Format of Talk

Average Retention Rates

- 5%  Lecture
- 10%  Reading
- **20%**  Audio-Visual
- 30%  Demonstration
- 50%  Group Discussion
- 75%  Practice
- 90%  Teaching Others

Passive Teaching Methods

Participatory Teaching Methods

*Adapted from National Training Laboratories. Bethel, Maine
Format of Talk

Average Retention Rates

5% Lecture
10% Reading
20% Audio-Visual
30% Demonstration
50% Group Discussion
75% Practice
90% Teaching Others

*Adapted from National Training Laboratories. Bethel, Maine

Figure 2:
Why, What, and How

Presidential Executive Order on Reducing Regulation and Controlling Regulatory Costs (EO 13771) requires that:

*[For] every one new regulation issued, at least two prior regulations be identified for elimination...*

- I want to enable regulators to be better rule-makers and rule-managers.
- Dynamic data visualization how I intend to manage the CFR’s volume.
- My automatic CFR aggregator algorithm simplifies the identification of *statistically similar texts.*
The Code of Federal Regulations (CFR)

- The CFR was about 50k pages and required a little more than a year to read in 1970.\(^1\)
- Today, it is 180k+ pages and requires almost 4 years to read.

\(^1\)According to Dr. Patrick McLaughlin.
CFR Organization (i.e. the technical challenge)

Each of the 49 CFR Titles is (pseudo) organized by

- Volume
  - Subtitle
    - Chapter

Each Chapter is organized by

- Subchapter
  - Part
    - Subpart

Each Subpart is organized by

- Subject
  - Section
Additional Uses

Advocacy leadership suggested I consider helping small businesses to identify the regulations most relevant to them.²

I’ve opened the interface up to any uploads as well as copy and pasted text.

- Small business proposals/plans
- Proposed regulations
- Popular media/transcripts

²Major Clark
Text Analysis Preliminaries

Text analysis involves preprocessing, classification, clustering, information extraction, and visualization.

1. After assembling a corpus, the first step is to remove stop words (and, the, of) and other words that convey little topic distinctivity (such as therefore, next, however, etc.).
2. The next step is to stem the vocabulary, i.e. trim them to their roots so that words with the same root are combined (fight, fights, fought).
3. The third step is to create n-grams, which are sets of words that co-occur impossibly often, and thus denote a single idea (White House, Supreme Court, etc.).

---

3 See Olga Scrivner, David Banks, Julia Silge, and Simone Teufel for great introductory material drawn upon here.
Zipf’s Law

The $i^{th}$ most frequent term in a given language has frequency proportional to $\frac{1}{i}$:

So if the most frequent term *the* occurs 100 times, then the second most frequent term *of* should appear 50 times, and *and* 25 times, etc.

The rule supposedly applies to

- Notes in musical performances
- Frequency of access to web pages
- Income distributions amongst top earning 3% individuals
- Korean family names
- Size of earth quakes
- Word senses per word

While the law doesn’t seem to hold very well for many English texts, it works quite well for regulatory texts.
Importance

The purpose of visualization is insight, not pictures.

Ben Schneidermen, 1999

Figure 3:
Importance is measured by the tf-idf

Suppose you are interested in CFR entries related to the word is prescription due to a new regulation you are considering.

- \( tf \) measures the number of times \( prescription \) appears in a given document.
- \( df \) measures the number of times it across the corpus (i.e. all documents).
- \( N \) is the number of unique words in the corpus.
- \( idf = \frac{N}{df} \) is the inverse document frequency.
- \( tf \times idf \) thus measures the importance of a word to in characterizing a given document.
Cosine Distance

Figure 4:
Cosine Distance

Figure 5:
Demonstration
Parallel processing is straightforward to implement and my tests show that it already promises about a 7 fold speed improvement. Unfortunately it wasn’t ready for today’s talk.

GPU processing could realistically increase processing speeds by 1000 fold although it is much more difficult to implement.

Topic modelling using LSA/LDA is implemented at the command-line level but I’m still working on making the topic descriptions either graphically or semantically interesting.\(^4\)

Sentiment analysis graphs are already programmed in, I just didn’t have time to polish the visualizations.

Add panning, zooming, and section selection to the collapsible trees.

\(^4\)I’d like to generate simple sentences that highlight the highly weighted words.
Thank you!