How One Metis Student
Made a Tabloid News Classifier.

Topic modeling clickbait in a world of information overload.

Andrew Tom | June 2, 2017 | awww slack! @andrew.tom

"A fiercely independent press remains fundamental to the stability of our society...Journalists represent the people as much as electeds."

Dan Rather, former host of 60 minutes.

Media Consolidation:
THE ILLUSION OF CHOICE

Media has never been more consolidated. **6 media giants** now control a staggering **90%** of what we read, watch, or listen to.

THESE SIX COMPANIES ARE:

GE	NEWS-CORP	DISNEY	VIACOM	TIME WARNER	CBS
Notable Properties:	Notable Properties:	Notable Properties:	Notable Properties:	Notable Properties:	Notable Properties:
COMCAST	FOX	ABC	MTV	CNN	SHOWTIME
NBC	WALL STREET JOURNAL	ESPN	NICK JR	HBO	SMITHSONIAN CHANNEL
UNIVERSAL PICTURES	NEW YORK POST	PIXAR	BET	TIME	NFL.COM
FOCUS FEATURES		MIRAMAX	CMT	WARNER BROS	JEOPARDY
		MARVEL STUDIOS	PARAMOUNT PICTURES		60 MINUTES

Exclusive To All Newspa~

DIANA WAS
ALIVE HOU
BEFORE SH

A SERIES of previously un-
seen CCTV images have
revealed that Diana was alive
and well before she was
tragically killed in Paris, ten
years ago.

INSIDE ● Do you think Diana is dead? Or have you seen her in
Morocco? Call now: 0900 000 0000. If not, call: 0900 000
0000. Calls cost (cont. p .94)

**BREAKIN
TRUM
TWEET**

THE WHITE HOUSE
2:56pm ET

THE WHITE HOUSE
WASHINGTON

LIVE

THE INTELLIGENCE REPORT

FOX BUSINESS

**WHITE HOUSE: WE HAVE A POROUS
BORDER WITH BROKEN FENCES**

DOW
▶ 20,953.29
▲ 3.40 0.02%
MARKET ALERT

SILVER 16.58 ▼ 0.25 **S&P** 2386.98 ▼ 4.19

NIKKEI	19,402.39	
SHANGHAI	3,159.17	
HANG SENG	23,374.17	

res MSCI Mexico (EWW) 45.35 ▼ 0.11   Whirlpool (WHR) 179.94 ▼ 10.26
Comcast (CMCSA) 73.60 ▲ 0.16   Mattel (MAT) 27.15 ▼ 4.41   Norwegian

CNBC **5:56A PACIFIC**

Journalists are fallible.

Problem:
Illustrate distinctions between
<u>broadsheet and tabloid journalism</u>.

What Can We Measure? The Spectrum of Print Journalism

Broadsheets

The New York Times

Friday, June 2, 2017 | Today's Paper | Video | 55°F | FTSE 100 +0.31% ↑

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Style Food Travel Magazine T Magazine Real Estate ALL

Trump Pulls U.S. Out of Climate Accord, Saying It's a Threat to U.S. Economy

Environment Will Suffer as Result, Allies and Rivals Say

By MICHAEL D. SHEAR

Tabloids

ENQUIRER   MEGYN KELLY   O.J. SIMPSON   GOT A TIP?

GOTCHA!

BILL COSBY BLINDNESS SCAM — BUSTED BY NATIONAL ENQUIRER

STAKEOUT SHOOTS DOWN PERVERT'S CLAIMS AS SEX ATTACK TRIAL LOOMS!

COURT RULING!

JIM CARREY TO FACE TRIAL OVER EX-GIRLFRIEND'S DEATH
Star still haunted by tragic suicide!

FAMILY PHOTOS!

BRAD PITT & SON ATTEND FAMILY THERAPY
Superstar struggling to stay involved!

What Can We Measure? The Spectrum of Print Journalism

Broadsheets

- Disseminate news to the public
- Traditional newsgathering and longer articles
- Focus on 'serious' topics like politics and international news

Tabloids

- Deliver interesting content
- Tend to more racy and sensationalistic
- Focus on 'less serious' topics like culture and celebrity gossip

Unleash the NLP:
Confirm the distinctions between broadsheet and tabloid journalism via **clustering and topic modeling**.

Process

Step 1

Scraped 80,000+ articles
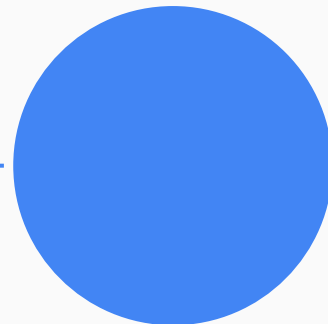(broadsheet and tabloid)
→ TFIDF Vect

Step 2

Multinomial Naive Bayes
Score: 93%
Language is distinct

Step 3

HDP → LDA over each corpus
(20 Topics, 3 Passes):
Topics more descriptive

Broadsheet Corpora
(2005–2015)

Latent topics:
- Economy
- Context
- Politics

Two word clouds demonstrating word frequency across the broadsheet and tabloid corpora.

Top 3 Latent Topics for the Tabloid Corpora (2017)

"Whenever trump communicates, things get sensational."

Latent Topic 2: (21.8%)	0.012*"heat" + 0.012*"feed" + 0.012*"delivered" + 0.011*"reading" + 0.011*"street" + 0.010*"latest" + 0.010*"thanks" + 0.009*"like" + 0.008*"trump" + 0.006*"video"
Latent Topic 13: (10.0%)	0.008*"trump" + 0.003*"president" + 0.003*"said" + 0.003*"news" + 0.003*"information" + 0.002*"video" + 0.002*"people" + 0.002*"great" + 0.002*"right" + 0.002*"donald"
Latent Topic 5: (8.2%)	0.013*"trump" + 0.005*"president" + 0.003*"reading" + 0.003*"did" + 0.003*"feed" + 0.003*"heat" + 0.003*"delivered" + 0.003*"spicer" + 0.003*"twitter" + 0.003*"street"

Latent topics:

- Anytime Trump communicates
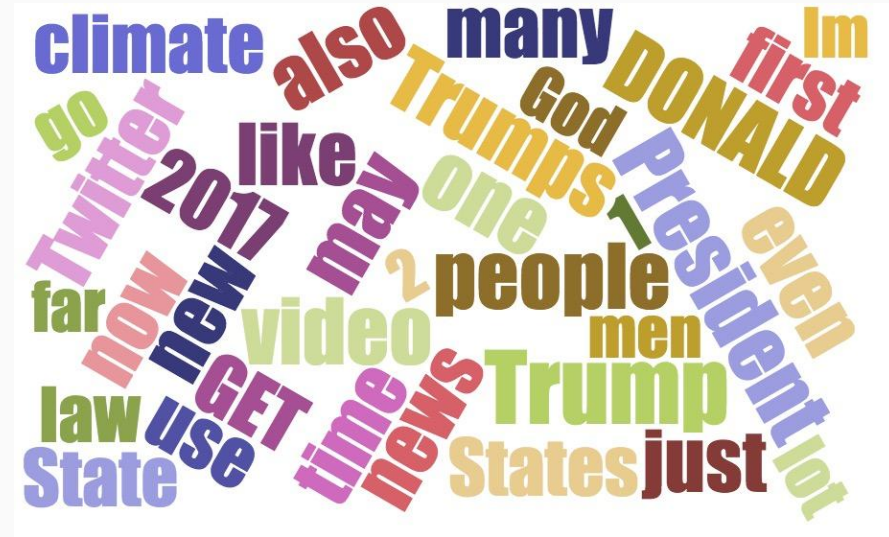
Tabloid Corpora
(2017)

Top 3 Latent Topics for the Broadsheet Corpora (2005-2015)

Latent Topic 8: Economy (13.5%)	0.012*"percent" + 0.011*"year" + 0.009*"said" + 0.008*"billion" + 0.007*"company" + 0.007*"sales" + 0.006*"quarter" + 0.006*"million" + 0.005*"bank" + 0.005*"market"
Latent Topic 15: Issues in Context (7.0%)	0.007*"new" + 0.005*"year" + 0.004*"like" + 0.004*"company" + 0.003*"city" + 0.003*"school" + 0.003*"old" + 0.003*"prime" + 0.003*"san" + 0.003*"people"
Latent Topic 13: Politics (9.1%)	0.015*"president" + 0.012*"obama" + 0.011*"house" + 0.009*"state" + 0.009*"democrats" + 0.008*"republican" + 0.008*"senate" + 0.007*"republicans" + 0.007*"health" + 0.007*"tax"

Broadsheet Corpora
(2005–2015)

Tabloid Corpora
(2017)

Two word clouds demonstrating word frequency across the broadsheet and tabloid corpora.

Future Work

Show where you are in the process and what's left to tackle

New Question
How can we measure deviation from broadsheet journalism?

New Goal
Develop a classifier that can classify an article's text on the spectrum of print journalism.

Classify Print Sources
See where things fall!

Reference a more objective baseline?

Logistic Regression? Random Forest Classifier? Doc2Vec?

INSIDE

INK *The Rise Of The Most Enduring And Trusted Player In Modern Journalism*

PART ONE OF AN EIGHT-PART SERIES

HIGHLIGHTS

...tion Just Prefers Feel ...f Newsprint In Hands

...TIONAL, *Page 14B*

'ONION' PRINT REVENUES UP 5,000%

Demand for digital content "simply isn't there yet," say

Experts: Digital Media Revolutio Still Another 70 Or 80 Years Awa

NEW YORK—Confirming that the current m landscape has so far shown virtually no sign of ...ing away from the lucrative print trade, a group o ...perts told reporters today

What we care about is what we'll get.

Questions?

Appendix

Assumptions & Limitations

- That every article from a given source adheres to the 'seriousness' or 'sensationalist' genre of the publication.
  - Clearly not true
- The corpora covered different time intervals
  - Topic modeling was not comparable
- No intuition for sensitivity along the 'print media' spectrum
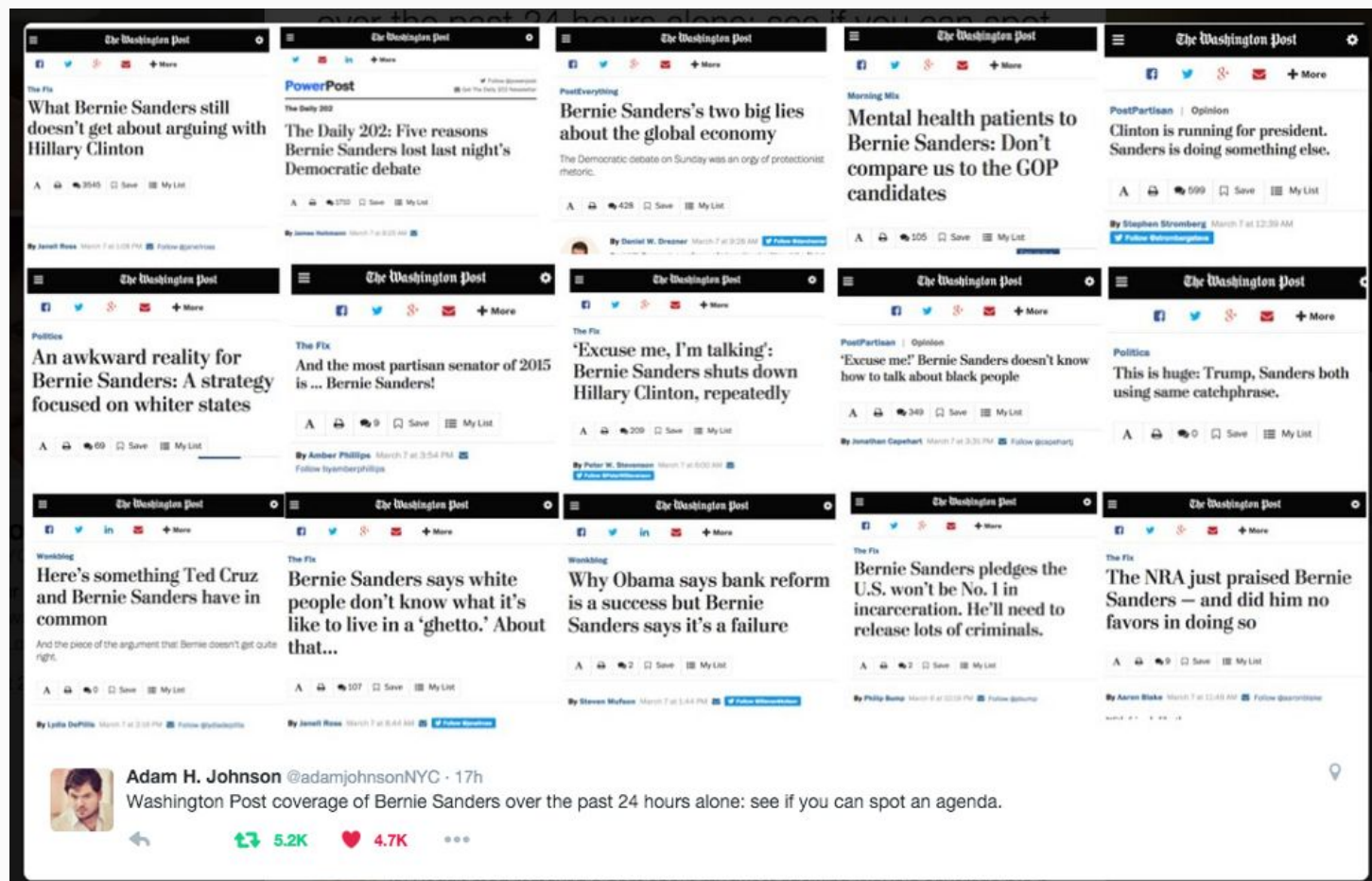
A collection of clickbait-y headlines from the New York Times in 2016. (P.S. I Scraped from 2005-2015)

2009: Michael Jackson had "six months to live."

In January 2009, the Enquirer reported that pop superstar Michael Jackson had only six months to live because of drug and alcohol addiction.

The singer passed away six months later, in June.

An example in 2009 when the National Enquirer reporting was accurate.

Source Selection

Broadsheet Sources

- New York Times
- Reuters (via circular reporting)
- Associated Press (via circular reporting)

Tabloids

- 30+ Sources including:
  - National Enq
  - addictinginfo.org
  - liberalamerica.org
  - politicususa.com
  - viraltube.nl
  - And more from Melissa Zimdars' list.

Melissa Zimdars is an assistant professor of communication and media at Merrimack College.

The Real Process

I can't even word cloud

Yo this HDP is 🔥

LDA FTW

Step 3
LDA over each corpus (20 Topics, 3 Passes): **Topics more descriptive**

Step 1.7: I hate AWS

Step 1
Multinomial Naive Bayes Score: 93%

Step 2
Hierarchical Dirichlet Process over each corpus (150 topics): **Topics not descriptive**

OMG MemoryErrors

FINALLY THIS MAKES SENSE

Wtf is NLP

Did an LSI, or 2

To aggregate or not to aggregate?