Analysis of News Articles With NLP and K-Means Clustering

Angela De La Mora
Regis University

Introduction

This project analyzes **247,568 news articles** downloaded in HTML format from two online newsgroups using web scraping, with permission of newsgroup owner and slowing down the server request process to prevent overloading.

All valuable information was extracted from the article text or the newsgroup post title.

Project main objectives:

- Extract information from articles to build metadata files.
- Use K-Means clustering to group similar articles.

Extracted information:

- Date
- Location
- Article type
- Language
- Publisher
- URL's

Main Python libraries used:

- BeautifulSoup
- Langdetect
- Matplotlib
- NLTK
- Numpy
- Pandas
- Requests
- Scikit-Learn
- Selenium
- Wordcloud

Dataset Profile

Dataset Description: 250K online news articles about LGBT issues, 2000-2019.

Date Range: August 2000 – January 2019

Corpus 1: 116,081 articles (589 MB) **Corpus 2:** 131,487 articles (643 MB)

Unique Articles: 150,846 (613 MB after duplication removal)

Top Languages: English (144,816), Spanish (2,451), French (1,319), German (965), Italian (772), Portuguese (342)

Top Countries: USA (87,253), UK (14,836), Canada (6,857), Australia (4,794), India (3,168), Spain (2,409), New Zealand (1,767), France (1,504)

Sources: The articles were gathered by non-profit LGBT advocacy newsgroups from *Yahoo! Groups* and *Groups.io* and distributed to group members.

---

* The groups had multiple moderators through time. The last moderator, who owns both groups, closed one in 2015 and left the other orphan in 2019. While permission was obtained to conduct this analysis, I will share the sources once I have cleared all potential ethical and legal issues to prevent that the newsgroups are shut down on copyright grounds.

Example of Article

This is an example of what the beginning of a typical article looks like.

There is no order or uniform placement of the metadata within the text body, but most articles contain date, publisher, headline, author, URL, and location.

```
Syracuse.com June 27 2002

http://www.syracuse.com/newsflash/regional/index.ssf?/newsflash/get_st
ory.ssf?/cgi-free/getstory_ssf.cgi?n0937_BC_NY--
SexualOrientation&&news&nystatenews

Prospects dim for sexual orientation non-discrimination act

By JOEL STASHENKO
The Associated Press
6/27/02 12:00 AM

ALBANY, N.Y. (AP) -- Advocates for banning discrimination in New York
state based on sexual orientation say the issue appears all but dead
in 2002.

"We're not holding out much hope," said Joe Grabarz, executive
```

Project Steps

1) Web scrape articles from newsgroups.
2) Create metadata files for each corpus containing id, corpus, date, and title.
3) Identify and remove duplicated articles.
4) Clean articles text (i.e., remove newsgroup signature, advertising, etc.)
5) Extract and clean a mini-corpus (Corpus 3) contained in Corpus 1.
6) Combine all metadata files into a single file.
7) Normalize text (lowercase) and remove stopwords.
8) Perform stemming and lemmatization.
9) Extract information about place and article type from title field in metadata file.
10) Get state name if extracted location is the USA.
11) Extract headlines from newsgroup titles.
12) Detect article language.
13) Extract URLs from article text.
14) Extract publisher and/or copyright information from article text.
15) Detect URLs that are still active online as originally referenced.
16) Perform descriptive statistics and create visualizations.
17) Perform K-Means clustering using titles.
18) Perform K-Means clustering using the normalized lemmatized articles.
19) Enjoy a cup of coffee!

1: Web Scrape Articles

Web-scraped using Python with permission of the groups' owner from:

Yahoo! Groups (Corpus 1) *Groups IO* (Corpus 2)

- Care was placed to not overload the platform's servers by delaying each page request.
- The bulk of the download was conducted at the end of 2017 over a period of two months.
- Remaining articles were downloaded in 2018 and 2019.
- Information gathered includes article body, and date and title as published on newsgroups.
- Article images downloaded for archival but not used for this project.
- HTML removed with BeautifulSoup and copies were saved as plain-text.

2: Create Metadata Files

As articles were downloaded, a metadata file was created for each corpus containing Post ID, timestamp, date posted to the newsgroup, and title from newsgroup.

An example of what these starter metadata files look like:

```
POSTID,TIMESTAMP,DATE,TITLE
000001,965700480,20000807,FW: Human Rights Act Influence Felt at Court [LawZONE]
000002,965700660,20000807,"FW: April Ashley: ""...I'm still treated as a joke..."" [The Express]"
000003,965700840,20000807,FW: Anti-gay Scouts lose support... [Sun-Sentinal - South Florida]
000004,965700840,20000807,FW: Same sex unions being considered in 7 states... [The Bermuda Sun - Faith]
000005,965700960,20000807,FW: Gay sex law setback for Labour... [The Independent]
000006,965700960,20000807,FW: Sex secret of husband... [THE NEWS OF THE WORLD]
000007,965700960,20000807,FW: Western Methodists preach open-mindedness... [Denver Post]
```

After project completion the final metadata file looks like this:

```
CORPUS,POSTID,TIMESTAMP,DATE,TITLE,LOCATION,TYPE,TITLE_CLEAN,LANG,URLS,SOURCES
corpus_1,000001,965700480,20000807,FW: Human Rights Act Influence Felt at Court [LawZONE],,,Human Rights Act Influence Felt at Court,en,http://www.lawzone.co.uk/cgi-bin/item.cgi?id=22982&d=101&h=175&f=173&dateformat=%o%20%B%20%Y | h
corpus_1,000002,965700660,20000807,"FW: April Ashley: ""...I'm still treated as a joke..."" [The Express]",,,"April Ashley: "" I'm still treated as a joke """,en,http://www.image100.com.She | http://www.lineone.net/express/00/07/28/
corpus_1,000003,965700840,20000807,FW: Anti-gay Scouts lose support... [Sun-Sentinal - South Florida],,,gay Scouts lose support,en,"http://www.sun-sentinel.com/news/daily/detail/0,1136,3250000000113111,00.html","1999, Sun-Sentinel
corpus_1,000004,965700840,20000807,FW: Same sex unions being considered in 7 states... [The Bermuda Sun - Faith],USA|BERMUDA,,Same sex unions being considered in 7 states,en,http://www.bermudasun.bm/cgi-local/edpull.pl?cat=14Faith&o
corpus_1,000005,965700960,20000807,FW: Gay sex law setback for Labour... [The Independent],,LEGAL,Gay sex law setback for Labour,en,http://www.independent.co.uk/news/UK/Politics/2000-07/sexlaw300700..html,2000 Independent Digital (U
corpus_1,000006,965700960,20000807,FW: Sex secret of husband... [THE NEWS OF THE WORLD],,INTERNATIONAL|NEWS,Sex secret of husband,en,,
corpus_1,000007,965700960,20000807,FW: Western Methodists preach open-mindedness... [Denver Post],"CO,USA",,Western Methodists preach open-mindedness,en,http://www.denverpost.com:80/news/news0729f.htm,
```

Fields added to the final metadata file include corpus, post ID, timestamp, date posted to newsgroup, post title given by newsgroup moderator, location, article type, cleaned title (proxy for headline), language, referenced URLs, and publisher.

3: Remove Duplicate Articles

The newsgroups were initially independent from one another but contained references to the same news articles. They were later combined under one owner and one group was closed in 2015 after years of duplication.

To identify article duplication, the following strategy was used:

- Compare titles in Corpus 2 against Corpus 1 within a date range of ±2 days and remove article from Corpus 1 if identical titles are found.

- If no title match found, compare articles in Corpus 2 against Corpus 1 within a ±2 day range using Cosine Similarity and remove article from Corpus 1 if duplicate found.

- The Cosine Similarity threshold was set at 0.94 after a few trial and error iterations to balance the number of false positives and negatives.

- Duplicate detection is performed with normalized text, but without removing stopwords.

- If two texts are identical, their Cosine Similarity will be 1.

- Two documents that have nothing in common will have a Cosine Similarity of 0.

4: Clean Articles Text

Many articles include signature text added at the bottom of the text body by newsgroup moderators and organizations, or as platform advertising. For example:

```
cut_points = ['Kindly appreciate that Brenda',
              "Moderator's Note: Thanks to Brenda",
              'UKPFC-NEWS is operated by Press for Change',
              'This message comes to you from Press for Change',
              'Download Yahoo! Messenger now',
              '---------------------- Yahoo! Groups Sponsor',
              'Do You Yahoo',
              'Gesendet von Yahoo',
              'Chat with friends online'
              ]
```

These strings of text are used as cut points to slice the article, saving the text before the slice point and removing everything after it:

BEFORE:

```
The MGRM will continue its dialogue with EU officials, as well as
with the European institutions, to ensure that the provisions of the
Directive are properly implemented in the local sphere prior to
Malta's accession to the EU.

Edgar Sammut
Media Officer

01.V.2002

For more information or comments kindly contact MGRM©-s International
Secretary, Mr Christian Attard, on 99.85.85.16.

by Edgar Sammut
--

*********************************************************************
"Whoever told them that the truth shall set them free
was obviously and grossly unfamiliar with federal law."
-- John Ashcroft
*********************************************************************

_____
Do You Yahoo!?
Get your free @yahoo.com address at http://mail.yahoo.com
```

AFTER:

```
The MGRM will continue its dialogue with EU officials, as well as
with the European institutions, to ensure that the provisions of the
Directive are properly implemented in the local sphere prior to
Malta's accession to the EU.

Edgar Sammut
Media Officer

01.V.2002

For more information or comments kindly contact MGRM©-s International
Secretary, Mr Christian Attard, on 99.85.85.16.

by Edgar Sammut
--

*********************************************************************
"Whoever told them that the truth shall set them free
was obviously and grossly unfamiliar with federal law."
-- John Ashcroft
*********************************************************************

_____
```

5: Extract Extra Corpus

- A third mini-corpus was found inside of Corpus 1.
- Instead of one article per post, some posts had multiple articles in daily digest form.
- These articles were identified by keywords referencing the word “digest”.
- Some articles required manual cleaning because of their low state of organization.
- To create the third corpus, 437 individual articles were split from 62 digest-style posts.
- Corpus 3 was compared against Corpus 1 and Corpus 2 for duplication removal.
- Final unique article count from Corpus 3 is 346 articles.

6: Combine Metadata Files

After text cleaning and duplication removal, metadata files were combined into a single file.

View of final Pandas dataframe containing all metadata information after project completion:

	CORPUS	POSTID	TIMESTAMP	DATE	TITLE	LOCATION	TYPE
0	corpus_1	000001	965700480	20000807	FW: Human Rights Act Influence Felt at Court [...		
1	corpus_1	000002	965700660	20000807	FW: April Ashley: "...I'm still treated as a j...		
2	corpus_1	000003	965700840	20000807	FW: Anti-gay Scouts lose support... [Sun-Senti...		
3	corpus_1	000004	965700840	20000807	FW: Same sex unions being considered in 7 stat...	USA|BERMUDA	
4	corpus_1	000005	965700960	20000807	FW: Gay sex law setback for Labour... [The Ind...		LEGAL
5	corpus_1	000006	965700960	20000807	FW: Sex secret of husband... [THE NEWS OF THE ...		INTERNATIONAL|NEWS

/

	TITLE_CLEAN	LANG	URLS	SOURCES
Human Rights Act Influence Felt at Court	en	http://www.lawzone.co.uk/cgi-bin/item.cgi?id=2...	2000 Sift plc
April Ashley: " I'm still treated as a joke "	en	http://www.image100.com.She | http://www.lineo...	Express Newspapers, 2000
gay Scouts lose support	en	http://www.sun-sentinel.com/news/daily/detail/...	1999, Sun-Sentinel Co. & South Florida Interac...
Same sex unions being considered in 7 states	en	http://www.bermudasun.bm/cgi-local/edpull.pl?c...	2000
Gay sex law setback for Labour	en	http://www.independent.co.uk/news/UK/Politics/...	2000 Independent Digital (UK) Ltd
Sex secret of husband	en		

7: Normalize / Remove Stopwords

- A copy of each corpus is created and all punctuation and numbers removed.
- Text converted to lowercase.
- Articles are tokenized during the process (each word is a token).
- Stopwords corpus imported from NLTK, then stopwords removed from articles.

This is London -
http://www.thisislondon.co.uk/dynamic/news/top_story.html?in_review_id=31005
6&in_review_text_id=253999
Tuesday, August 22, 2000

Pc suspended over 'abusing gay officer'
by Lucy Lawrence

A police constable specialising in investigating hate crimes has been suspended over alleged homophobic abuse of a colleague.

The officer, who is believed to work for the Hammersmith and Fulham Community Safety Unit, is alleged to have made abusive and homophobic remarks to a gay constable.

The suspension of the officer, who is now being investigated by the Met's central area complaints unit, will embarrass senior officers who set up CSUs only last year to focus on homophobic crime, race crime and domestic violence. Former Commissioner Sir Paul Condon launched the CSUs in every borough, saying: "The CSUs are an indication of the priority the Met is placing on tackling hate crime. Officers working within these units are to be more accountable than ever before."

The gay officer involved in the incident is being supported by the Lesbian and Gay Police Association. A spokesman said: "It is outrageous that these things go on. It means we really have to be vigorous to make sure we are selecting the right people to do the job and that the training the officers receive is robust enough not just to impart the new information but also to weed out people who are not suitable." He added the prompt action to suspend the officer showed that Scotland Yard is taking this type of allegation seriously, although several gay officers a week are calling the association's helpline to report homophobic abuse.

Earlier this year, two constables in Greenwich were suspended for an alleged campaign of persecution against a gay colleague.

A Scotland Yard spokeswoman said of the latest allegations: "An officer was suspended from duty on 18 August in connection with an investigation by the central area complaints unit into alleged remarks made between two officers on 14 August."

← BEFORE

AFTER:

london thisislondon co uk dynamic news top story review id review text id tuesday august pc suspended abusing gay officer lucy lawrence police constable specialising investigating hate crimes suspended alleged homophobic abuse colleague officer believed work hammersmith fulham community safety unit alleged made abusive homophobic remarks gay constable suspension officer investigated met central area complaints unit embarrass senior officers set csus last year focus homophobic crime race crime domestic violence former commissioner sir paul condon launched csus every borough saying csus indication priority met placing tackling hate crime officers working within units accountable ever gay officer involved incident supported lesbian gay police association spokesman said outrageous things go means really vigorous make sure selecting right people job training officers receive robust enough impart new information also weed people suitable added prompt action suspend officer showed scotland yard taking type allegation seriously although several gay officers week calling association helpline report homophobic abuse earlier year two constables greenwich suspended alleged campaign persecution gay colleague scotland yard spokeswoman said latest allegations officer suspended duty august connection investigation central area complaints unit alleged remarks made two officers august associated newspapers ltd august

8: Stemming and Lemmatization

PorterStemmer and WordNetLemmatizer imported from NLTK.

ORIGINAL:

```
This is London -
http://www.thisislondon.co.uk/dynamic/news/top_story.html?in_review_id=
31005
6&in_review_text_id=253999
Tuesday, August 22, 2000


Pc suspended over 'abusing gay officer'
by Lucy Lawrence

A police constable specialising in investigating hate crimes has been
suspended over alleged homophobic abuse of a colleague.

The officer, who is believed to work for the Hammersmith and Fulham
Community Safety Unit, is alleged to have made abusive and homophobic
remarks to a gay constable.

The suspension of the officer, who is now being investigated by the
Met's
central area complaints unit, will embarrass senior officers who set up
CSUs
only last year to focus on homophobic crime, race crime and domestic
violence. Former Commissioner Sir Paul Condon launched the CSUs in every
borough, saying: "The CSUs are an indication of the priority the Met is
placing on tackling hate crime. Officers working within these units are
to
be more accountable than ever before."

The gay officer involved in the incident is being supported by the
Lesbian
and Gay Police Association. A spokesman said: "It is outrageous that
these
things go on. It means we really have to be vigorous to make sure we are
selecting the right people to do the job and that the training the
officers
receive is robust enough not just to impart the new information but also
to
weed out people who are not suitable." He added the prompt action to
suspend
the officer showed that Scotland Yard is taking this type of allegation
seriously, although several gay officers a week are calling the
association's helpline to report homophobic abuse.

Earlier this year, two constables in Greenwich were suspended for an
alleged
campaign of persecution against a gay colleague.

A Scotland Yard spokeswoman said of the latest allegations: "An officer
was
suspended from duty on 18 August in connection with an investigation by
the
central area complaints unit into alleged remarks made between two
officers
on 14 August."

© Associated Newspapers Ltd., 22 August 2000
```

STEMMING:

```
london thisislondon co uk dynam news top stori review id review text id
tuesday august pc suspend abus gay offic luci lawrenc polic constabl
specialis investig hate crime suspend alleg homophob abus colleagu offic
believ work hammersmith fulham commun safeti unit alleg made abus
homophob remark gay constabl suspens offic investig met central area
complaint unit embarrass senior offic set csu last year focu homophob
crime race crime domest violenc former commission sir paul condon launch
csu everi borough say csu indic prioriti met place tackl hate crime
offic work within unit account ever gay offic involv incid support
lesbian gay polic associ spokesman said outrag thing go mean realli
vigor make sure select right peopl job train offic receiv robust enough
impart new inform also weed peopl suitabl ad prompt action suspend offic
show scotland yard take type alleg serious although sever gay offic week
call associ helplin report homophob abus earlier year two constabl
greenwich suspend alleg campaign persecut gay colleagu scotland yard
spokeswoman said latest alleg offic suspend duti august connect investig
central area complaint unit alleg remark made two offic august associ
newspap ltd august
```

LEMMATIZATION:

```
london thisislondon co uk dynamic news top story review id review text
id tuesday august pc suspended abusing gay officer lucy lawrence police
constable specialising investigating hate crime suspended alleged
homophobic abuse colleague officer believed work hammersmith fulham
community safety unit alleged made abusive homophobic remark gay
constable suspension officer investigated met central area complaint
unit embarrass senior officer set csus last year focus homophobic crime
race crime domestic violence former commissioner sir paul condon
launched csus every borough saying csus indication priority met placing
tackling hate crime officer working within unit accountable ever gay
officer involved incident supported lesbian gay police association
spokesman said outrageous thing go mean really vigorous make sure
selecting right people job training officer receive robust enough impart
new information also weed people suitable added prompt action suspend
officer showed scotland yard taking type allegation seriously although
several gay officer week calling association helpline report homophobic
abuse earlier year two constable greenwich suspended alleged campaign
persecution gay colleague scotland yard spokeswoman said latest
allegation officer suspended duty august connection investigation
central area complaint unit alleged remark made two officer august
associated newspaper ltd august
```

9: Place and Article Type

Newsgroup moderators added information to post titles about location and/or article type. While there was variation across years and individual moderators, this information was frequently added within square brackets:

```
[People/Entertainment] [Australia] How the other half lives,AUSTRALIA,,How the other half lives,en,http://www.n
"[Blog/Entertainment] [UK] Cabaret in East London, With a Twist",UK,,"Cabaret in East London, With a Twist",en,
"[News] [UT, USA] Forum discusses LGBT meanings and modes","UT,USA",NEWS,Forum discusses LGBT meanings and mode
"[People/Religion] [FL,USA] The Rev. Rebecca Steen, First Congregational Church, Fort Lauderdale",,,"The Rev. R
"[News/Entertainment] [NY, USA] Cho amps up the raunch -- delightfully so",USA,,Cho amps up the raunch -- delig
"[News/People/Arts] [CA, USA] Under the cover of Original Plumbing's top model and the magazine's release",USA,
[Blog/News] [USA] Bilerico Radio: Fred Karger interview and ENDA roundtable,USA,BLOG|INTERVIEW,Bilerico Radio:
"[News/People] [KS, USA] Bisexuals face additional challenges",USA,,Bisexuals face additional challenges,en,htt
"[News] [TX, USA] Regional roundup (UNT to offer unisex bathrooms on campus)","TX,USA",NEWS,Regional roundup (U
[Blog/Commentary] [USA] Sorry I Haven't Written,USA,,Sorry I Haven't Written,en,http://www.edgeboston.com/inde
"[Commentary] [USA] Gay activist: ENDA may lead to homosexual outing, quotas in the workplace",USA,COMMENTARY,"
```

After stopwords, numbers and special characters were removed from titles, all remaining unique words were manually normalized using two JSON files (one for place names, the other one for article types) to catch misspellings, variations and abbreviations:

```
"AU": "AUSTRALIA",
"AUATRALIA": "AUSTRALIA",
"AUS": "AUSTRALIA",
"AUSTRAILIA": "AUSTRALIA",
"AUSTRAKIA": "AUSTRALIA",
"AUSTRALIA": "AUSTRALIA",
"AUSTRALIE": "AUSTRALIA",
"AUSTRALIEN": "AUSTRALIA",

"PHILIPPINES": "PHILIPPINES",
"PHILLIPINES": "PHILIPPINES",
"PHILLPPINES": "PHILIPPINES",
```

```
"ENNTERTAINMENT": "ENTERTAINMENT",
"ENTERAINMENT": "ENTERTAINMENT",
"ENTERTAIMENT": "ENTERTAINMENT",
"ENTERTAINMANT": "ENTERTAINMENT",
"ENTERTAINMENT": "ENTERTAINMENT",
"ENTERTAINNMENT": "ENTERTAINMENT",

    "BRETAGNA": "UK",
    "BRETAGNE": "UK",
    "BRITAIN": "UK",
    "BRITANIQUE": "UK",
    "BRITISH": "UK",
```

```
"COMENTARY": "COMMENTARY",
"COMM": "COMMENTARY",
"COMMENARY": "COMMENTARY",
"COMMENATRAY": "COMMENTARY",
"COMMENATRY": "COMMENTARY",
"COMMENETARY": "COMMENTARY",
"COMMENTARTY": "COMMENTARY",
"COMMENTARY": "COMMENTARY",
"COMMENTARYT": "COMMENTARY",
"COMMENTARYUSA": "COMMENTARY",
"COMMENTS": "COMMENTARY",
"COMMMENTARY": "COMMENTARY",
```

Titles were then matched against the dictionary and if a key related to place or article type was found, its value was added to the corresponding metadata column.

10: Extract States (USA Only)

Titles related to the USA frequently include information about the state:

```
."[News] [MO, USA] Belton man pleads not guilty in fatal shooting allegedly connected to
."[News] [MO, USA] Dad 'shot wife after cross-dressing row'",USA,NEWS,Dad 'shot wife afte
.[Blog/Commentary] [USA] Dressed as a Woman But Your ID Says You're a Man? You Can't Ente
."[News] [IL, USA] Chicago gay bar insists patron match the gender on their ID","IL,USA",
."[News] [IL, USA] Gay Bar's New ID Policy Is a Drag",USA,POLICY|NEWS,Gay Bar's New ID Po
."[News] [MO, USA] Exhibit shows transgender photos, stories",USA,NEWS|PHOTOGRAPHY,"Exhib
."[News] [FL, USA] Autopsy finds low level of medication in transgender man hit, killed b
."[News] [OH, USA] Ohio House Passes Equal Rights Legislation",USA,LEGAL|NEWS,Ohio House
."[Blog/News] [USA] Criminal Charges: Card crime committed by voyeurs, transvestites and
."[News/Events] [TN, USA] TTPC fundraising dinner set for Oct. 17","TN,USA",NEWS,TTPC fun
```

To tag articles by state, a JSON file was created with all US state names and abbreviations:

```
"AK": "ALASKA",
"AL": "ALABAMA",
"AR": "ARKANSAS",
"AZ": "ARIZONA",
"CA": "CALIFORNIA",
"CO": "COLORADO",
"CT": "CONNECTICUT",
```

Keys and values were matched against the title:

```
corpus_1,008439,1079067600,20040312,UK: Hansard re Commons Standing Committee on the Gerbil,UK,,Hansard Commons Standing Comm
corpus_1,008442,1079067600,20040312,US - US Cracks Down on 'Andro' Performance Supplement... [Reuters - Mar 12/04],USA,,US Cr
corpus_1,008443,1079067600,20040312,Britain - Gender Recognition Bill - Members of House of Commons Standing Committe 'A...'
corpus_1,008446,1079067600,20040312,Britain - Thomas Boyd/Debbie Gould - SEX-SWAP CLEANER IN LONDON HOTEL SUICIDE... [The Dai
corpus_1,008447,1079067600,20040312,Britain - Gender Recognition Bill - House of Commons Standing Committee A Debate (Morning
corpus_1,008451,1079154000,20040313,US - Philanthropist's group receives global honor for work onGLBT rights... [The Rocky Mo
corpus_1,008452,1079154000,20040313,Switzerland - NGOs Worldwide Support Brazil's Historic UN SexualOrientation Resolution...
corpus_1,008453,1079240400,20040314,US: Denver Protects Trans Students,"CO,USA",,Denver Protects Trans Students,en,http://www
```

If a match was found, the state abbreviation + “USA” were added to the metadata file:

TITLE	LOCATION	TYPE	TITLE_CLEAN
s TG - Ohi...	OH,USA		Hetero couple denied marriage: one is TG - Ohi...
to male jail	OH,USA		'Female' inmate on way to male jail
nsgender E...	UK	INTERNATIONAL|ANNOUNCEMENTS|EVENT	Announcement: 2003 International Transgender E...
e and More...		FILM	Being a Kid is a Drag-on-Screen: Moand MoGende...
nination Bill	MD,USA	ACTION	Call to Action: BaltimoDiscrimination Bill
ion in Lou...	LA,USA	RELEASES|CASE	Press Release: ACLU Criticizes Decision in Lou...
pestry Jou...	GA,USA	EDUCATION|SUBMISSIONS	Call for Submissions: Transgender Tapestry Jou...
nination Bill	MD,USA	ACTION	Call to Action: BaltimoDiscrimination Bill

11: Extract Headlines

To extract headlines, titles were cleaned by removing place, article type and other data. This was a difficult step because many variations are found in the structure of titles, for example:

```
,France: LUDWIG TROVATO FILM MAKER AND WRITER,FRANCE,FILM,France: LUDWIG TROVATO FILM MAKER AND WRITER,en,http://www.exaequoreims.com/news.php?id
,CA: Transsexual Day of Pride,,,CA: Transsexual Day of Pride,en,http://www.ctffr.org/id33.html | http://www.ctffr.org/,
,UK: Further progress of the GerBil in committee,UK,,Further progress of the GerBil in committee,en,http://www.parliamentlive.tv/,
,China - 1st Chinese transsexual gets marriage certificate... [Xinhua News Agency: Shenzhen Daily - Mar 17/04],CHINA,NEWS,1st Chinese transsexual
,UK: Hansard re Commons Standing Committee on the Gerbil,UK,,Hansard Commons Standing Committee on the Gerbil,en,http://www.publications.parliame
,"US,OH: Greene v. Bowles opinion",,OPINION,"US,OH: Greene v. Bowles opinion",en,http://pacer.ca6.uscourts.gov/cgi-bin/getopn.pl?OPINION=a0078p.0
,China - China allows transsexual to marry... [redif.com - Mar 17/04],CHINA,,China allows transsexual to marry,en,http://in.rediff.com/news/2004/
,India - Hijra Neelam - Three charged with eunuch 's 2003 killing... [The Indian Express - Mar 17/04],INDIA,,Hijra Neelam - Three charged with eu
,Gender Recognition Bill - Committee Stage - Hansard,,,Gender Recognition Bill - Committee Stage - Hansard,en,http://www.publications.parliament.
,Scotland: Equality Network news - Your Scotland project,SCOTLAND,NEWS,Scotland: Equality Network news - Your Scotland project,en,http://www.equa
,US - Charlie could have what amounts to his sex-change operation as soon as Monday... [The Greeley Tribune - Mar 17/04],USA,,Charlie could have
,"US - For Children of Gays, Marriage Brings Joy... [The New York Times - Mar 19/04]","NY,USA",NEWS,"For Children of Gays, Marriage Brings Joy",e
```

```
US: Health out of the closet,USA,HEALTH,Health out of the closet,en,http://www.usnews.com/usnews/issue/040913/health/13med.htm,
"US,CA: County supervisors divided over health care for indigent residents",,HEALTH,"US,CA: County supervisors divided over health
[Film Review] Stage Beauty,,FILM|REVIEWS,Stage Beauty,en,http://www.telegraph.co.uk/arts/main..html?xml=/arts/2004/09/03/bfhell03.xm
[Humor] UK: THE COLOURFUL LIFE OF CAR RENTAL,UK,,THE COLOURFUL LIFE OF CAR RENTAL,en,http://www.responsesource.com/releases/rel_disp
[Film Review] Stage Beauty,,FILM|REVIEWS,Stage Beauty,en,http://iccoventry.icnetwork.co.uk/0800whatson/entertainment/tm_objectid=14
"India: Dance, music & claps at Bhopal love parade",INDIA,MUSIC,"India: Dance, music & claps at Bhopal love parade",en,http://www.te
[Opinion] Michael Moore: Why Democrats Shouldn't Be Scared,,OPINION,Michael Moo Why Democrats Shouldn't Be Scared,en,,
[News] [California] Transgender People -- a Study in Courage,,PEOPLE|STUDIES|NEWS,a Study in Courage,en,"http://www.latimes.com/news
[Opinion] How Kerry Became a Girlie-Man,,OPINION,How Kerry Became a Girlie-Man,en,http://www.nytimes.com/2004/09/05/arts/05RICH.html
[NEWS] A potentially expensive proposition,,NEWS,A potentially expensive proposition,en,http://www.pridesource.com/article..html?art
```

```
,FW: 'A private family life' [The Telegraph - leaders],,,'A private family life',en,http://www.dailytelegraph.co.uk:80/dt?ac=001840572447371&rtmo=Qx03Sa0R
,"FW: Catholic priesthood is becoming mainly gay, US theologianclaims... [The Guardian]",USA,,"Catholic priesthood is becoming mainly gay, US theologiancla
,FW: Politicians tested in equality search... [The Irish Times ],IRELAND,,Politicians tested in equality search,en,http://www.ireland.com:80/newspaper/irel
,FW: Homosexuals terrorized during apartheid era... [Earth Times NewsService],,,Homosexuals terrorized during apartheid era,en,"http://www.earthtimes.org:
,FW: North Yorkshire Police Get Gay Helpline... [Sunday Sun],,,North YorkshiPolice Get Gay Helpline,en,"http://www.sundaysun.co.uk/cfm/body_news_story.cfm
,FW: Officials go 'on message' to answer the [HRA] critics... [ TheTelegraph],,,Officials go 'on message' to answer the critics,en,http://www.telegraph.co.
,FW: RTE rejects claim on gay content... [The Irish Times - IRELAND],IRELAND,,RTE rejects claim on gay content,en,http://www.ireland.com/newspaper/ireland/
,FW: Female Genital Mutilation: A Guide to Laws and PoliciesWorldwide... [The Earth Times/BOOK REVIEW],,LEGAL|REVIEWS,Female Genital Mutilation: A Guide to
,FW: Group sex for gays to be legal after Euro ruling... [The Times],,GROUPS|LEGAL,Group sex for gays to be legal after Euro ruling,en,http://www.the-times
,FW: Britain's century-old sex laws were branded unfair today... [London Evening News],UK,LEGAL|NEWS,old sex laws webranded unfair today,en,http://www.this
,FW: Gay athletes hope games will boost Olympics bid... [The Guardian],,,Gay athletes hope games will boost Olympics bid,en,"http://www.guardianunlimited.
```

```
,APPLICATIONS FOR NATIONAL GAY AND LESBIAN TASK FORCE MESSENGER-ANDERSON JOURNALISM INTERNSHIP/SCHOLARSHIP P
,New Organization for Transgender/Transsexual Veterans,USA,NEWS,New Organization for Transgender/Transsexual
,Minnesota Gender Law Narrowed,"MN,USA",LEGAL,Minnta Gender Law Narrowed,en,http://www.gaycitynews.com/gcn31
,Key West makes move to protect civil rights of transgender people,"FL,USA",PEOPLE,Key West makes move to pr
```

12: Detect Language

Language detection was not just the project's easiest step. It was the **only** easy step!

A Python library called **langdetect** was used to process the text body of all articles:

```
def detect_language(self, path_out):
    articles = [article for article in os.listdir(self.path_in) if article.endswith('.txt')]
    df = pd.DataFrame(columns=['POSTID', 'LANG'])

    for article in articles:
        try:
            with open(os.path.join(self.path_in, article), 'r') as f:
                txt = f.read()
        except:
            with open(os.path.join(self.path_in, article), 'r', encoding='utf-8') as f:
                txt = f.read()

        try:
            language = lang.detect(txt)
        except:
            language = ''

        id = article.split('.')[0]
        df = df.append(pd.DataFrame([[id, language]], columns=['POSTID', 'LANG']))
        print(id, '--->', language)

    df = df.reset_index(drop=True)
    df.to_csv(path_out, sep='\t', index=None)
```

Article counts by language →

```
('en', 144816)
('es', 2451)
('fr', 1319)
('de', 965)
('it', 772)
('pt', 342)
('', 23)
('ca', 7)
('nl', 4)
('no', 1)
('id', 1)
('sk', 1)
```

Languages were detected through each article's text, but titles could have also been used:

```
ES: Interior reforma el protocolo de detenciones para evitar malos tratos,SPAIN,,Interior reforma el protocolo de detenciones para evitar malos tratos,e
"ES: Argentina elige film sobre ""hermafrodita"" para el Oscar",SPAIN|ARGENTINA,FILM,"Argentina elige film sob""hermafrodita"" para el Oscar",es,http://
DE [Deutschland] Bericht: Ho mos ändern Geschlecht im Iran,GERMANY|IRAN,,DE Bericht: Ho mos ändern Geschlecht im Iran,de,http://www.queer.de/news_detail
ES [Argentina] Juez argentino autoriza cambio de sexo a menor de edad,SPAIN|ARGENTINA,,ES Juez argentino autoriza cambio de sexo a menor de edad,es,http
"ES [Argentina] El filme ""XXY"" repre sentará al país en los Oscar y Goya",SPAIN|UK|ARGENTINA,,"ES El filme ""XXY"" repsentará al país en los Oscar y G
DE: Medien: Bei RTL2 heißt es ab 1. Oktober Probewohnen für ein Gratis-Haus im Fernsehen,SPAIN|FRANCE,,DE: Medi Bei RTL2 heißt es ab 1. Oktober Probewohn
ES: Denuncia a la Armada por discriminarle tras su boda con una transexual,SPAIN,,Denuncia a la Armada por discriminarle tras su boda con una transexual,
ES [Iran] La Fundació Tàpies abo rda la historia de Irán en fotos,SPAIN|UK|IRAN,,ES La Fundació Tàpies abo rda la historia de Irán en fotos,es,http://ww
ES [España] UN COLECTIVO TRANSEXU AL PIDE QUE LOS PRIMEROS TRATAMIEN TOS PARA EL CAMBIO DE SEXO SEAN E N LOS CENTROS DE ATENCIÓN PRIMARIA,SPAIN,,ES UN C
EN [USA] The New Issue: Trans Politics/ Why Our Government Can't Take The Heat,USA|UK,POLITICS|GOVERNMENT|NEWS,EN The New Issue: Trans Politics/ Why Our
```

13: Extract URLs

Regular expressions were used to process the large variance in URL structure:

```
txt = re.sub(r'www.', 'http://www.', txt, flags=re.IGNORECASE)  # format if no http and only www present
txt = re.sub(r'http://http://www.', 'http://www.', txt, flags=re.IGNORECASE)  # reconfigure http part
txt = re.sub(r'http', ' http', txt)  # ensure http is not connected to any previous text

txt = re.sub(r'[-=#.*+~_]{7,}', ' ', txt, flags=re.IGNORECASE)  # replace line dividers
txt = txt.replace('(', ' ').replace(')', ' ').replace('[', ' ').replace(']', ' ').replace('<', ' ').replace('>', ' ')
txt = txt.replace('\n\n', ' ').replace('\n', '')

txt = re.sub(r'.html', '.html ', txt, flags=re.IGNORECASE)
txt = re.sub(r'.html \?', '.html?', txt, flags=re.IGNORECASE)
txt = re.sub(r'.html &', '.html&', txt, flags=re.IGNORECASE)

urls = re.findall(r'http[s]?://(?:[A-Z]|[0-9]|[{}!$#-~_@.&+]|[!*\(\),] |(?:%[0-9A-F][0-9A-F]))+', txt, re.IGNORECASE)
```

Some URLs spanned multiple rows of text, as seen below:

```
Syracuse.com June 27 2002

http://www.syracuse.com/newsflash/regional/index.ssf?/newsflash/get_st
ory.ssf?/cgi-free/getstory_ssf.cgi?n0937_BC_NY--
SexualOrientation&&news&nystatenews

Prospects dim for sexual orientation non-discrimination act

By JOEL STASHENKO
The Associated Press
6/27/02 12:00 AM

ALBANY, N.Y. (AP) -- Advocates for banning discrimination in New York
state based on sexual orientation say the issue appears all but dead
in 2002.

"We're not holding out much hope," said Joe Grabarz, executive
```

Fine-tuning is needed. Words like “retrieved” and day of the week should be excluded:

```
http://www.sfgate.com:80/cgi-bin/article.cgi?file=/chronicle/archive/2000/08/16/DD60603.DTLRETRIEVED:
http://www.ireland.com:80/newspaper/ireland/2000/1021/courts7.htmRETRIEVED:-
http://www.ananova.com/news/story/sm 101158.html?nav src=newsIndexHeadlineTuesday,
http://www.lawzone.co.uk/cgi-bin/item.cgi?id=31166&d=101&h=175&f=173&dateformat=%o%20%B%20%YSaturday,
```

14: Extract Publisher

Extracting the publisher's information was not an easy task. While this information exists in the text of almost all articles, its location and order has too much variance.

The easiest solution was to use copyright information, which is included at the bottom of most articles, as proxy for publisher:

ORIGINAL TEXT:

```
Police have yet to release the name of the victim or a description of
any possible suspects. However, if you have information, call the
Jacksonville Sheriff's Office at (904)630-0500.

Updated: August 8, 2002 7:00 PM

© 2002 First Coast News. All rights reserved.
```

EXTRACTED SOURCE:

URLS	SOURCES
:tp://tinyurl.com/1...	2002 First Coast News

The copyright symbol has been removed and only the text before the first period is kept.

15: Process URLs

- 300,000 URLs were extracted and checked online for existence.
- Headlines were searched in the HTML response and the URL **Confirmed** if found.
- Other statuses are **Forbidden**, **Invalid, Not Found**; and, for all other issues, **Unknown**.
- The URLs list was split and run in parallel on two computers and multiple VMs.

A log file was created to catch the live status of all the URLs:

```
  1 CORPUS   POSTID   STATUS   URL
  2 corpus_1     000001   Invalid http://www.the-times.co.uk/news/pages/tim/2000/07/27/timnwsnws01015.html
  3 corpus_1     000001   Unknown http://www.lawzone.co.uk/cgi-bin/item.cgi?id=22982&d=101&h=175&f=173&dateformat
  4 corpus_1     000002   Invalid http://www.lineone.net/express/00/07/28/features/f0200woman-d.html
  5 corpus_1     000002   Invalid http://www.image100.com.She
  6 corpus_1     000003   Not Found    http://www.sun-sentinel.com/news/daily/detail/0,1136,32500000000113111,00.h

255954 corpus_2     090901   Forbidden    http://www.autostraddle.com/100-black-lesbian-bisexual-queer-and-transge
255955 corpus_2     090902   Confirmed    http://ideas.time.com/2014/02/28/dont-applaud-jared-letos-transgender-mar
255956 corpus_2     090903   Confirmed    http://www.huffingtonpost.com/jason-rozek/it-turns-out-my-partner-is-a-w
255957 corpus_2     090904   Confirmed    http://www.huffingtonpost.com/j-mase-iii/im-trans-and-i-have-a-rig_b_486
255958 corpus_2     090905   Unknown http://www.huffingtonpost.com/christopher-edwards/an-open-letter-to-barneys_
255959 corpus_2     090906   Not Found    http://www.lifesitenews.com/news/deconstructing-gods-plan
255960 corpus_2     090907   Confirmed    http://www.huffingtonpost.com/lauren-lubin/trans-people_b_4870501.html

300195 corpus_3     000463   Not Found    http://www.ngltf.org/about/messenger.htm
300196 corpus_3     000463   Unknown http://www.ngltf.org
300197 corpus_3     000465   Unknown http://www.gaycitynews.com/gcn31/minnesota.html
300198 corpus_3     000467   Unknown http://www.miami.com/mld/miamiherald/living/people/gay_lesbian/4842184.htm
300199 corpus_3     000474   Unknown http://www.tgender.net/mailman/listinfo/gain-all
```

Only 51K (1/6 of all URLs) are still accessible online and include the original content:

```
('Unknown', 143929)
('Not Found', 65039)
('Confirmed', 50992)
('Invalid', 37445)
('Forbidden', 2793)
```

16: Statistics and Visualizations

Top 20 Countries

```
('USA', 87253)
('UK', 14836)
('CANADA', 6857)
('AUSTRALIA', 4794)
('INDIA', 3168)
('SPAIN', 2409)
('NEW_ZEALAND', 1767)
('FRANCE', 1504)
('IRELAND', 979)
('ITALY', 931)
('MALASYA', 823)
('THAILAND', 795)
('EUROPE', 792)
('AFRICA', 705)
('GERMANY', 680)
('TURKEY', 664)
('PAKISTAN', 655)
('JAPAN', 485)
('BRAZIL', 484)
('PHILIPPINES', 477)
```

Top 20 USA States

```
('TX', 3488)
('VA', 1311)
('NY', 1256)
('CA', 1237)
('PR', 1032)
('WA', 915)
('TN', 903)
('UT', 714)
('DC', 640)
('WI', 614)
('MA', 528)
('GA', 478)
('SC', 351)
('SD', 279)
('IL', 256)
('PA', 232)
('VT', 199)
('WV', 192)
('FL', 184)
('WY', 128)
```

Top 20 Words in Articles

```
('said', '509088')
('transgender', '486382')
('people', '412843')
('gender', '363093')
('right', '307934')
('woman', '273831')
('gay', '273300')
('year', '250127')
('one', '237777')
('would', '233648')
('trans', '215988')
('sex', '196604')
('say', '194980')
('state', '179918')
('time', '172618')
('school', '169235')
('community', '164596')
('law', '163864')
('new', '163000')
('like', '161016')
```

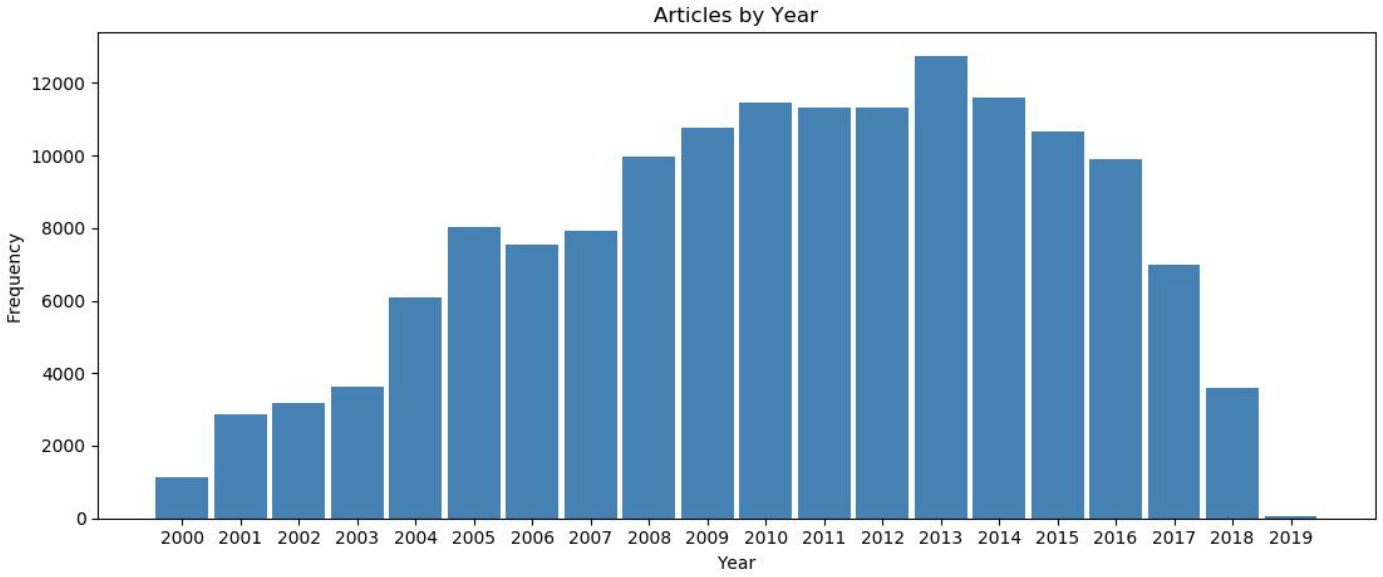
Top 20 Words in Titles

```
('transgender', 29844)
('trans', 12262)
('gay', 9295)
('gender', 8665)
('rights', 7027)
('sex', 6265)
('woman', 5900)
('bill', 5583)
('new', 5322)
('lgbt', 5306)
('discrimination', 4768)
('change', 4377)
('people', 4053)
('man', 3663)
('law', 3386)
('school', 3129)
('anti', 3108)
('transsexual', 3089)
('court', 2884)
('policy', 2652)
```

16: Statistics and Visualizations

```
('2000', 1145)
('2001', 2862)
('2002', 3173)
('2003', 3614)
('2004', 6079)
('2005', 8026)
('2006', 7539)
('2007', 7918)
('2008', 9981)
('2009', 10763)
('2010', 11449)
('2011', 11324)
('2012', 11315)
('2013', 12747)
('2014', 11590)
('2015', 10659)
('2016', 9885)
('2017', 6987)
('2018', 3584)
('2019', 62)
```

← Articles by Year



Articles by Language →

```
('en', 144816)
('es', 2451)
('fr', 1319)
('de', 965)
('it', 772)
('pt', 342)
('', 23)
('ca', 7)
('nl', 4)
('no', 1)
('id', 1)
('sk', 1)
```

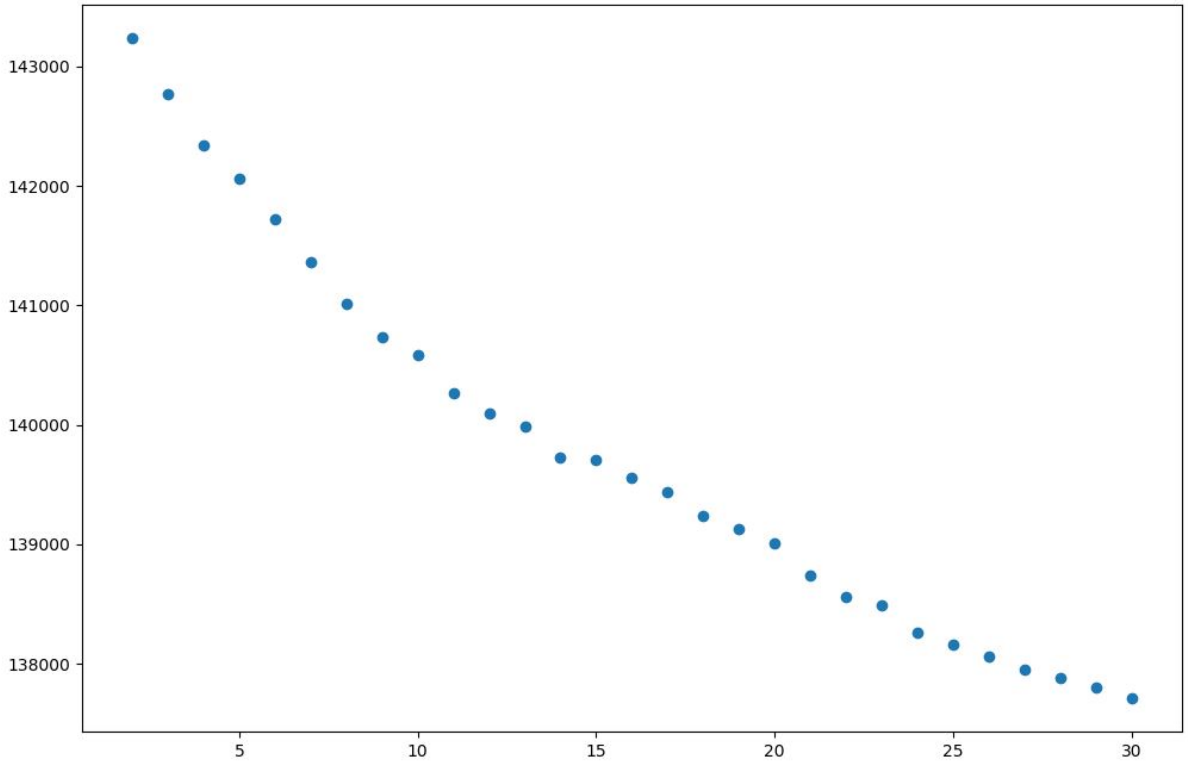
16: Statistics and Visualizations

← Word Cloud of Articles

Word Cloud of Titles →

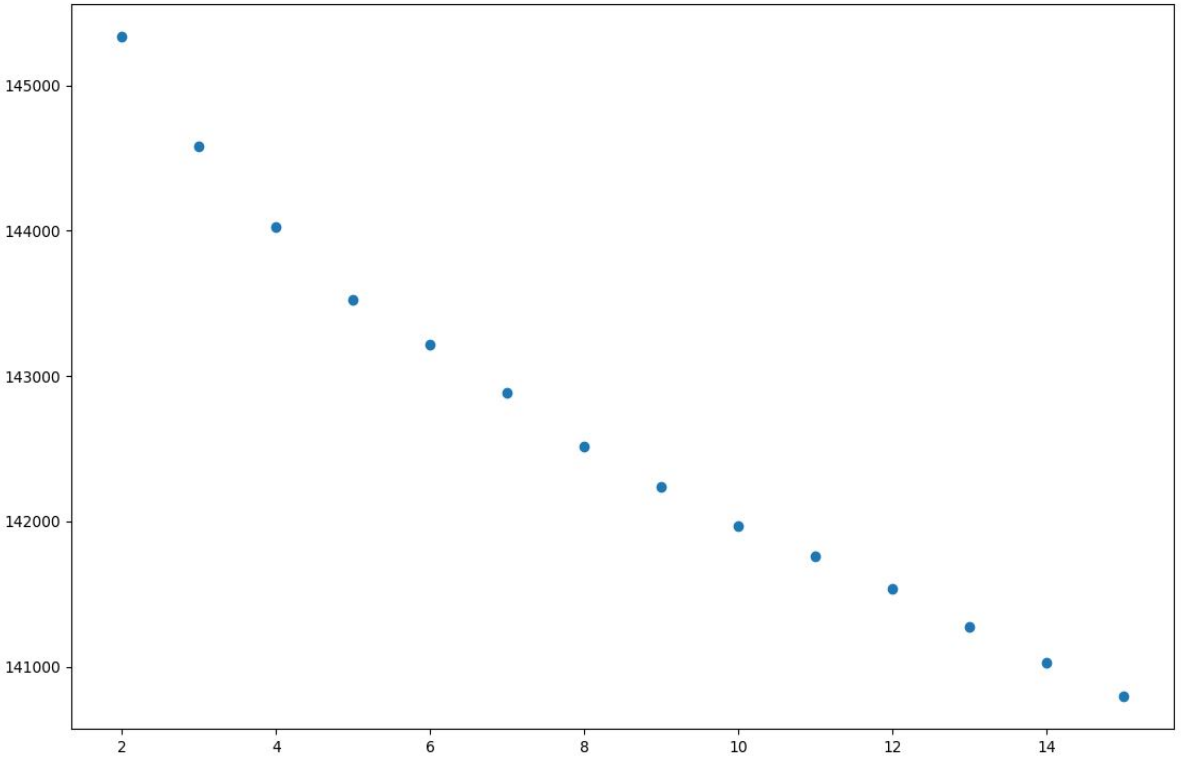
17: K-Means Clustering Titles

- Clustering on all articles was performed using scikit-learn's TfidfVectorizer (**30 clusters**).
- Ignore all terms that appear **less than two times** in corpus.
- TF-IDF matrix shape: **144,816 rows by 24,269** features.
- Completion time: **33 min.** (AMD Ryzen 2700X / 16 GB RAM).
- No satisfactory results were found on the WSS graph using the elbow technique.



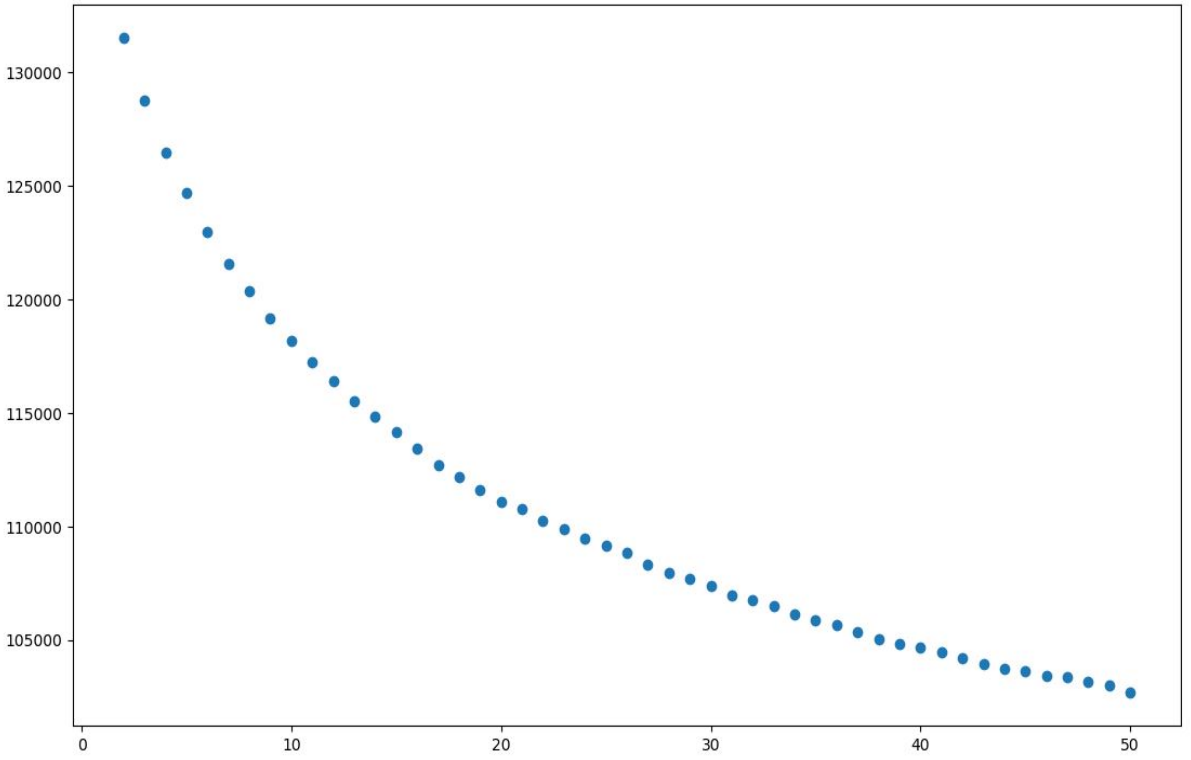
18: K-Means Clustering Articles

- Clustering on all articles was performed using scikit-learn's TfidfVectorizer (**15 clusters**).
- Ignore all terms that appear **less than two times** in corpus.
- TF-IDF matrix shape: **150K rows by 450K features**.
- Completion time: **27 hours** (AMD Ryzen 2700X / 16 GB RAM).
- No satisfactory results were found on the WSS graph using the elbow technique.



18: K-Means Clustering Articles

- Clustering on all articles was performed using scikit-learn's TfidfVectorizer (**50 clusters**).
- Ignore all terms that appear in **less that 10% and more than 30%** of corpus.
- TF-IDF matrix shape: **150K rows by 359 features**.
- Completion time: **24 hours** (AMD Ryzen 2700X / 16 GB RAM).
- Elbow found after **dimensionality reduction** and cluster number increase (~15 clusters).

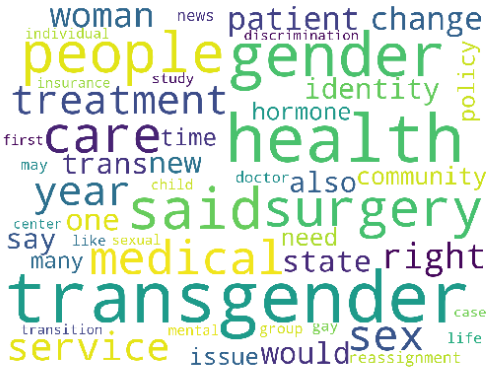


18: K-Means Clustering Articles

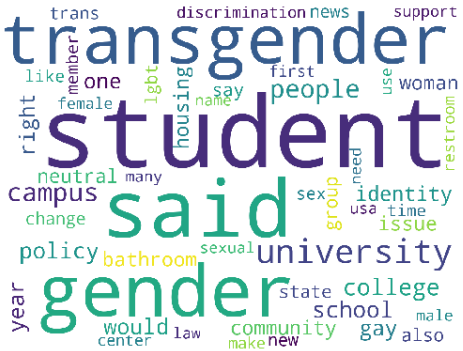
WORD CLOUDS BY SELECTED CLUSTER NUMBERS (C#)

C1: Non-English

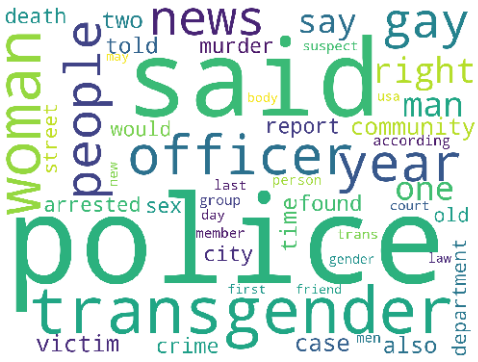
C2: Medical/Health



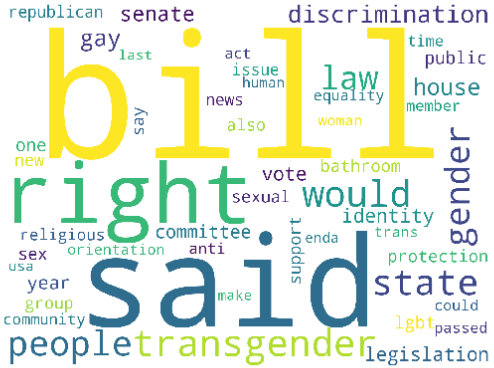
C3: Education (Adults)



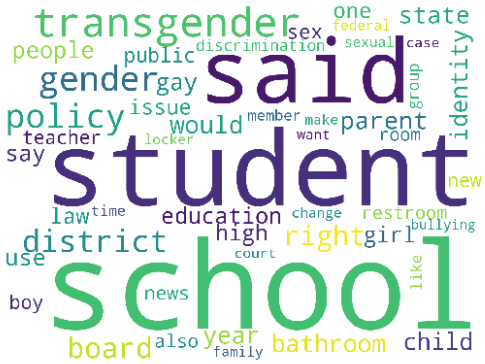
C5: Police



C7: Legislation



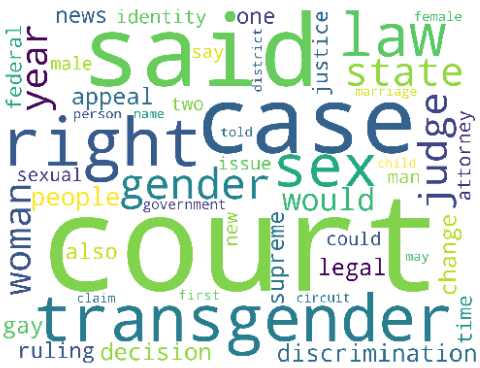
C8: Education (Kids)



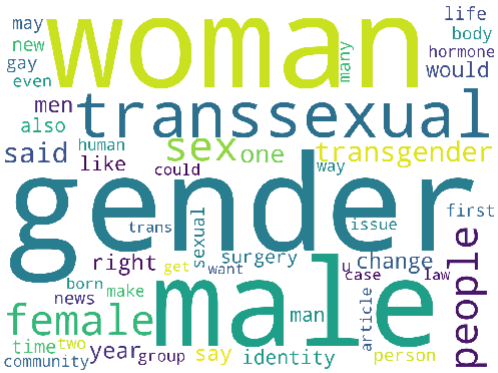
18: K-Means Clustering Articles

WORD CLOUDS BY SELECTED CLUSTER NUMBERS (C#)

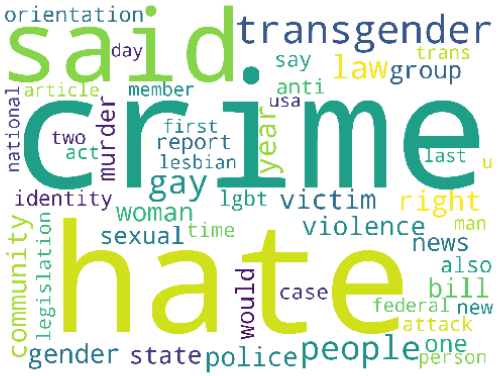
C9: Legal Issues



C11: Gender Issues

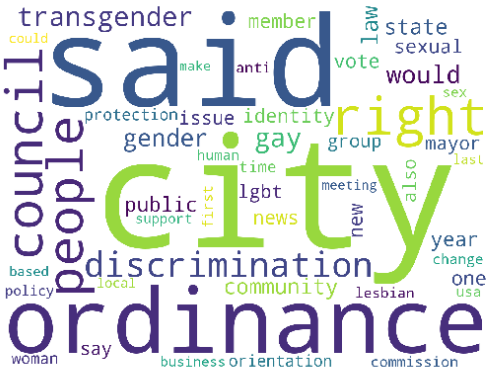


C12: Hate Crimes

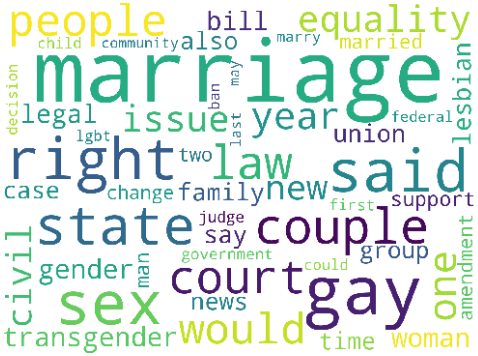


C13: Transgender

C14: Local Ordinances



C15: Marriage Equality



Summary

• The most time-consuming activity was downloading the articles due to the delays introduced to web-scraping between server requests (~2 months).

• Text cleaning and data extraction was the second most time-consuming phase (~1 month).

• Once the data was cleaned (normalized, stemmed/lemmatized, stopwords and non-related text removed, etc.), the K-Means clustering process was relatively fast (~4 days).

• Checking the live status of 300K URLs would have taken about 15 days, but it was reduced to 3 days after distributing the operations to 10 different computers/VMs.

• Dimensionality reduction was important for increased performance and better cluster model fitting. Resulting clusters show somewhat clear topics, as inferred from the word clouds.

• Non-English articles should be excluded from the clustering process in future iterations.

• Extra words may need to be added to the stopwords list. For example, the word “said” appears in over half of the clusters. It suggests that stories rely heavily on accounts told by people quoted in the articles. The same applies to the word “transgender”, which appears frequently. Identifying and removing these words might result in better clustering results.

Future Actions

Cleaning / Text Processing

- Split embedded metadata (publisher, author, headline, etc.) and article body.
- Further explore articles manually and clean non-related text.
- Addition of new stopwords.
- Fine-tune URL extraction

Descriptive Statistics

- Get article word counts.
- Compare non-confirmed URLs against Wayback Machine records at Archive.org.

NLP

- Part-of-Speech (PoS) tagging.
- Sentiment analysis.
- Latent Semantic Analysis (LSA).

Machine Learning

- Fine-tune K-Means clustering model.
- Manually label a few thousand articles to perform supervised learning.

Implementation

- Create search engine using Information Retrieval (IR) and implement as web app.

Other

- Ethical and legal issues such as copyright / fair use for research and publication.


Thank You!