Text Mining the Federal Budget

Janet Carson

Who might want to use this?

Officials and Staffers

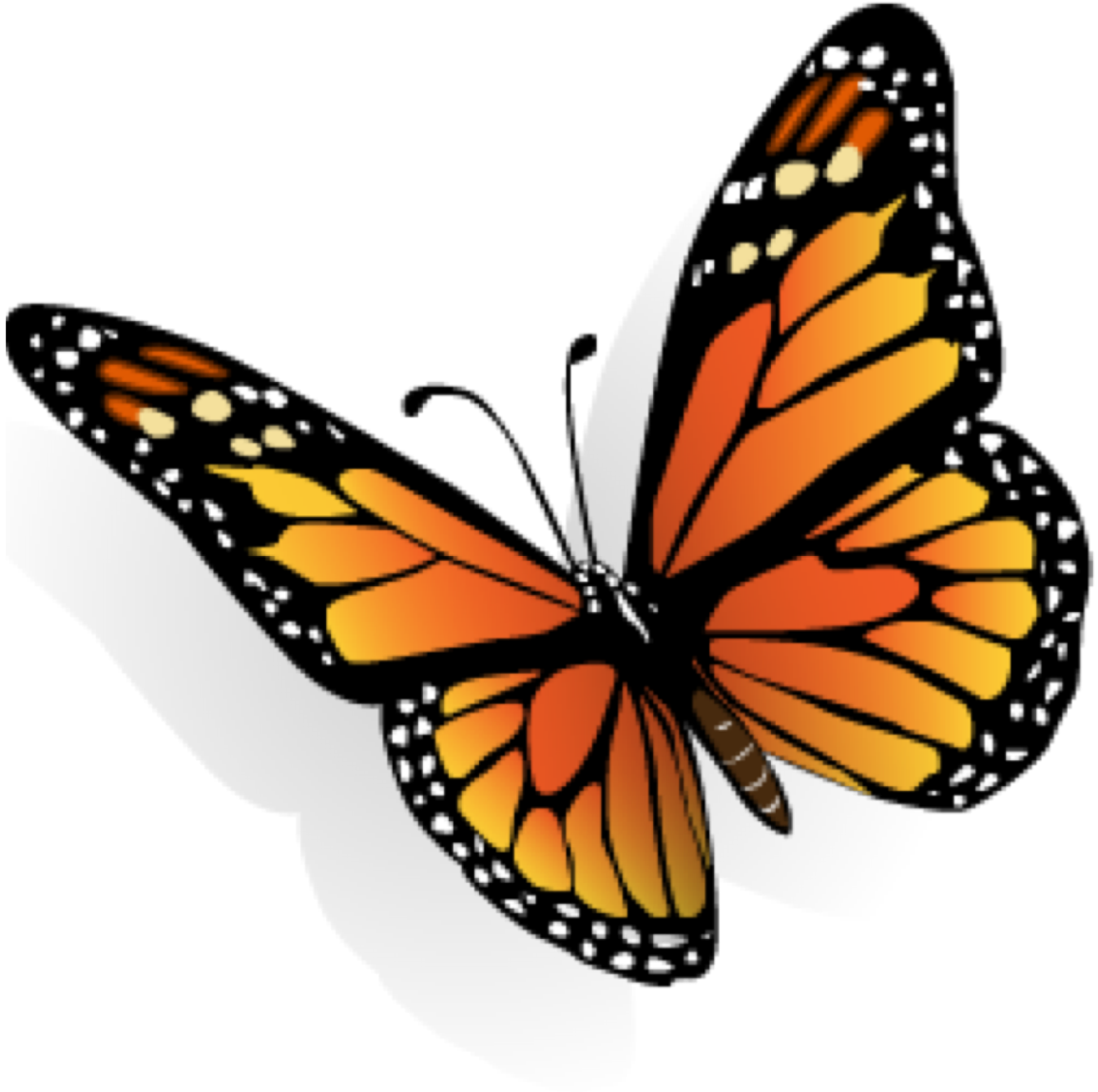
Journalists

Citizens and Activists

Lobbyists and Donors

231. None of the funds made available by this Act or prior Acts are available for the construction of **pedestrian fencing**—(1) within the Santa Ana Wildlife Refuge; (2) within the Bentsen-Rio Grande Valley State Park; (3) within La Lomita Historical park; (4) within the **National Butterfly Center**; or (5) within or east of the Vista del Mar Ranch tract of the Lower Rio Grande Valley National Wildlife Refuge.

How often does a word like ‘butterfly’ appear in the federal budget?



Does anyone really read this?

- 465 page PDF
- BeautifulSoup
- Gensim
- 1248 Documents
- 207416 Words

What's in an entry in the Federal budget?

Numbers

Stop Words

Budget and legal words

Unique words

A simple bag of words model is not suited to analyzing the budget

- Word similarity using Glove and Word2Vec

Clustering Words with Glove

Public Health

[‘outbreak', 'avian', 'virus', 'tuberculosis', 'polio', 'malaria', 'illness', 'epidemic', 'infection’]

???

['redesignate', 'disaggregate', 'adulterate', 'securitize', 'prorate', 'subleasing', 'urbanize', 'repurpose’]

Butterfly*

['species', 'bird', 'wild', 'olympic', 'cat', 'holder', 'turtle', 'paralympic', 'event’]

Word Mover Distance

Obama speaks to the media in Illinois.

The President greets the press in Chicago.

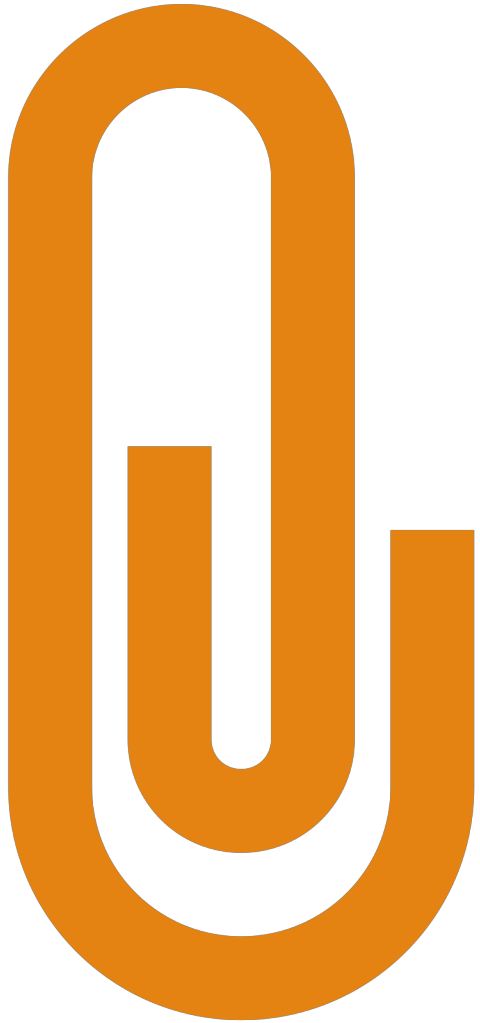
Results were…

- I ran it on 950 items, length 10 to 100 words
- Pulled out cut-n-paste and similarly structured legalese, and left me with 830 outliers…

- Because the entire budget is outliers.

Thank You

Appendix



Most of the dataset is boilerplate

Repairs and Restoration

For the repair, alteration, and improvement of archives facilities, and to provide adequate storage for holdings, $7,500,000, to remain available until expended.

Section 7034 of Special Provisions for the State Department goes on for six pages

1600 Words Appear In Exactly One Section of the Federal Budget

- Moderately
- Nonreimbursable
- Redesignation
- Postage
- Supreme
- Biosimilar
- Centrocercus
- Cylinder
- Beginning
- Butterfly
- Cambodia
- Grape
- Jupiter
- Yacht
- Incinerator
- Reef
- Coin
- Expeditious

Almost half the entries have a word not found elsewhere in the budget

How to define an outlier?

Repeated uncommon words

None of the funds provided by this Act shall be available to promote the sale or export of **tobacco** or **tobacco** products, or to seek the reduction or removal by any foreign country of restrictions on the marketing of **tobacco** or **tobacco** products, except for restrictions which are not applied equally to all **tobacco** or **tobacco** products of the same type

Multiple uncommon words

For acquisition of lands within the exterior boundaries of the **Cache**, **Uinta**, and **Wasatch** National Forests, **Utah**; the **Toiyabe** National Forest, Nevada; and the Angeles, San **Bernardino**, **Sequoia**, and **Cleveland** National Forests, California; and the **Ozark**-St. **Francis** and **Ouachita** National Forests, **Arkansas**....