



University of
Sheffield

Text as Data

Data Journalism

Dani Madrid-Morales

19 April 2023

A WORLD
TOP 100
UNIVERSITY

Today's Learning Outcomes

1. To list different ways in which **text data can be incorporated in news stories**.
2. To apply a **simple workflow** (gather > clean > analyze > visualize) in the use of text data for data journalism.
3. To create **visualizations from text data** using widely adopted web-based tools.

Today's data files

You can access some of the files we will need today, including these slides from here

<https://bit.ly/uc3m-data>



Unstructured vs. structured data

You might be familiar with this data format

Artist	Popularity	Track Name	Release Data	Genre	Subgenre
Ed Sheeran	66	I Don't Care (with Justin Bieber)	14/06/2019	pop	dance pop
Maroon 5	67	Memories (Dillon Francis Remix)	13/12/2019	pop	dance pop
Zara Larsson	70	All the Time (Don Diablo Remix)	05/07/2019	pop	dance pop
The Chainsmokers	60	Call You Mine - The Remixes	19/07/2019	pop	dance pop
Lewis Capaldi	69	Someone You Loved (Future Humans Remix)	05/03/2019	pop	dance pop
Ed Sheeran	67	Beautiful People (feat. Khalid)	11/07/2019	pop	dance pop
Katy Perry	62	Never Really Over (R3HAB Remix)	26/07/2019	pop	dance pop
Sam Feldt	69	Post Malone (feat. RANI)	29/08/2019	pop	dance pop
Avicii	68	Tough Love	14/06/2019	pop	dance pop
Shawn Mendes	67	If I Can't Have You (Gryffin Remix)	20/06/2019	pop	dance pop
Ed Sheeran	58	Cross Me	21/06/2019	pop	dance pop
Ellie Goulding	67	Hate Me (R3HAB Remix)	16/08/2019	pop	dance pop
Loud Luxury	67	Body On My	29/03/2019	pop	dance pop

Text data is “unstructured”

Administration

APRIL 13, 2023

Remarks by President Biden at
Banquet Dinner

 BRIEFING ROOM  SPEECHES AND REMARKS


Dublin Castle
Dublin, Republic of Ireland

9:09 P.M. IST

THE PRESIDENT: Please — (applause) — thank you very much. Please — please sit down. My colleagues with me know me, and they're never going to let me forget this.

First of all, what I haven't done yet is I'm going to ask — I won't take all the time to introduce everyone, but I'd like to ask all the members and former members of the Congress and the min- — from previous administrations who are here as part of our delegation to please stand up and let everybody see — all of them. Chris, stand up. (Applause.)

They're the reason why even if I didn't believe what I believed, I'd have to believe what I believe. (Laughter.) Because they're all strong, strong, strong supporters of Ireland, and peace and security in Europe.


CNN International 📍 @cnni · 34 min

After winning a best actress Oscar for her role in “Everything Everywhere All At Once” last month, Michelle Yeoh is preparing to step back into the Star Trek universe to reprise her role as Emperor Philippa Georgiou in the new “Star Trek: Section 31” movie



cnn.com
Michelle Yeoh set to return as Emperor Philippa Georgiou in new 'Sta...
Live long and prosper, Michelle Yeoh.

4 15 46 20,4m

CNN International 📍 @cnni · 45 min

A 13-year-old in Ohio has died after “he took a bunch of Benadryl,” trying a dangerous TikTok challenge that’s circulating online, according to a CNN affiliate and a GoFundMe account from his family



cnn.com
A 13-year-old died in Ohio after participating in a Benadryl TikTok 'ch...
A 13-year-old in Ohio has died after "he took a bunch of Benadryl," trying a dangerous TikTok challenge that's circulating online, ...

9 20 40 23,9m

Most common types of data used by data journalists

Table 2. Kind of data (multiple coding possible, $n = 222$).

Kind of data	%
Geodata	47.3
Financial data	45.0
Measured values	38.3
Sociodemographical data	35.1
Personal data	30.2
Metadata	15.8
Poll or survey data	12.6

Loosen, W., Reimer, J., & De Silva-Schmidt, F. (2020). Data-driven reporting: An on-going (r)evolution? An analysis of projects nominated for the Data Journalism Awards 2013–2016. *Journalism*, 21(9), 1246–1263. doi: 10.1177/1464884917735691

From “unstructured” data ...

“The party arranged by the great leader comrade Kim Jong Il for the art group of Korean schoolchildren in Japan was held at the Mansudae Art Theatre on Monday evening.

Attending the party were Kim Su Jin, director of the education department of the central standing committee of the General Association of Korean Residents in Japan (Chongryon), and members of the art group of Korean schoolchildren in Japan led by Kim Sun, chief of the central standing committee, staying in the socialist fatherland.

“The Chinese Embassy here donated 7 tons of rice to the Kim Jong Suk Creche. Economic and Commercial Counsellor Chen Yufa and other officials of the Embassy conveyed the rice to the creche on Thursday. Kim Mun Song, Vice-Chairman of the State External Economic Affairs Commission, and officials of the creche were present at the ceremony.”

KCNA

“The “civilian”-veiled fascist clique of South Korea sentenced Kang Wi Won who was the chairman of the fifth-term South Korean Federation of University Student Councils (Hanchongryon) at the puppet Kwangju district court on November 25 and sentenced him to a prison term of six years and suspension of qualification of three years, a South Korean newspaper said. The fascist clique arrested him in July, claiming that he organised the inaugural meeting of the fifth-term Hanchongryon and took the lead in the struggle for independence, democracy and reunification..”

“The South Korean fascist clique detained Prof. Pak Ji Hyeon, a professor at the Press Graduate School of Kwangju, on November 28 on the charge of violation of the National Security Law”, according to a radio report from Seoul. Pak authored and published “method on true understanding and discourse”, a textbook for high school and university students, in April last year, in which he vindicates the Juche idea. Earlier, on the 27th, he was brought to the puppet Kwangju district court for questioning. The professor said the charge the prosecution imposed on him was an unjustifiable “application of law” and that he would take a strong counteraction against the authorities' suppression.”

... to structured data

	Total	doc1	doc2	doc3	doc4
according	1	0	0	0	1
affairs	1	0	1	0	0
application	1	0	0	0	1
april	1	0	0	0	1
arranged	1	1	0	0	0
arrested	1	0	0	1	0
art	3	3	0	0	0
association	1	1	0	0	0
attending	1	1	0	0	0
authored	1	0	0	0	1
authorities	1	0	0	0	1
brought	1	0	0	0	1
central	2	2	0	0	0
ceremony	1	0	1	0	0

A hand holding a pen is positioned over a document, with a red overlay covering the entire image. The text 'Using text data in data journalism' is written in white, bold, sans-serif font across the center of the image.

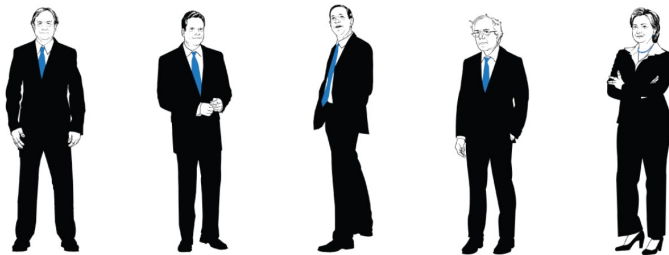
Using text data in data journalism

Data Journalism Using Text – Counting & Comparing

Politics






Deconstructing the #demdebate: Clinton, Sanders control conversation

Oct. 14, 2015



The five Democrats running for president debated for the first time Tuesday night in Las Vegas. The two front-runners owned the conversation. Hillary Clinton and Bernie Sanders together said **56 percent** of the words spoken by candidates.

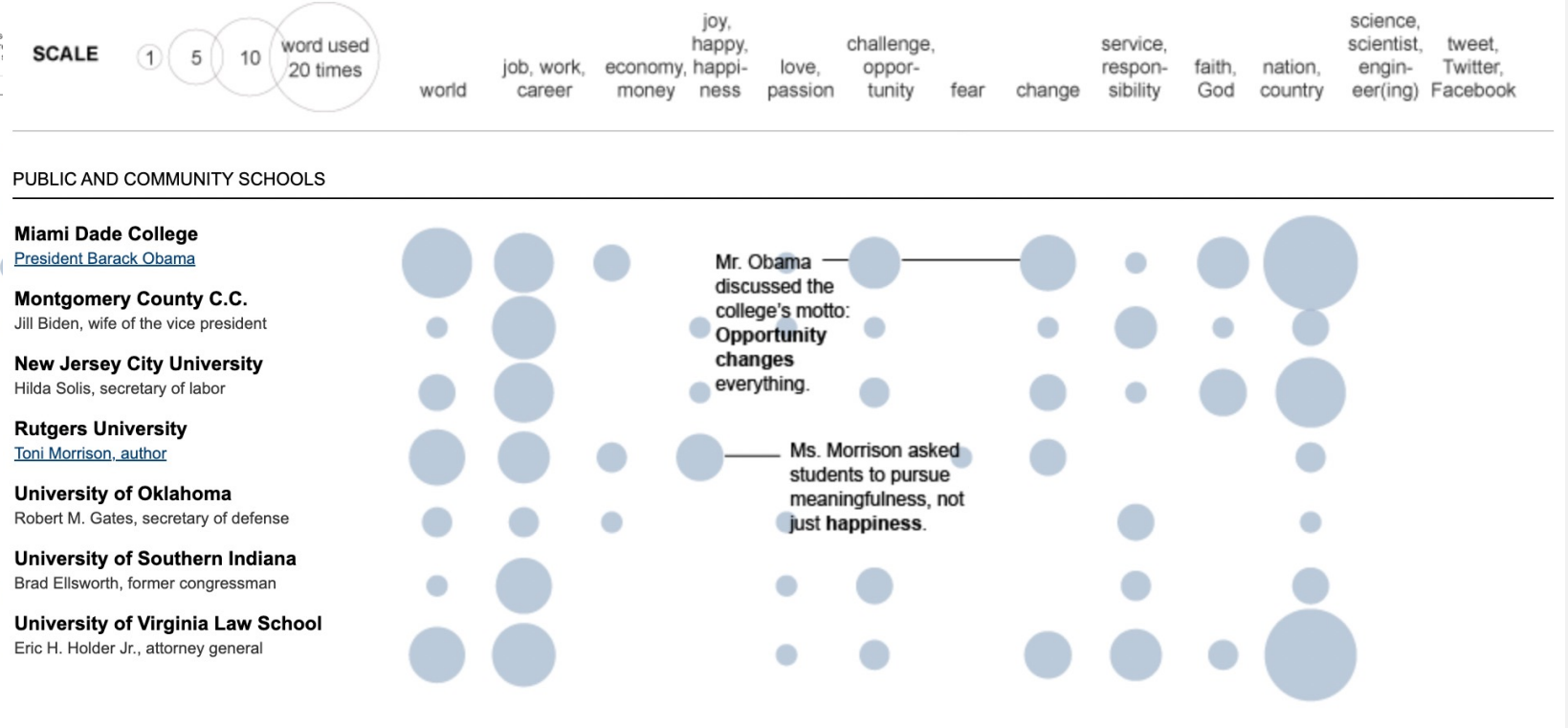
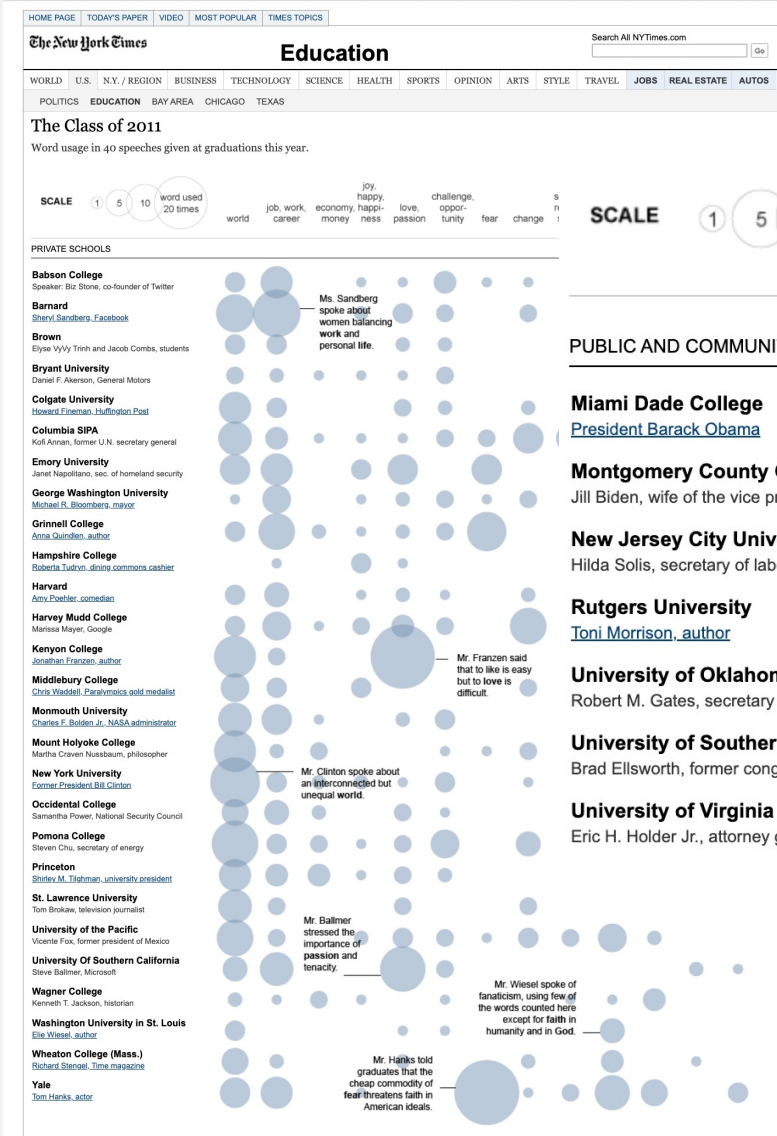
WHAT THEY TALKED ABOUT (BY NUMBER OF MENTIONS)

	 Sanders	 Clinton	 O'Malley	 Webb	 Chafee	Total
Climate change	3	3	3	1	3	13
Guns	5	1	4	2	1	13
Iraq	2	1	1	1	4	9
Middle class	5	3				8
NRA		2	4	1		7
Syria	2	3		1	1	7
Wall Street	5			1		6
Healthcare	4		1			5
Vietnam				4		4
1 percent	4					4

Source:

<https://www.washingtonpost.com/graphics/politics/2016-election/debates/oct-13-speakers/>

Data Journalism Using Text – Counting & Comparing



Source:

<https://archive.nytimes.com/www.nytimes.com/interactive/2011/06/10/education/commencement-speeches-graphic.html>

Data Journalism Using Text – Counting & Comparing (Over Time)

Analysis

London mayor: Commons speeches reveal candidates' differing issue focus

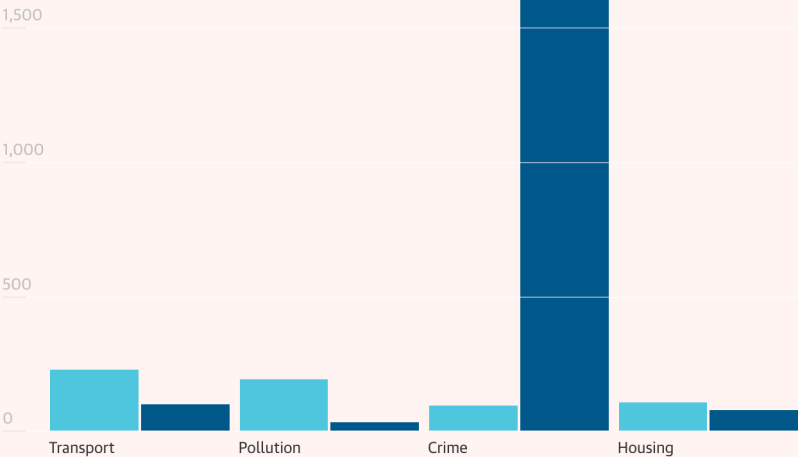
Caelainn Barr

Hansard records suggest Zac Goldsmith spoke on crime less than other manifesto promises while Sadiq Khan discussed pollution less

Mayoral words

Number of times mayoral candidates have used these words or spoken about related legislation in parliament since 2010

Goldsmith Khan

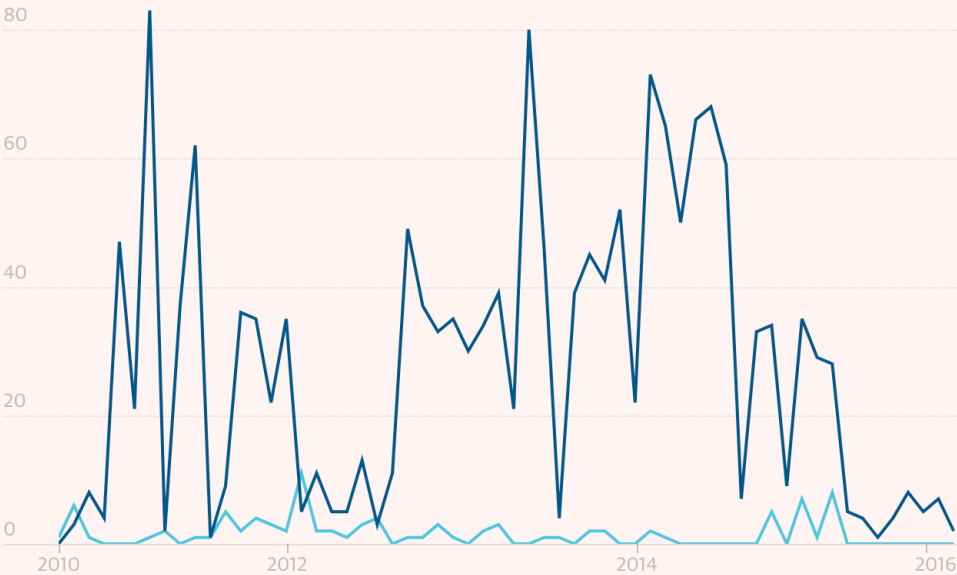


Guardian graphic | Source: Guardian analysis of theyworkforyou.com Hansard data

Crime

Number of times mayoral candidates have talked about crime or spoken about related legislation in parliament since 2010

Goldsmith Khan



Guardian graphic | Source: Guardian analysis of theyworkforyou.com Hansard data

Source: <https://www.theguardian.com/politics/datablog/2016/may/03/london-mayor-data-indicates-candidates-differing-focus-on-issues>

A hand holding a pen is positioned over a document. The entire image is covered with a semi-transparent red overlay. The text 'A workflow for using text data in stories' is written in white, bold font across the center of the image.

A workflow for using text data in stories


Activity – Reproducing an example

Sections

Search

NATIONAL POST

1



Trudeaus rang in New Year at luxury Jamaica estate owned by Trudea...

TRENDING 🔥

2



Coca-Cola's cocaine connection is worth billions

TRENDING 🔥

World / News

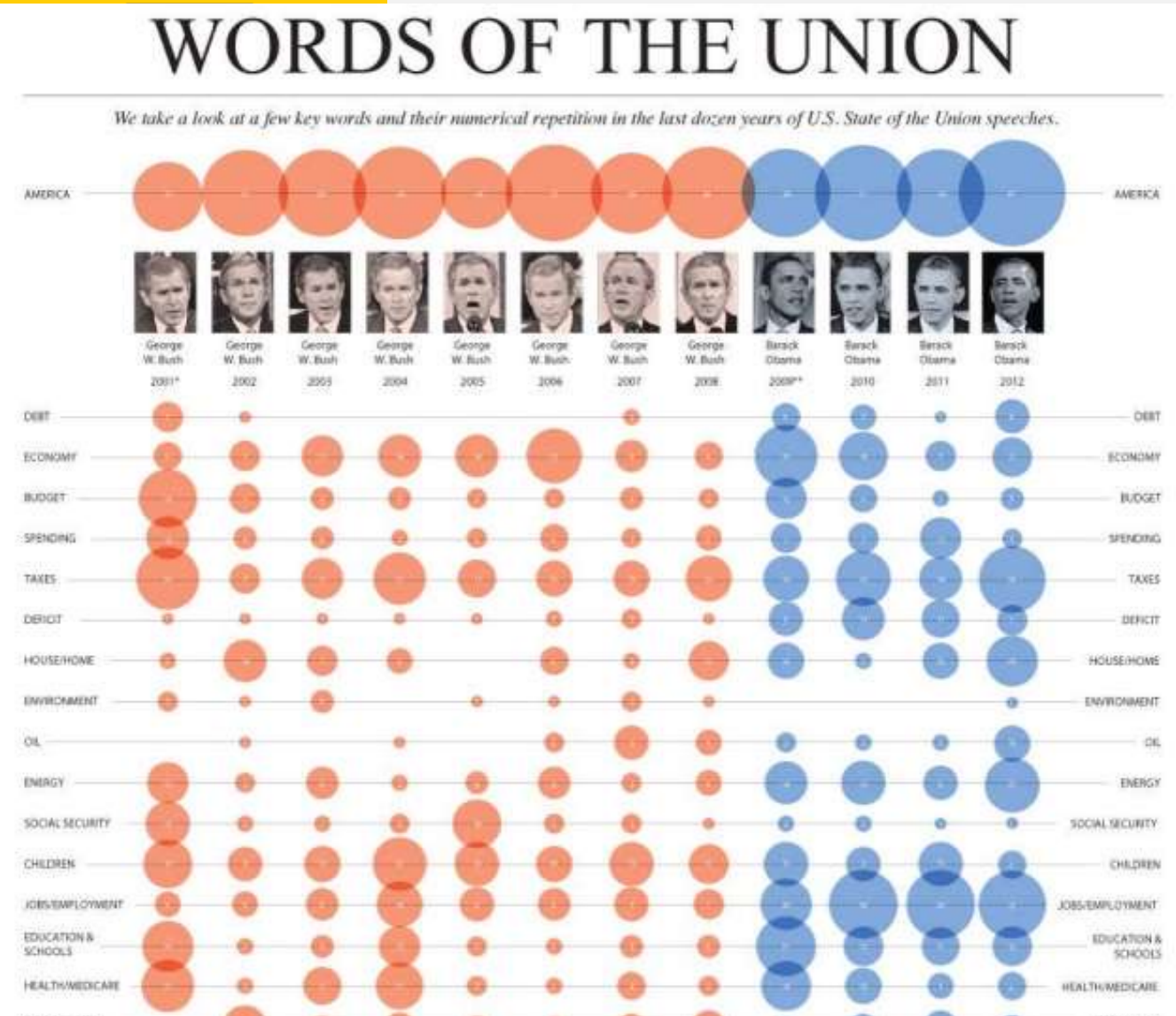
Graphic: The State of the Union's Words

We take a look at a few key words and their numerical repetition in U.S. State of the Union speeches.

Richard Johnson

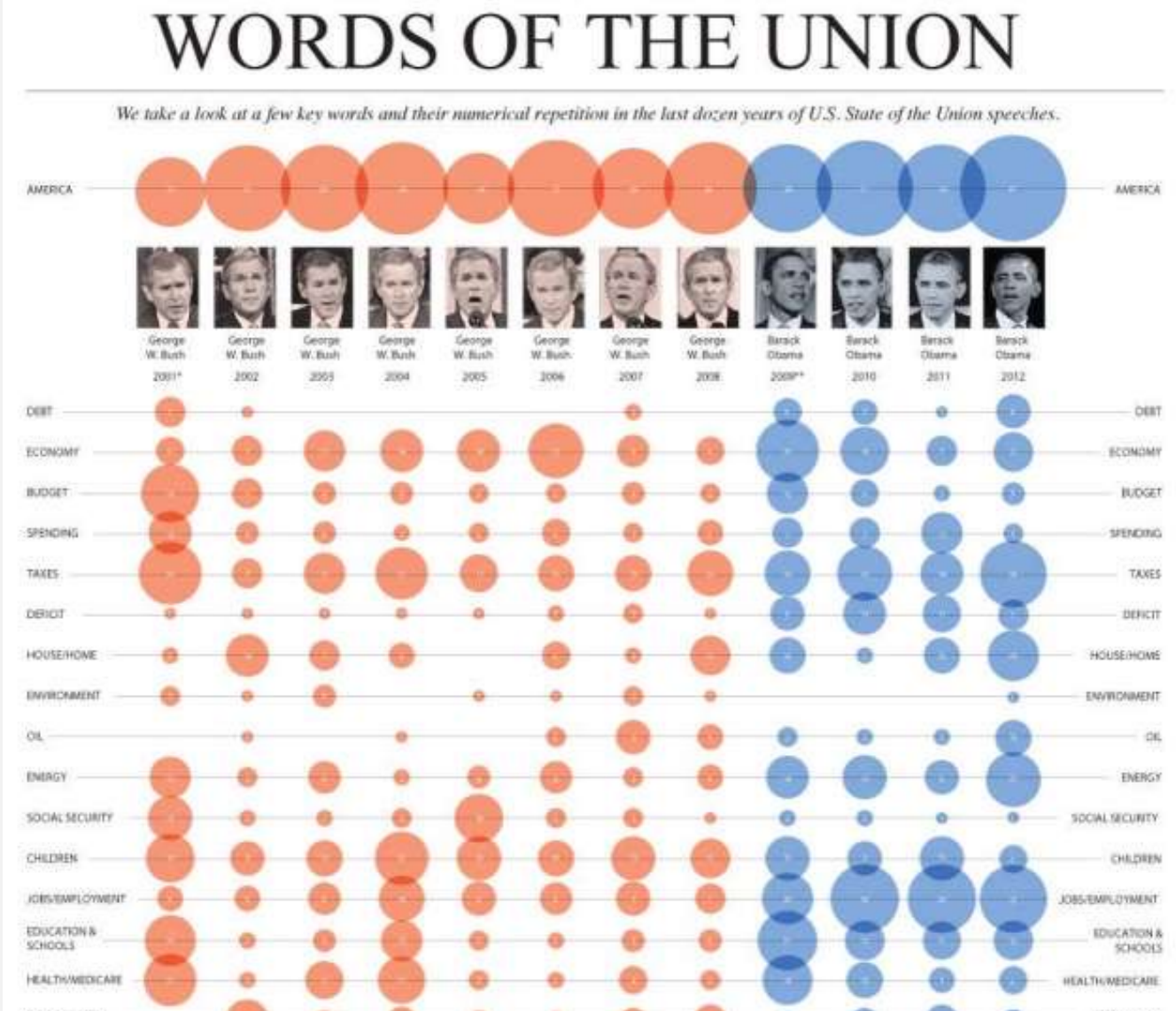
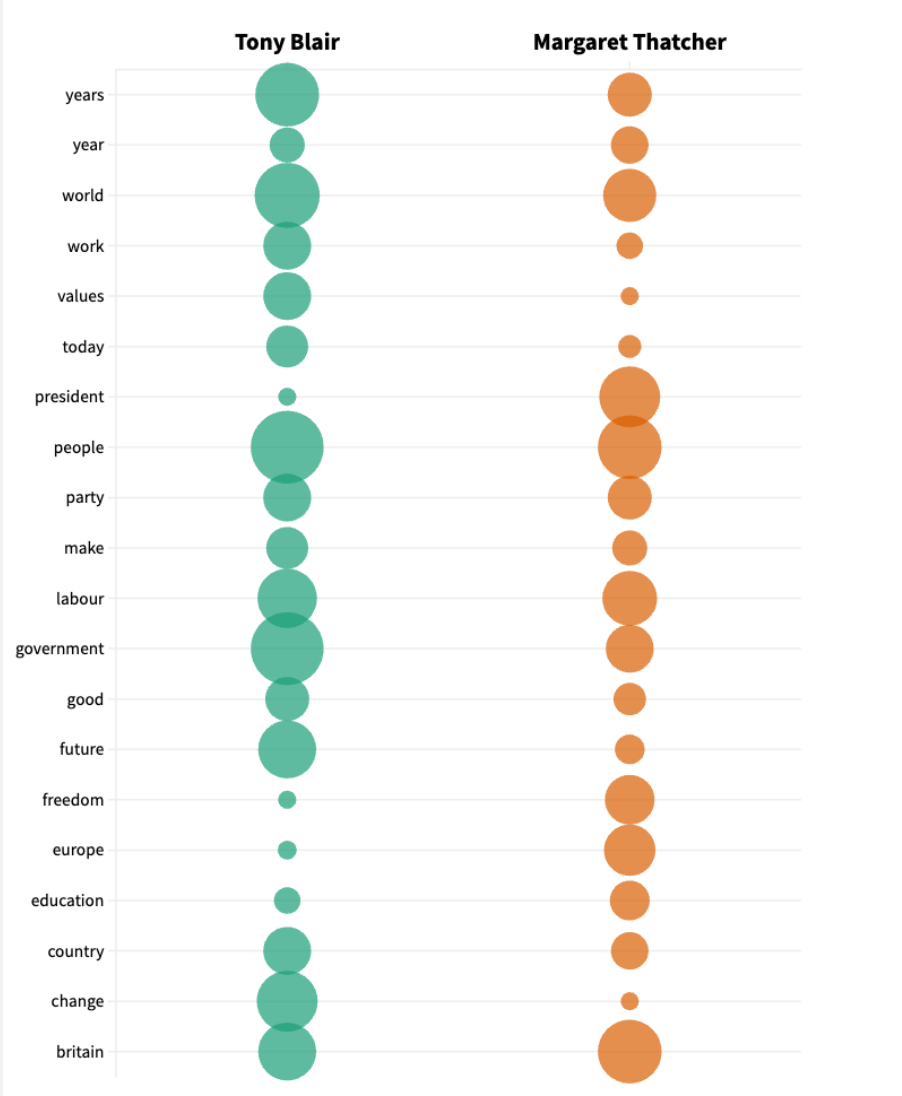
Published Feb 12, 2013 • < 1 minute read

[Join the conversation](#)



Source: <https://nationalpost.com/news/graphics/graphic-the-state-of-the-unions-words>

Activity – Reproducing an example

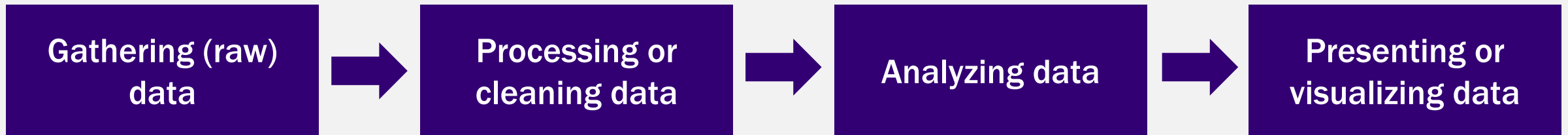


Source: <https://nationalpost.com/news/graphics/graphic-the-state-of-the-unions-words>

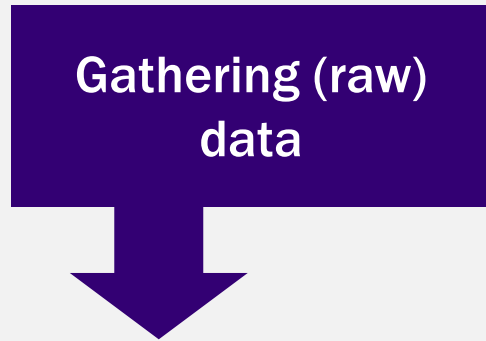
Activity – Reproducing an example: Recommendations

- We will go over the steps together, but if something does not work for you, you will find the files needed to catch up with the group on Aula Global
- Here's the files you have:
 - 01_blair.txt
 - 01_thatcher.txt
 - 02_stopwords.txt
 - 03_tokenizer-table.csv
 - 04_topwords.xlsx

Data journalism workflow



A workflow for writing news stories with text data



- Sources of text data include social media, news websites, government websites, databases...
- Text data are rarely (never) found in tabular format (= structured format), instead you'll find it in different formats such as PDF, HTML, TXT...
- You might need to use some special techniques to gather text data such as web scraping, or making requests through an API.

Activity – Text Data Gathering: Text Scraping with Google Sheets

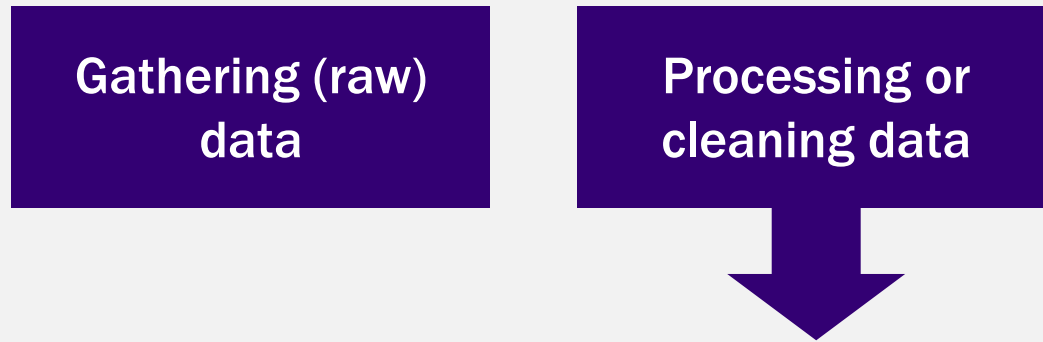
1. We are going to scrape data from a website with speeches by British politicians: <http://www.britishpoliticalspeech.org/speech-archive.htm>
2. Let's look for speeches by Margaret Thatcher, and click on “Leader's speech, Bournemouth 1990”. Copy the URL to the speech.
3. Open a Google Sheet and on Cell A1 paste the URL.
4. We are going to use the =IMPORTXML() formula on Google which requires 2 parameters, a formula an XPath.
5. To get the XPath, using Chrome, right-click on any word of the speech and select ‘Inspect’. Then, within the HTML text, right-click on the part you want to scrape, select ‘Copy’ and then ‘Copy XPath’.
6. Now, back to Sheets, on B1 paste the formula =IMPORTXML(A1, "//*[@id='main-content']/div/div[2]/div[2]").

Activity – Text Data Gathering: Text Scraping with Google Sheets

7. Now, try to do the same with Tony Blair's 2005 Leader's speech in Brighton.
8. We could leave the data as is, but in this case, we are going to save the speeches in separate TXT files, one for each speech.
9. Select rows B1 to BA1, copy the content and paste it as plain text on a Google Doc using the 'Paste without formatting' option.
10. Download this document as a .txt file named 'Thatcher.txt'.
11. Repeat the same operation with the other speech, and save it as 'Blair.txt'.

If you had any issues with these steps, you can download the files '01_blair.txt' and '01_thatcher.txt'.

A workflow for writing news stories with text data



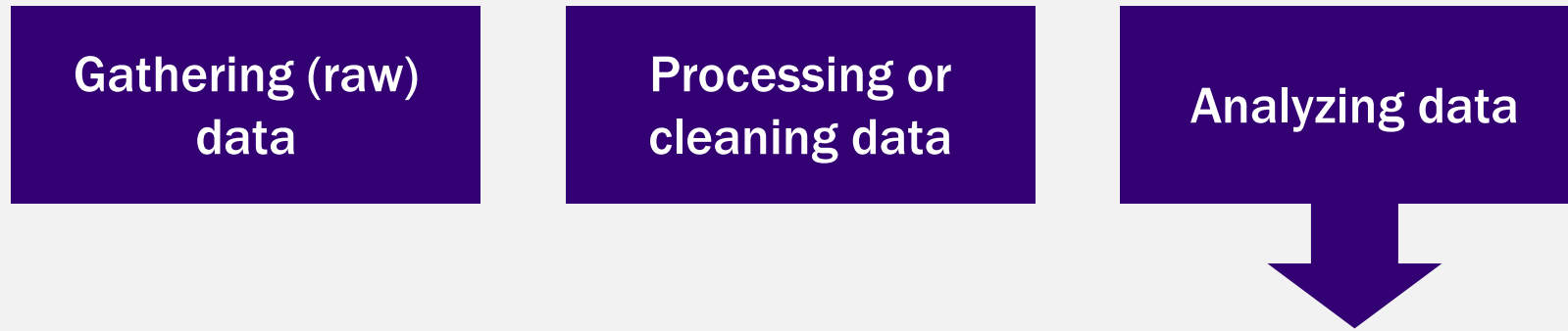
- Unstructured text data needs to be processed before it can be turned into structured text data. The most common steps include: 1) lowercasing documents; 2) removing punctuation and digits; 3) removing stopwords
- Most of these steps can be completed with an online tool such as {Lexos} <http://lexos.wheatoncollege.edu>
- The last step involves breaking down documents into ‘tokens’ (or words), a process known as ‘tokenization’.

Activity – Text Data Cleaning Using {Lexos}

**** IMPORTANT NOTE:** If too many people use {Lexos} at the same time, the server can slow down. So do have patience during this demonstration ******

1. On the landing page of {Lexos}, drag and drop the two TXT files we just created.
2. To begin the data cleaning stage, on the top menu, click 'Prepare' > 'Scrub'.
3. By default, you'll see the basic data cleaning options selected, including 'Make Lowercase', 'Remove digits' and 'Remove Punctuation'.
4. We also want to list some stop words that we want excluded from our data. We are going to use: <https://algs4.cs.princeton.edu/35applications/stopwords.txt>. The file is on Aula Global, upload it to {Lexos}. Make sure to select 'Stop'
5. Next, click on 'Apply' to complete the data cleaning stage.
6. In the 'Prepare' menu, select 'Tokenize' to turn your texts into structured data.

A workflow for writing news stories with text data



- Structured text data can be used to examine the frequency in which words occur in a document, compare different groups/individuals, and see how the use of words has evolved over time.
- A simple frequency analysis can be done with {Lexos}.
- Some more advanced tools allow other forms of analysis such as similarity analysis, sentiment analysis and lexical diversity.

Activity – Frequency Analysis

**** IMPORTANT NOTE:** If too many people use {Lexos} at the same time, the server can slow down. So do have patience during this demonstration ******

1. On the 'Tokenize' page of {Lexos}, select the option 'Raw' and click on 'Generate'.
2. To get a copy of your word frequencies, click 'Download' and select 'Documents as Rows'.
3. Open the CSV with Excel or Sheets for a visual inspection and a preliminary analysis based on the top mentioned words by each politician.
4. {Lexos} provides a tool ('Analyze' >> 'Top Words') to do a more robust analysis. Click on the '?' to get more information on how each parameter works.

If you had any problem with these steps, you can download '03_tokenizer-table.csv'

Activity – Frequency Analysis

1. On Excel, start by going to cell A1 and typing “Words”.
2. On Excel, let’s first find the top 20 words for each leader. ‘Sort’ the frequencies for Blair from Z to A. Highlight in yellow the top 20 words. Repeat the same process with Blair.
3. Now, you could sort all your cells by colour. This can be done by going to the top menu, selecting ‘Sort & Filter’ and then ‘Custom Sort’. In ‘Custom Sort’ you will need to change ‘Values’ to ‘Cell Colour’.
4. Delete the ‘Total’ and ‘Average’ columns. We won’t need them. Select the top words we highlighted in yellow, copy them and paste them to a new spreadsheet. Save the new spreadsheet as “topwords.xlsx”.
5. Next, we need to turn this table from a wide to long format. We need 3 columns: words, values and politician. You will need to do this manually in the next step.

If you had any problem with these steps, you can download ‘04_topwords.xlsx’

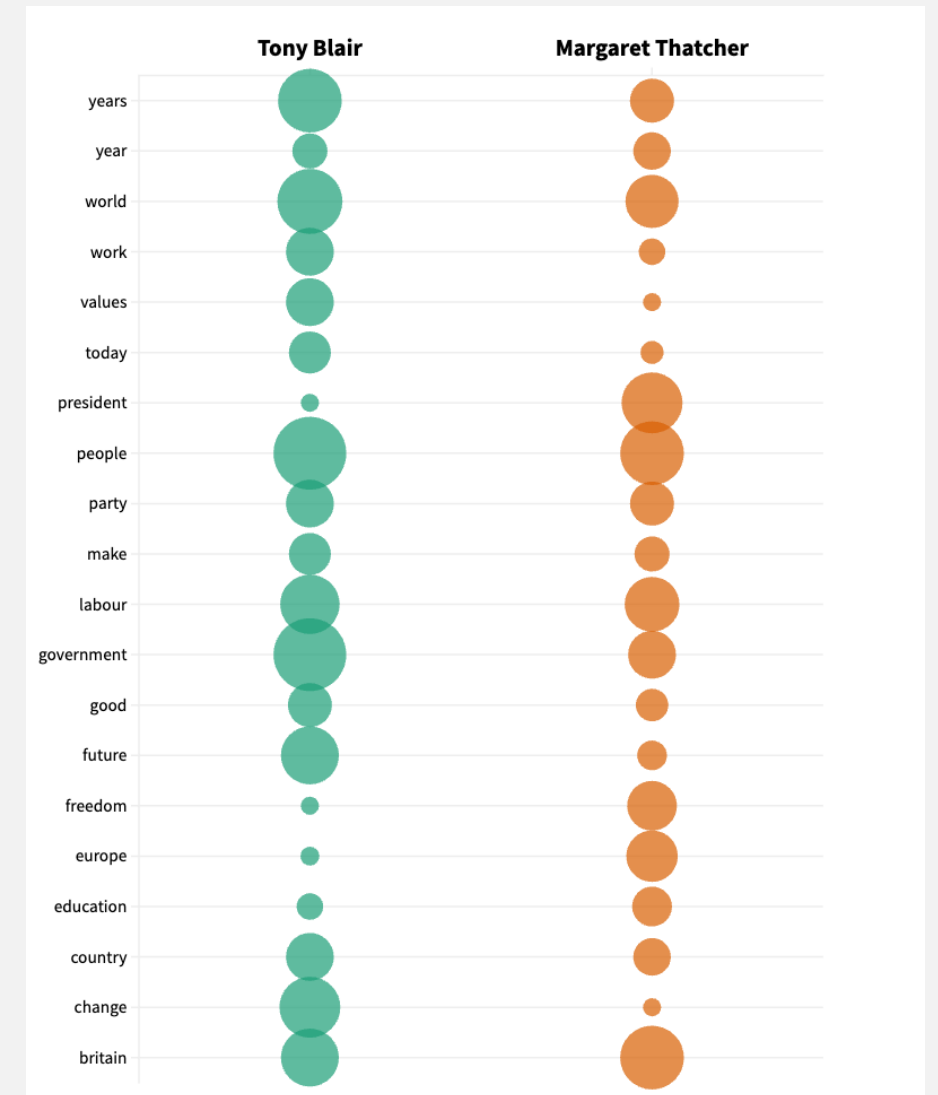
A workflow for writing news stories with text data



- The basic rules of good data visualization apply when visualizing data that was originally in the form of text: simplicity, clarity and accuracy.
- While there are some text specific types of visualization (such as word clouds), oftentimes, the best visualizations are ‘old fashioned’ line charts, bubble charts and bar charts.
- {Lexos} offers a suit of visualizations, but other online platforms are better equipped to visualize your results.

Activity – Create a count plot using bubble charts

1. Transform the list of top words from a wide to long format with three columns: word, count and name of politician.
2. Launch Flourish, create a 'New visualization' and select 'Bubble chart' as your template. Clear the existing data and upload the file we just created.
3. Set up your Data so X-values = politician, Y-values = word, Size = count and Color = politician.
4. Remove the Legend and the label in the X axis (Axis Title = Custom).
5. Change the label in the X axis to a 0° angle and change the font to bold. Do the same in the Y axis, if needed.



Today's Learning Outcomes

By now, you should feel (a little bit) more comfortable with...

1. ... listing different ways in which **text data can be incorporated in news stories**.
2. ... applying a **simple workflow** (gather > clean > analyze > visualize) in the use of text data for data journalism.
3. ... creating **visualizations from text data** using widely adopted web-based tools.