Natural Language Processing

The Choice Between an Open Internet and a Sovereign One

By Fahmida Y Rashid

Classifying the Digital Deciders

- There are two visions of the Internet, a global and open Internet and a national, “sovereign” Internet
  - Global and open: all sites are available to users,
    - a free flow of communication and ideas.
  - National and sovereign: each country decides what it wants available to its population.
    - Often cited as a way to monitor users, suppress unpopular ideas
    - Can also be economic protectionism, letting home-grown technologies thrive without competing with others.

Research Question

- Assumption #1: Countries that have not yet said which Internet they favor are equally likely to say one or the other.
- Assumption #2: Countries may pick a position based on what their like-minded peers say. (Be like your friends!)
- **Given these assumptions,** can we look at what countries say to determine if they have more in common with countries that favor an open Internet or with countries that favor a sovereign internet?

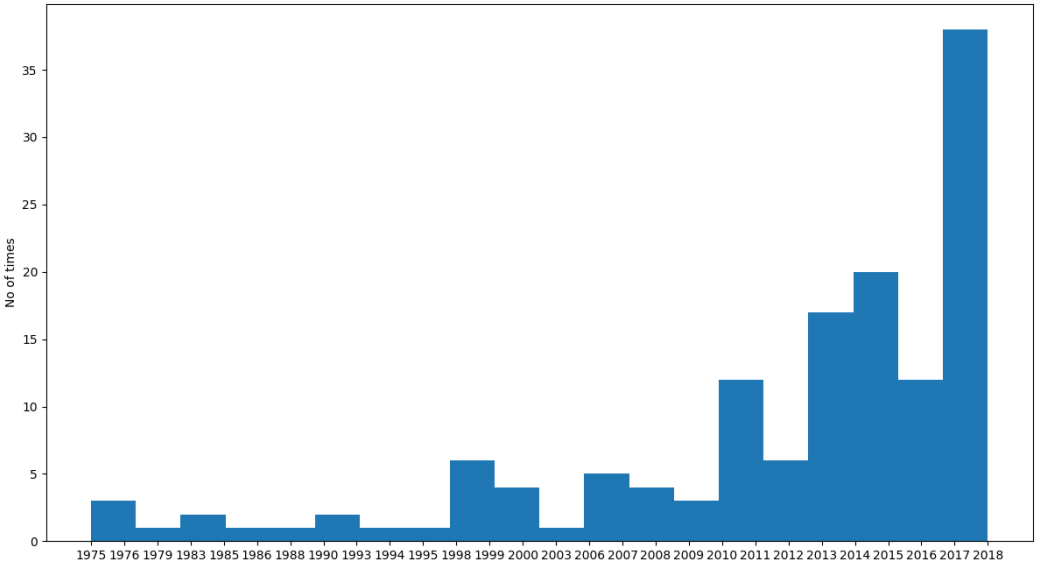
Where the Data Came From

- Every year, when the United Nations General Assembly meets, there is a General Debate, where countries can give a speech about anything that is important to them.
- The “state of the nation” speech can cover a range of topics.
- Harvard Dataverse has the text of every General Debate speech from 1970 to 2018.

Citation: Alexander Baturo, Niheer Dasandi, and Slava Mikhaylov, "Understanding State Preferences With Text As Data: Introducing the UN General Debate Corpus" Research & Politics, 2017

Data Summary

- Cybersecurity speeches cluster towards the end of the timeframe covered by the data. Most speeches that contain cybersecurity-related terms were in the last five years.
- “Security” results in a lot of false matches, because the UN associates security with political, social, and economic topics, too.



Methodology

- Create a corpus of cybersecurity speeches, economics, and political speeches
- TF-IDF to see how much of the speech has relevant terms.
- Comparison analysis with cosine similarity
- Creating a classifier

- About 46 speeches for cybersecurity were from “open” countries, to 23 from “sovereign.”
- I kept the cosine similarity scores by speech – and then looked at whether there were consistently leaning one way or the other.

Analyzing the Data

- Even the cybersecurity speeches did not have high TF-IDF scores for cybersecurity terms. (this makes sense since it is a state of the nation speech)
- When looking at the individual speeches, there were no real differences, but when rolled up into country, there were clear indications of being one way than others.
- The economics speeches gave a stronger sense of lean with the cosine similarity than the cybersecurity ones did. (this also makes sense)

Classifying the Digital Deciders

- I pulled out about 10 percent of the speeches I had flagged in the corpus for the classifier for my “test” data.
- It became a game of “guess the opposite.” If the classifier said sovereign, it was more likely to be open.
- I ran the “deciders” file, and the break down (about 2/3 as open, 1/3 as sovereign) is pretty comparable to what the training and test data looks like, but I don’t trust the classifier.

Observations and Notes

- The classifier was tripped up by the fact that speeches were so similar to each other. I should have trimmed the corpus down to individual sentences, not speeches.
- The ngram of (1,3) didn’t give better results than (1,2) for the cybersecurity speeches. That makes sense since there aren’t many 3-term phrases in this field.
- The cosine similarity analysis showed very small numbers, but you could see some “lean” when the speeches were rolled up into the overall country.