Learning for the Long Tail

Train AI 2018

Marti Hearst
UC Berkeley

With Philippe Laban
How Many Distinct People Were Mentioned In Accounts of the Cambridge Analytical Scandal?

Audience Exercise!
How Many Distinct People Were Mentioned In Accounts of the Cambridge Analytical Scandal?

We find > 700
Why Handle The Long Tail?

How to Handle Rare Cases?
Why Handle Rare Cases?

Avoid Damaging Errors

Don’t erroneously say someone died!
In 2014, John Allen died in a plane downed by a Russian-made missile. John Allen was also a retired U.S. General and Presidential Envoy in 2014. Confusing these people would have been problematic!
John Allens are Plentiful!
Why Do We Care?

newsLens
### World
- Syria Talks
- Trump Lawyer Michael C. 
- Israeli Police
- Salisbury Nerve Agent A.
- Rohingya Crisis
- FBI Director
- Nelson Mandela’s Ex-Wife.
- Bill Cosby Trial
- Saudi-Led Coalition
- Obamacare Repeal
- North Korea
- Junior Hockey Team’s Bus.
- Americas Summit’s Failure
- Brexit Talks
- Trump’s EPA Budget Prop.
- Catalan Independence
- Same-Sex Marriage
- EU Air Traffic Control Ag.
- Deputy National Security
- Solo Alps Ski Trip
- Putin’s Oligarch Allies
- Greek Fighter Jet Crashes
- Former Brazil President L.
- Missouri Governor’s Bid

### Business
- Cambridge Analytica Scandal
- Migrant Crisis
- Co-Pilot In Algeria’s Wors.
- US STOCKS-Wall St
- Trump Trade War
- VW Scandal
- Its Self-Driving Cars
Apr. 2018
Facebook CEO Mark Zuckerberg is about to testify to Congress
Open details

Apr. 06 2018
EU says Facebook confirmed data of 2.7 million Europeans 'improperly shared'
Open details

Apr. 04 2018
Zuckerberg will testify before Congress on April 11
Open details

Mar. 27 2018
Cambridge Analytica whistleblower: Vote Leave ‘cheating’ may have swayed Brexit referendum - video
Open details

Mar. 27 2018
Facebook CEO Mark Zuckerberg will reportedly testify before Congress
Open details

Mar. 27 2018
Zuckerberg will not appear before UK parliament committee
Open details

Mar. 26 2018
FTC Confirms Facebook Probe Over Privacy Practices
Open details
<table>
<thead>
<tr>
<th>Name</th>
<th>Mentions</th>
<th>Roles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steve Bannon</td>
<td>97</td>
<td>strategist, adviser, chief strategist, white house</td>
</tr>
<tr>
<td>Brittany Kaiser</td>
<td>79</td>
<td>director, former director, business development director, former business development director</td>
</tr>
<tr>
<td>Darren Grimes</td>
<td>75</td>
<td>founder, 23-year-old fashion student, friend, fashion student</td>
</tr>
<tr>
<td>Nigel Oakes</td>
<td>73</td>
<td>founder, company, parent company, executive</td>
</tr>
<tr>
<td>Ted Cruz</td>
<td>71</td>
<td>sen., senator, campaign, candidate</td>
</tr>
<tr>
<td>Arron Banks</td>
<td>62</td>
<td>donor, founder, eu founder, campaigner</td>
</tr>
</tbody>
</table>

Arron Banks

- businessman

- Open newsLens profile
- Open Wikidata entry

**Introductions**

- funder
- its
- a key financier in the Brexit campaign
- who has funded the Leave
- British businessman
- A prominent campaigner to leave the European Union
- EU
- the insurance magnate
- who campaigned for Britain to leave the European Union in a 2016 referendum
- EU founder
NEWSLENS SHOWS THE LONG TAIL IN NEWS

• Represent disparate articles as stories
• Can go back in time to see past stories
• Shows many points of view at once
• Shows the less famous people in context
<table>
<thead>
<tr>
<th>Category</th>
<th>Title</th>
<th>Read the story</th>
</tr>
</thead>
<tbody>
<tr>
<td>World</td>
<td>Ebola Outbreak</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Israeli Police</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Syria Talks</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Ukraine Crisis</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Zanu-PF</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Boko Haram Attack</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>South Sudan Conflict</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Zambia Opposition Leader</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Mexico’s 43 Missing Stud.</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Philadelphia Kidnapping ..</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Brexit Talks</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Same-Sex Marriage</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Oil Behind Turkana Fight</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>OK Legal Marijuana</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Free-Swinging Christie T.</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Bin Laden Raid</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Colorado Personhood Am.</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Georgia Senate Race</td>
<td>Read the story</td>
</tr>
<tr>
<td>Business</td>
<td>Virgin Spaceship Crash</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Takata Air Bags</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Pimco Total Return</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Apple’s New Music Strea..</td>
<td>Read the story</td>
</tr>
<tr>
<td>People</td>
<td>AC / DC’s Rudd Has Atte..</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Openly Gay Country Musicist</td>
<td>Read the story</td>
</tr>
<tr>
<td></td>
<td>Disney’s ’ Big Hero 6 ’ Sha..</td>
<td>Read the story</td>
</tr>
</tbody>
</table>
Hawaii Volcano Eruption

May 2018

- Dramatic footage shows lava engulfing car in Hawaii - video
- Dramatic video shows lava bursting from the ground in a volcano eruption in Hawaii
- Fountains of lava: Dramatic video shows scale of Mount Kilauea eruption in Hawaii (PHOTO/VIDEO)
- Aerial footage shows volcanic lava destroying homes in Hawaii - video

Apr. 29 2018

- Hawaii volcano destroys dozens of homes

May 2018

- Thousands evacuated after Hawaii's Kilauea volcano erupts

447 days with no activity

Feb. 05 2017

- 1/4

---

1. resident
2. geologist
3. spokeswoman
4. scientist

Janet Babb • 47 mentions
- geologist
- spokeswoman
- scientist

Darryl Oliveira • 42 mentions
- director
- chief defense director
- chief
- hawaii county

Jeremy Wilson • 17 mentions
-

Jessica Gautier • 16 mentions
- naitor

David Ige • 14 mentions
- governor
- hawaii governor
gov.

Jordan Sonner • 12 mentions
-

Tulsi Gabbard • 12 mentions
- rep.
- u.s. representative
- hawaii
- representative
Why Handle The Long Tail?

How to Handle Rare Cases?

Goal: Entity Linking People to Knowledge Base Entries
Journalists Write Using “Introductions”

A properly sourced news story states some background for anyone who is mentioned.

“Vatican analyst John Allen”

We use an “introduction” as very precise evidence for a person’s identity.
The first introduction below suggests someone who might not be known on wikidata.

“another victim”
“Oscar winning actress”
“the American actress”

The additional introductions enrich the description.
The top U.S. and NATO commander in Afghanistan
Marine Corps Gen.

CNN’s Vatican analyst
The Easier Case
“our much-loved colleague”

An Amsterdam-based British lawyer

one of the last two of the ten UK victims to be identified

(The Telegraph, July 20, 2017)
“our much-loved colleague”

An Amsterdam-based British lawyer

one of the last two of the ten UK victims to be identified

(The Telegraph, July 20, 2017)
Likely to be on Wikidata?

Classifier trained on introductions

Output "no match"

"our much-loved colleague"
The top U.S. and NATO commander in Afghanistan
Marine Corps Gen.
who held talks in Islamabad twice in the last six days
The top allied commander in Afghanistan
Likely to be on Wikidata?

Output “no match”

Likely to be on Wikidata?

Best match

Marine Corps Gen.
Marine Gen.
who held talks in Islamabad twice in the last six days
The top allied commander in Afghanistan

Best match
Likely to be on Wikidata?

Output “no match”

0.99

Found a match

0.05

0.05

0.05
Likely to be on Wikidata?

Output “no match”

“author of a biography of Desmond Tutu”

Nothing above threshold, No match
A CHALLENGING DATASET

• Extracted 1000 (article, people) pairs
• Each person has 2+ potential entries in Wikidata
• Manually labeled the pairs for ground truth
• 15% (293) are instances that do not match to WD
RESULTS

Easy cases: When there is a match in Wikidata
84.5% correct (91.5 using position in Wikidata)
(compared to a popular commercial system of 82.1)

Hard cases: When no match in Wikidata
50.8% correct
(compared to a popular commercial system of 14.5)
Summary: How We Handled the Long Tail

• Identified evidence that is:
  • High precision
  • Contextually determined
  • Can often label with just one instance, but also can be improved when additional evidence is present.

• Two-phase classification:
  • First decide yes-or-no for match to KB
  • This greatly improves accuracy on no-match cases.
Thank you!

newslens.berkeley.edu

Marti Hearst and Phillipe Laban
UC Berkeley