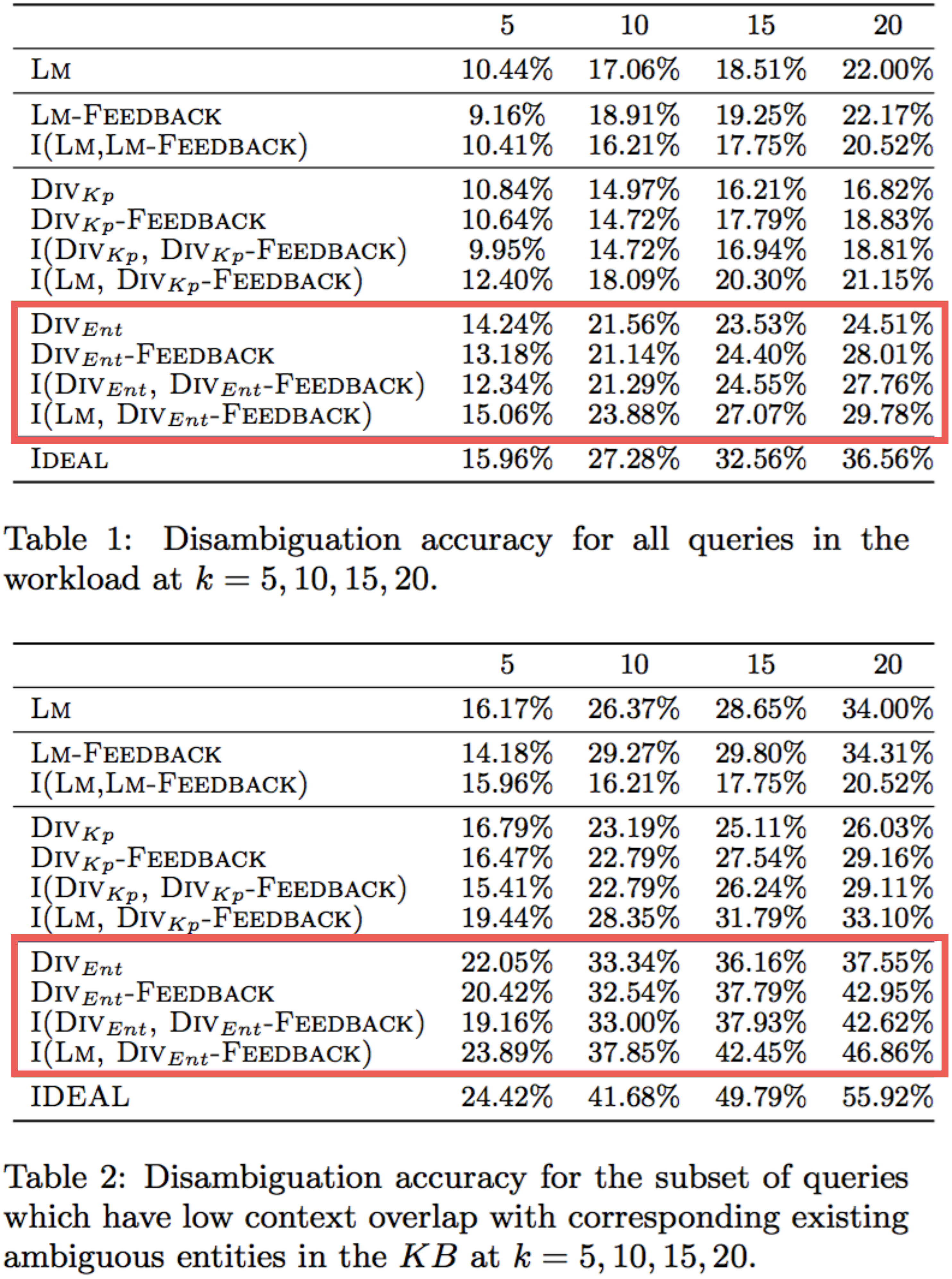
+ = Consequential Document

Experiments

- Document Collection: **Clueweb 09**
- Query Workload: **50** long-tail ambiguous queries
- Ground truth & NED: **AIDA** with YAGO2 (Wikipedia '14)

Query = ambig. mention + 3 keyphrases

- User Simulation:
  - FACC1 - high precision entity linking
    - Document Relevance
  - AIDA - entity descriptions
    - Top 1000 keyphrases based on MI score



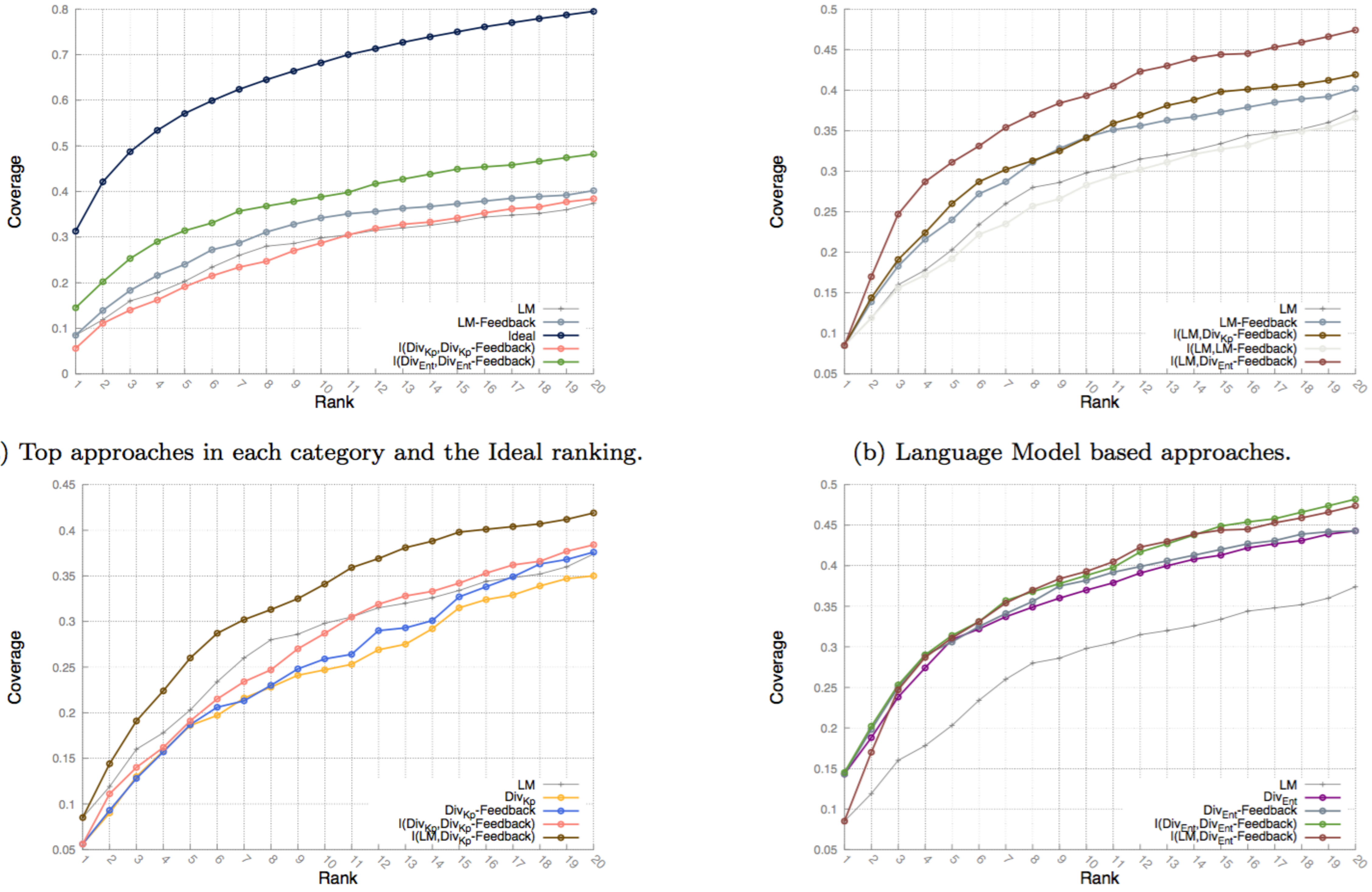
	5	10	15	20
Lm	10.44%	17.06%	18.51%	22.00%
Lm-Feedback	9.16%	18.91%	19.25%	22.17%
I(Lm,Lm-Feedback)	10.41%	16.21%	17.75%	20.52%
Div$_{Kp}$	10.84%	14.97%	16.21%	16.82%
Div$_{Kp}$-Feedback	10.64%	14.72%	17.79%	18.83%
I(Div$_{Kp}$, Div$_{Kp}$-Feedback)	9.95%	14.72%	16.94%	18.81%
I(Lm, Div$_{Kp}$-Feedback)	12.40%	18.09%	20.30%	21.15%
Div$_{Ent}$	14.24%	21.56%	23.53%	24.51%
Div$_{Ent}$-Feedback	13.18%	21.14%	24.40%	28.01%
I(Div$_{Ent}$, Div$_{Ent}$-Feedback)	12.34%	21.29%	24.55%	27.76%
I(Lm, Div$_{Ent}$-Feedback)	15.06%	23.88%	27.07%	29.78%
Ideal	15.96%	27.28%	32.56%	36.56%

Table 1: Disambiguation accuracy for all queries in the workload at $k = 5, 10, 15, 20$.

	5	10	15	20
Lm	16.17%	26.37%	28.65%	34.00%
Lm-Feedback	14.18%	29.27%	29.80%	34.31%
I(Lm,Lm-Feedback)	15.96%	16.21%	17.75%	20.52%
Div$_{Kp}$	16.79%	23.19%	25.11%	26.03%
Div$_{Kp}$-Feedback	16.47%	22.79%	27.54%	29.16%
I(Div$_{Kp}$, Div$_{Kp}$-Feedback)	15.41%	22.79%	26.24%	29.11%
I(Lm, Div$_{Kp}$-Feedback)	19.44%	28.35%	31.79%	33.10%
Div$_{Ent}$	22.05%	33.34%	36.16%	37.55%
Div$_{Ent}$-Feedback	20.42%	32.54%	37.79%	42.95%
I(Div$_{Ent}$, Div$_{Ent}$-Feedback)	19.16%	33.00%	37.93%	42.62%
I(Lm, Div$_{Ent}$-Feedback)	23.89%	37.85%	42.45%	46.86%
IDEAL	24.42%	41.68%	49.79%	55.92%

Table 2: Disambiguation accuracy for the subset of queries which have low context overlap with corresponding existing ambiguous entities in the KB at $k = 5, 10, 15, 20$.

Coverage



(a) Top approaches in each category and the Ideal ranking.

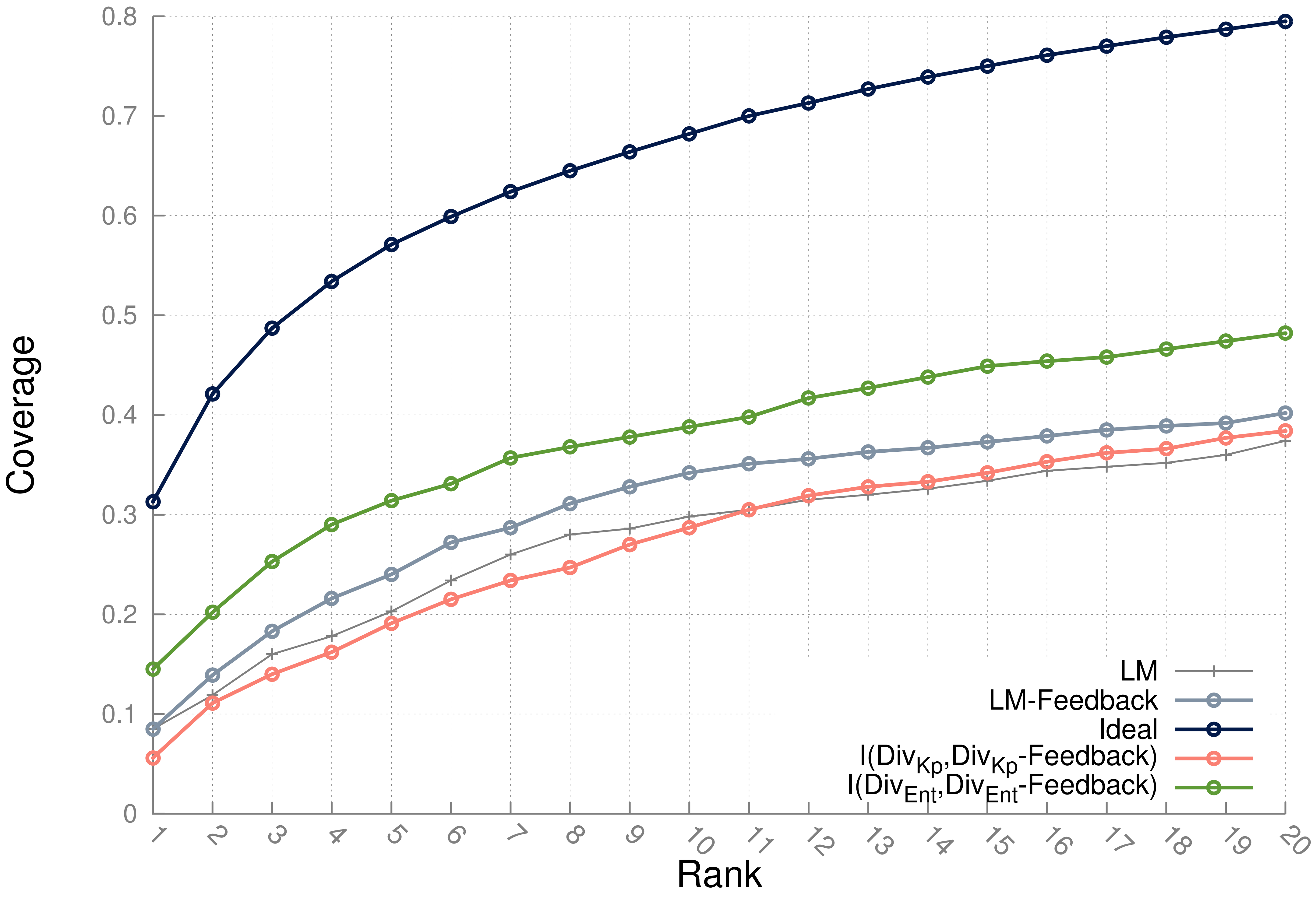
(b) Language Model based approaches.

(c) Keyphrase Diversification based approaches.

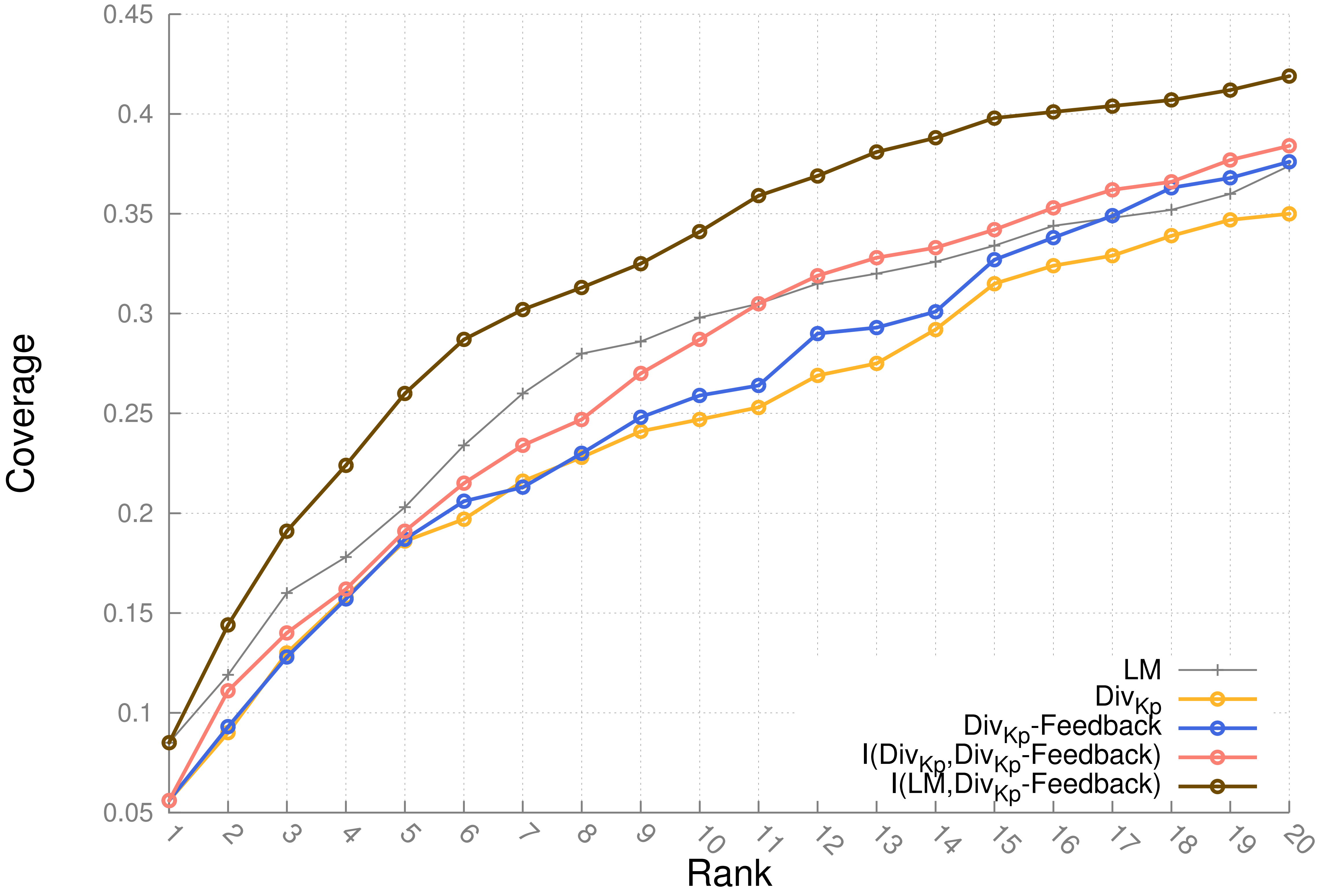
(d) Entity Diversification Approaches.

Figure 5: Keyphrase Coverage vs. Rank: The plots show the fraction of keyphrase coverage against the number of documents the user requests ($k = 1$ to 20).

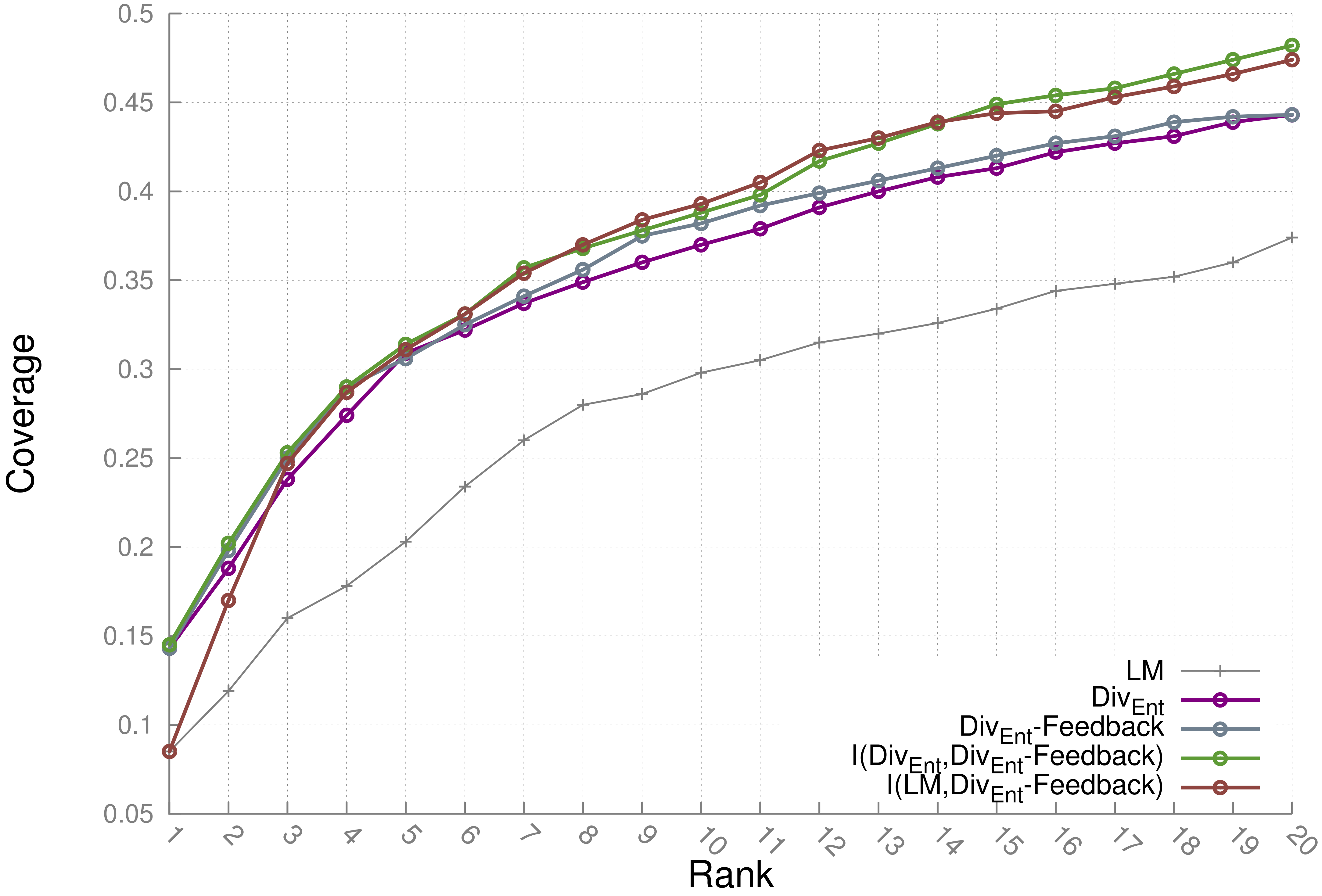
General Trend



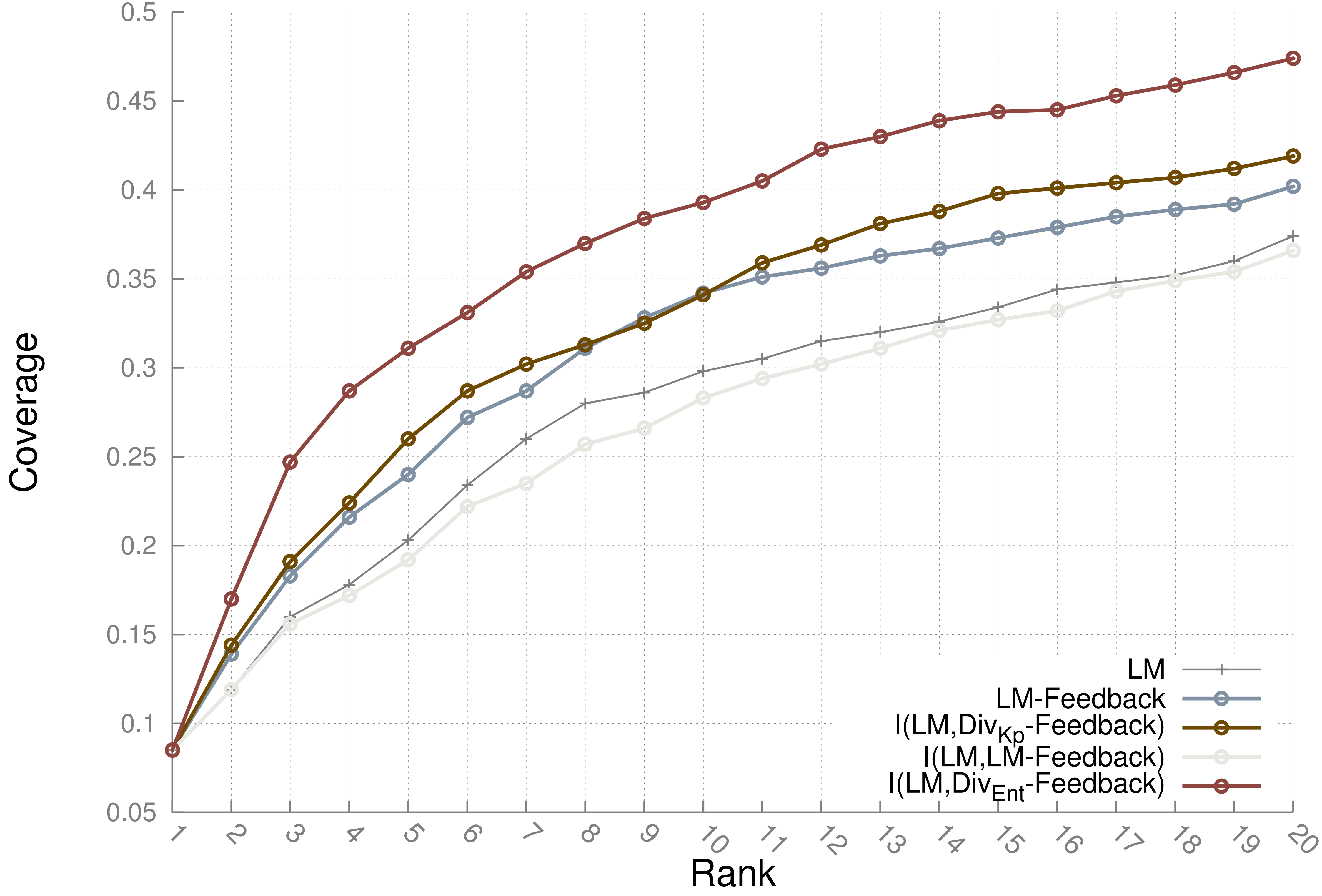
Keyphrase Diversification

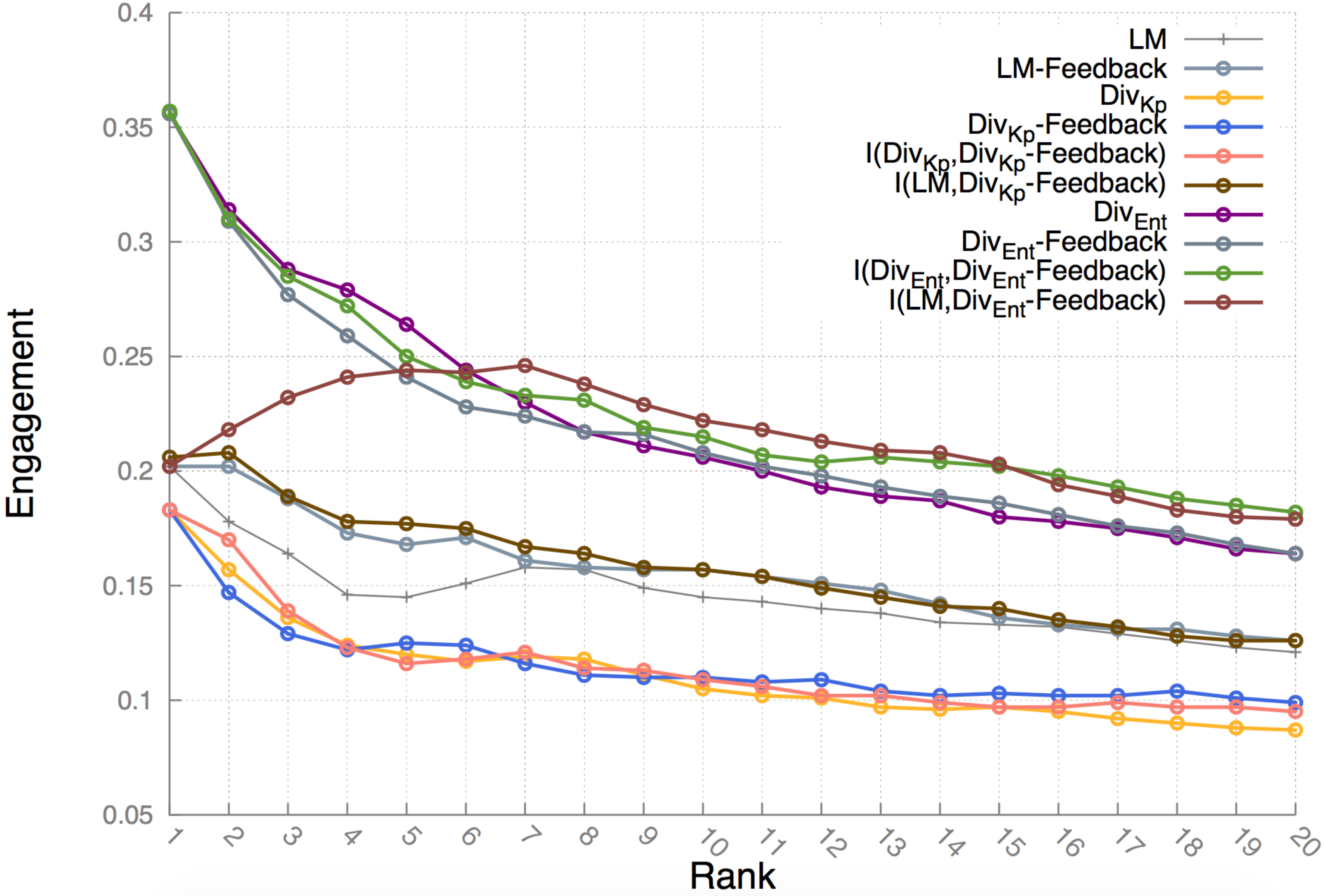


Entity Diversification



Effect of Rank 1





0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
Engagement
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Rank
LM
LM-Feedback
$\mathrm{Div}_{\mathrm{Kp}}$
$\mathrm{Div}_{\mathrm{Kp}}$-Feedback
I($\mathrm{Div}_{\mathrm{Kp}}$,$\mathrm{Div}_{\mathrm{Kp}}$-Feedback)
I(LM,$\mathrm{Div}_{\mathrm{Kp}}$-Feedback)
$\mathrm{Div}_{\mathrm{Ent}}$
$\mathrm{Div}_{\mathrm{Ent}}$-Feedback
I($\mathrm{Div}_{\mathrm{Ent}}$,$\mathrm{Div}_{\mathrm{Ent}}$-Feedback)
I(LM,$\mathrm{Div}_{\mathrm{Ent}}$-Feedback)

Takeaways

- Diversification based approaches are better
    - Diversifying over keyphrases is not good
- User feedback is helpful
- Interleaving helps when two contrasting approaches are used

- **I (LM, DivEnt-Feedback)** achieves the best balance between accuracy, coverage and engagement

Conclusion

- Entity annotations are needed for specialised ranking and mining techniques
- Longtail entities often found in archives are not present in Wikipedia, making NED tools less effective
- NED of ambiguous long tail entities can be tackled with a human in the loop approach