CFR Miner

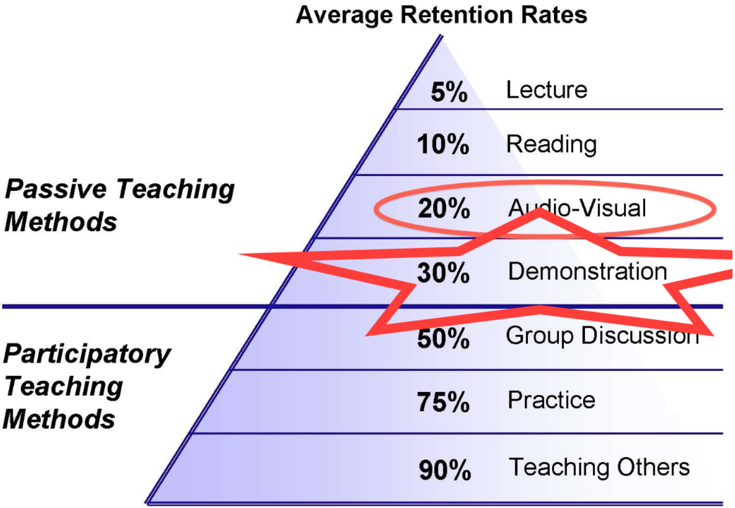
R. Schwinn

11/2/2017

Modifications

1. Mention the problem of granularity and the automated aggregator.
2. Describe work with the OIRA / regulatory group
3. Describe NIST and ITC programs
4. Integrate language complexity measures from

Format



*Adapted from National Training Laboratories. Bethel, Maine

Figure 1:

Why

1. Navigation of the CFR can be difficult for the general public and for agencies[1]
2. Regulatory reform can be burdensome to individuals, businesses, and AGENCIES
   - Presidential Executive Order on Reducing Regulation and Controlling Regulatory Costs (EO 13771) requires that,“[For] every new regulation issued, at least two prior regulations are to be identified for elimination…”
   - Section 6 of Executive Order 13563 requires agencies conduct periodic, retrospective review and analysis of existing regulations that “may be outmoded, ineffective, insufficient, or excessively burdensome, and to modify, streamline, expand, or repeal them … so as to make the agency’s regulatory program more effective and less burdensome in achieving regulatory objectives.”[2]

[1] especially in regard to old (>10yr), duplicative rules from other agencies.

[2] See Amy Bunk’s Drafting and Updating FOIA Regulations

Why

Numerous statutes apply. For example, Title 5, Section 610 requires periodic review of rules:

- Agencies must provide (publish) a plan for the periodic review of rules issued by the agency which have a significant economic impact upon a substantial number of small entities
    - The purpose of the plan is to identify which rules
        - should be continued without change
        - should be amended
        - should be rescinded
    - Review should include all rules promulgated within the previous 10 years

Additional Uses

Advocacy leadership suggested we consider ways to help small businesses identify the regulations most relevant to them. The interface is open to uploads and copy and pasted text.

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

How

- Natural language processing
- Dynamic data visualizations
- Automatic CFR aggregator

Demonstrations

- *Sally the Rule Writer* is tasked with reviewing Title 1, Part 457.[3] She would like to search for related rules promulgated by other agencies.[4]
- *Ralph the Researcher* is labor economist who would like to search the IRS code for information on the treatment of earnings of workers with and without dependents.
- *Contessa the Small Business Owner* would like to identify the CFR sections most related to the topics found in her business plan.

[3] Volume 1, Chapter 4...

[4] Names borrowed from W. Liberante.

Text Analysis[5]

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

1. After assembling a corpus, the first step 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 improbably often, and thus denote a single idea (White House, Supreme Court, etc.).

[5] See Olga Scrivner, David Banks, Julia Silge, and Simone Teufel for great introductory material.

Importance

The purpose of visualization is insight, not pictures.
-Ben Schneidermen, 1999

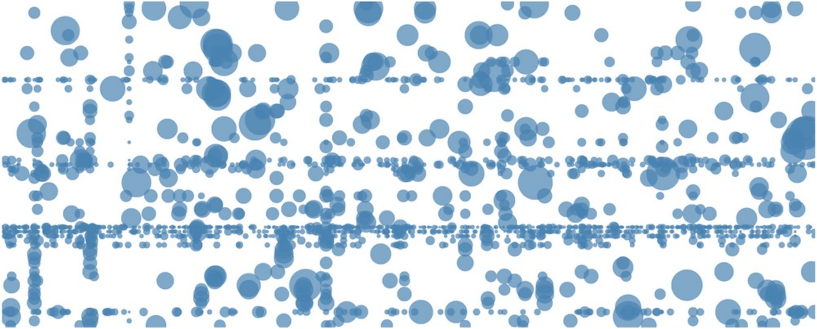


Figure 2:

Importance (θ)

Importance is measured by the tf-idf. Suppose you are interested in CFR entries related to the word *prescription*:

- $tf_{i,j}$ measures the number of times term i appears in document j. e.g. $tf_{prescription,4} = 10$ means that the word *prescription* appears in document 4 on 10 occasions.

Importance (θ)

Importance is measured by the θ or $tf * idf$. Suppose you are interested in CFR entries related to the word *prescription*:

- $tf_{i,j}$ measures the number of times term i appears in document j. e.g. $tf_{prescription,4} = 10$ means that the word *prescription* appears in document 4 on 10 occasions.
- df_i measures the percent of documents within which the word i appears across the entire corpus (i.e. all documents) of N documents. e.g. $df_{prescription} = 0.25$ means that the term *prescription* appears in 25% of the documents.

Importance (θ)

Importance is measured by the tf-idf. Suppose you are interested in CFR entries related to the word *prescription*:

- $tf_{i,j}$ measures the number of times term i appears in document j. e.g. $tf_{prescription,4} = 10$ means that the word *prescription* appears in document 4 on 10 occasions.
- df_i measures the percent of documents within which the word i appears across the entire corpus (i.e. all documents) of N documents. e.g. $df_{prescription} = 0.25$ means that the term *prescription* appears in 25% of the documents.
- $idf_i = \frac{1}{df}$ is the inverse document frequency. By inverting the percentage, weight is added to words that appear infrequently. e.g. $df_i = 0.25$ means that the tf is weighted by 4.On the other hand, if $df_i = 1$, and the term appears in all documents, the tf is weighted by 1.

Importance (θ)

Importance is measured by the tf-idf. Suppose you are interested in CFR entries related to the word *prescription*:

- $tf_{i,j}$ measures the number of times term i appears in document j. e.g. $tf_{prescription,4} = 10$ means that the word *prescription* appears in document 4 on 10 occasions.
- df_i measures the percent of documents within which the word i appears across the entire corpus (i.e. all documents) of N documents. e.g. $df_{prescription} = 0.25$ means that the term *prescription* appears in 25% of the documents.
- $idf_i = \frac{1}{df}$ is the inverse document frequency. By inverting the percentage, weight is added to words that appear infrequently. e.g. $df_i = 0.25$ means that the tf is weighted by 4.On the other hand, if $df_i = 1$, and the term appears in all documents, the tf is weighted by 1.
- $\theta_{i,j} = \frac{tf_{i,j}}{df_i} = tf_{i,j} \times idf_i$ thus measures the importance of a word in characterizing a given document. e.g. $\theta_{prescription,4} = 40$.

Importance (θ)

- $\theta_{i,j} = \frac{tf_{i,j}}{df_i}$ provides measure of the importance of a word in characterizing a given document within a corpus. e.g. $\theta_{prescription,4} = 40$.

Notice that if a word appears in every document, such as *the*, then the *idf* term does little to increase its relative importance. On the other hand, if the term appears in only a small percent of documents, then it's importance is magnified.

Zipf's Law

- The i$^{\text{th}}$ 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 among 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.

Cosine Distance

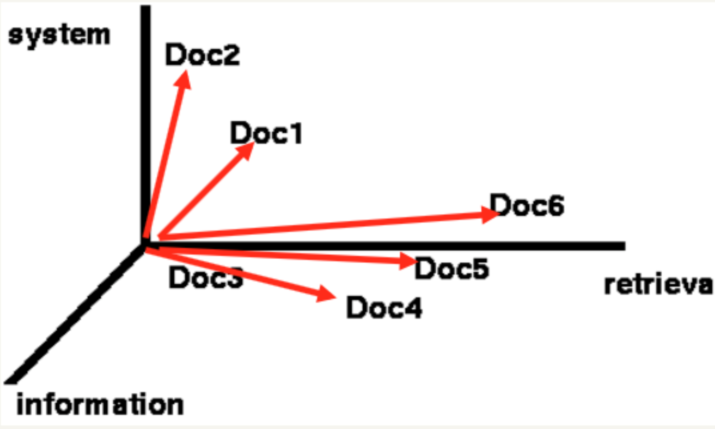


Figure 3:

Cosine Distance

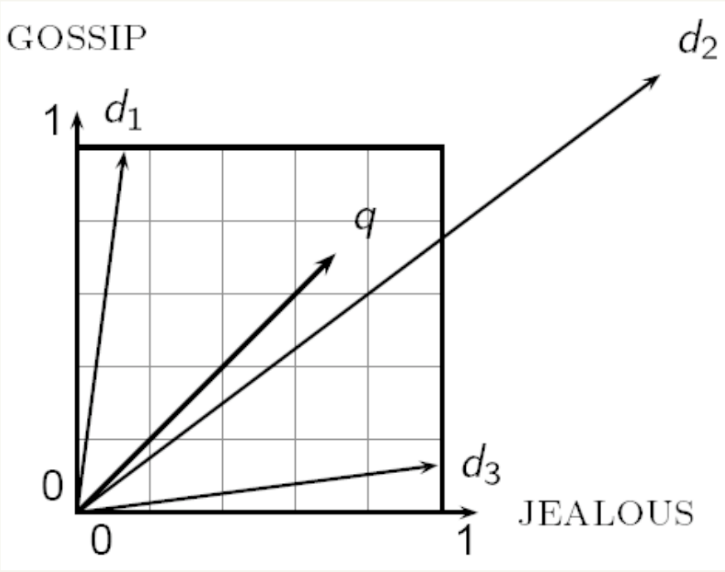


Figure 4:

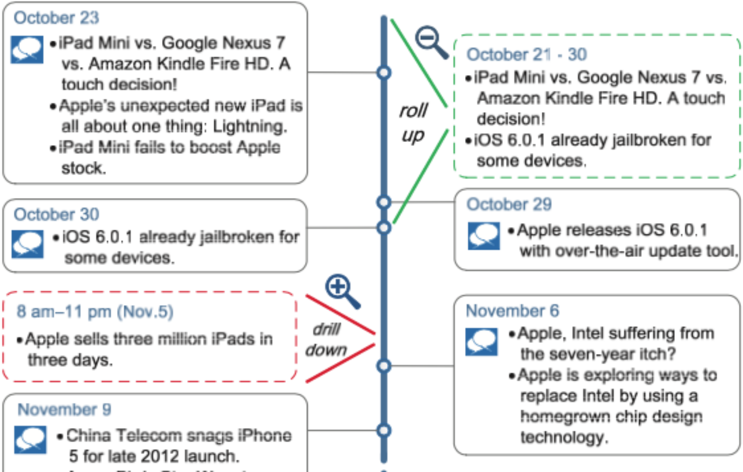
Next Steps Short List

- Parallel processing is straight forward and basic tests show that it already promises a 7-fold speed improvement.
- GPU processing could realistically increase processing speeds 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.[6]
- Sentiment analysis plots are implemented but not obviously useful.

[6] I'd like to generate simple sentences that highlight the highly weighted words.

Ideas for the future

Wendy mentioned a project that actively identifies changes to the CFR over time. Applying recent text summarization algorithms could provide interesting insights into agency activity over time.[7]



October 23

- iPad Mini vs. Google Nexus 7 vs. Amazon Kindle Fire HD. A touch decision!
- Apple's unexpected new iPad is all about one thing: Lightning.
- iPad Mini fails to boost Apple stock.

October 30

- iOS 6.0.1 already jailbroken for some devices.

8 am–11 pm (Nov.5)

- Apple sells three million iPads in three days.

November 9

- China Telecom snags iPhone 5 for late 2012 launch.

October 21 - 30

- iPad Mini vs. Google Nexus 7 vs. Amazon Kindle Fire HD. A touch decision!
- iOS 6.0.1 already jailbroken for some devices.

October 29

- Apple releases iOS 6.0.1 with over-the-air update tool.

November 6

- Apple, Intel suffering from the seven-year itch?
- Apple is exploring ways to replace Intel by using a homegrown chip design technology.

Thank you![8]

[8] Ask about DocketWrench

The Code of Federal Regulations (CFR)

- The CFR was about 50,000 pages and required a little more than a year to read in 1970.[9]
- Today, it is 180,000+ pages and requires almost 4 years to read.

[9] 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