Day 3: Quantitative methods for comparing texts

Kenneth Benoit

Quantitative Analysis of Textual Data

October 7, 2014

Sampling issues in existing measures

- Lexical diversity measures may take sample frames, or moving windows, and average across the windows
- Readability may take a sample, or multiple samples, to compute readability measures
- But rather than simulating the "sampling distribution" of a statistic, these are more designed to:
  - get a representative value for the text as a whole
  - normalize the length of the text relative to other texts

Bootstrapping text-based statistics

Simulation and bootstrapping

Used for:

- Gaining intuition about distributions and sampling
- Providing distributional information not distributions are not directly known, or cannot be assumed
- Acquiring uncertainty estimates

Both simulation and bootstrapping are numerical approximations of the quantities we are interested in. (Run the same code twice, and you get different answers)

Solution for replication: save the seed

Quantifying Uncertainty

- Critical if we really want to compare texts
- Question: How?
  - Make parametric assumptions about the data-generating process. For instance, we could model feature counts according to a Poisson distribution.
  - Use a sampling procedure and obtain averages from the samples. For instance we could sample 100-word sequences, compute reliability, and look at the spread of the readability measures from the samples
  - Bootstrapping: a generalized resampling method

Bootstrapping

- *Bootstrapping* refers to repeated resampling of data points with replacement
- Used to estimate the error variance (i.e. the standard error) of an estimate when the sampling distribution is unknown (or cannot be safely assumed)
- Robust in the absence of parametric assumptions
- Useful for some quantities for which there is no known sampling distribution, such as computing the standard error of a median

Bootstrapping illustrated

```
> ## illustrate bootstrap sampling
> set.seed(30092014)   # set the seed so that your results will match m
> # using sample to generate a permutation of the sequence 1:10
> sample(10)
 [1]  4  2  1  9  8  5  7  3  6 10
> # bootstrap sample from the same sequence
> sample(10, replace=T)
 [1] 8 6 6 2 5 8 4 8 4 9
> # boostrap sample from the same sequence with probabilities that
> # favor the numbers 1-5
> prob1 <- c(rep(.15, 5), rep(.05, 5))
> prob1
 [1] 0.15 0.15 0.15 0.15 0.15 0.05 0.05 0.05 0.05 0.05
> sample(10, replace=T, prob=prob1)
 [1] 4 1 1 2 8 3 1 6 1 9
```

Bootstrapping the standard error of the median

Using a user-defined function:

```
b.median <- function(data, n) {
    resamples <- lapply(1:n, function(i) sample(data, replace=T))
    sapply(resamples, median)
    std.err <- sqrt(var(r.median))
    list(std.err=std.err, resamples=resamples, medians=r.median)
}
summary(b.median(spending, 10))
summary(b.median(spending, 100))
summary(b.median(spending, 400))
median(spending)
```

Bootstrapping the standard error of the median

Using R's **boot** library:

```
library(boot)
samplemedian <- function(x, d) return(median(x[d]))
quantile(boot(spending, samplemedian, R=10)$t, c(.025, .5, .975))
quantile(boot(spending, samplemedian, R=100)$t, c(.025, .5, .975))
quantile(boot(spending, samplemedian, R=400)$t, c(.025, .5, .975))
```

Note: There is a good reference on using `boot()` from `http://www.mayin.org/ajayshah/KB/R/documents/boot.html`

Bootstrapping methods for textual data

- Question: what is the "sampling distribution" of a text-based statistic? Examples:
  - a term's (relative) frequency
  - lexical diversity
  - complexity
- Could use to compare subgroups, analagous to ANOVA or t-tests, for statistics such as similarity (below)

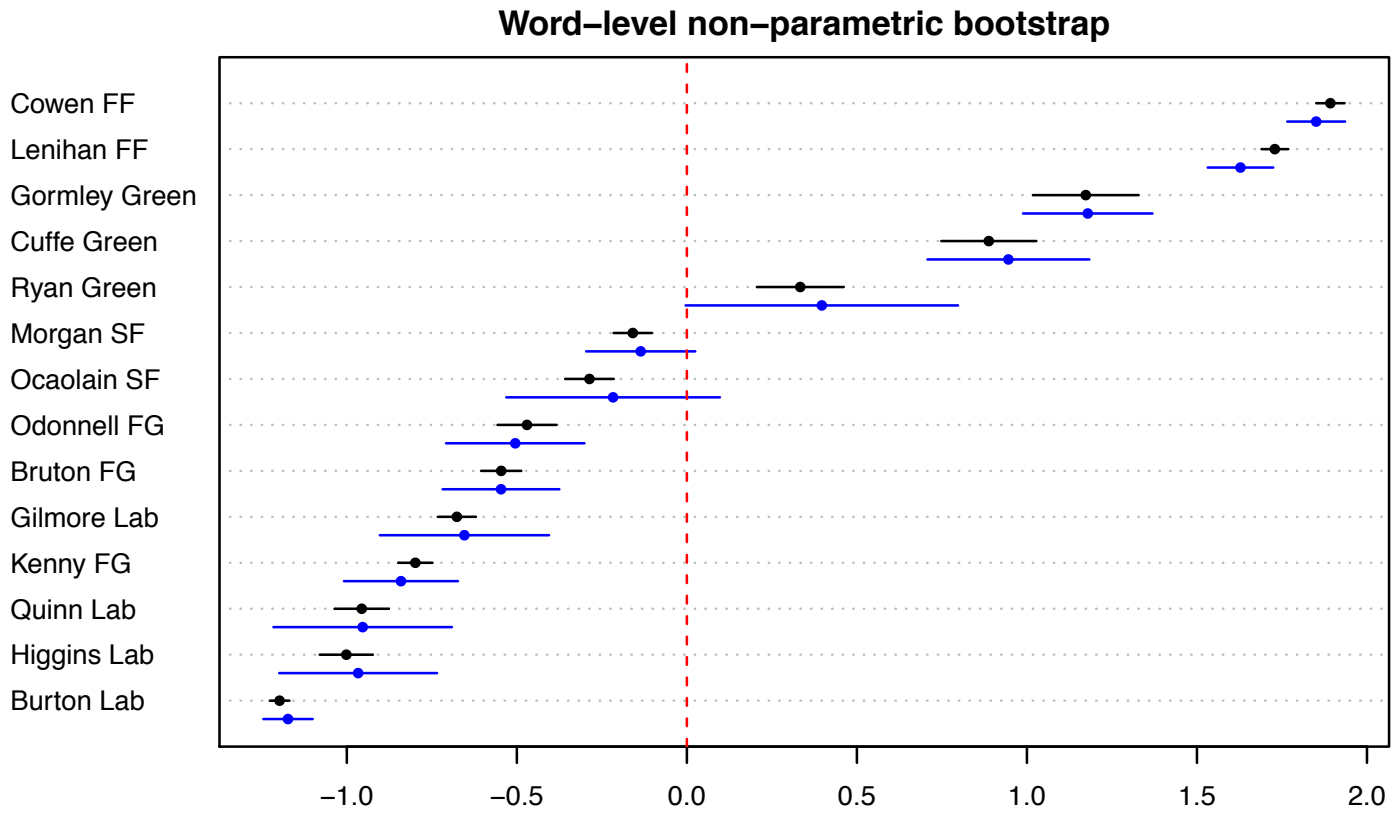
Guidelines for bootstrapping text

- Bootstrap by resampling tokens.
  Advantage: This is easily done from the document-feature matrix.
  Disadvantage: Ignores the natural units into which text is grouped, such as sentences
- Bootstrap by resampling sentences.
  Advantage: Produces more meaningful (potentially readable) texts, more faithful to data-generating process.
  Disadvantage: More complicated, cannot be done from dfm, must segment the text into sentences and construct a new dfm for each resample.

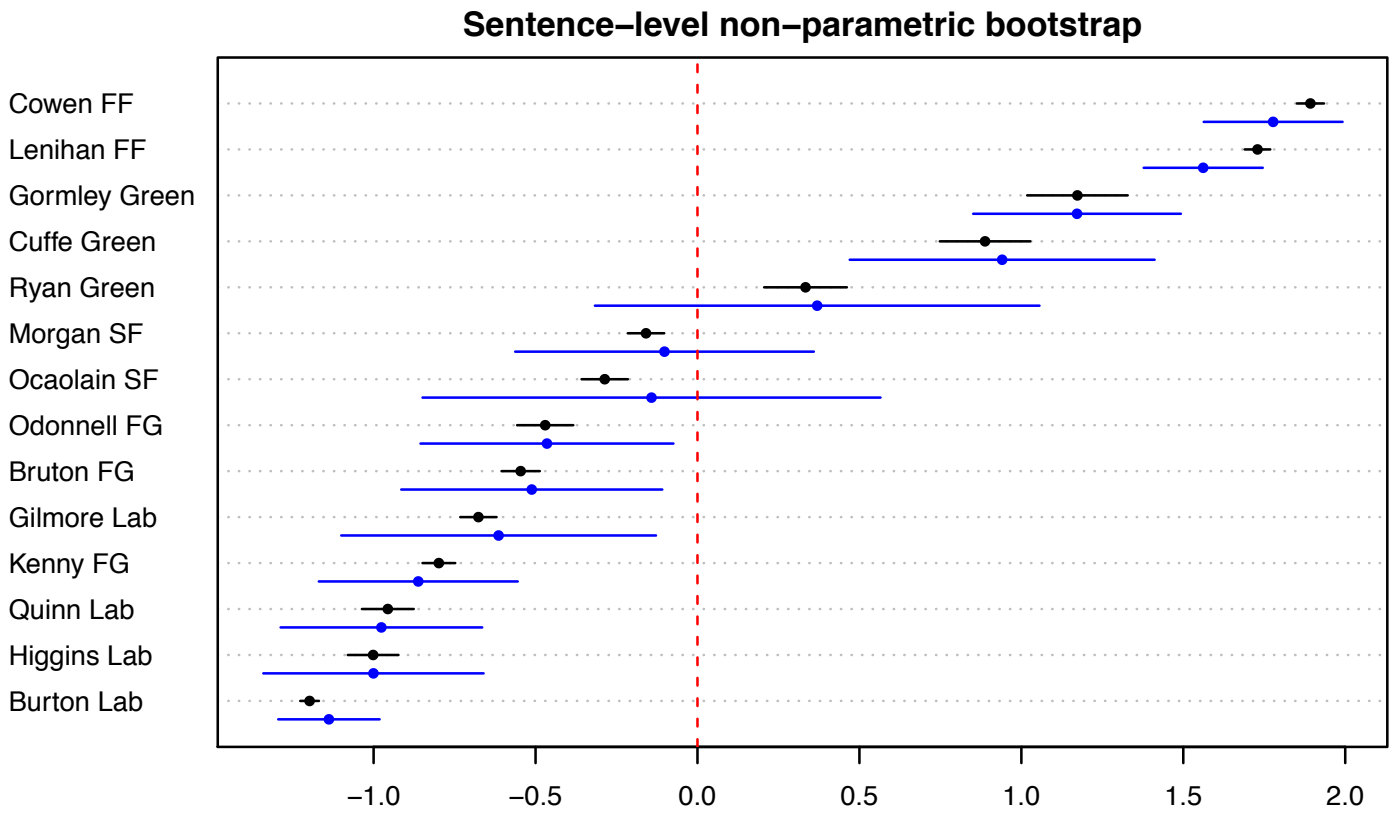
Guidelines for bootstrapping text (cont)

- Other options for bootstrapping text: generalize the notion of the "block" bootstrap. In block bootstrap, consecutive blocks of observations of length K are resampled from the original time series, either in fixed blocks (Carlstein, 1986) or overlapping blocks (Künsch, 1989)
  - paragraphs
  - pages
  - chapters
  - stratified: words within sentences or paragraphs

Different bootstrapping methods: example



Different bootstrapping methods: example



Different bootstrapping methods: example

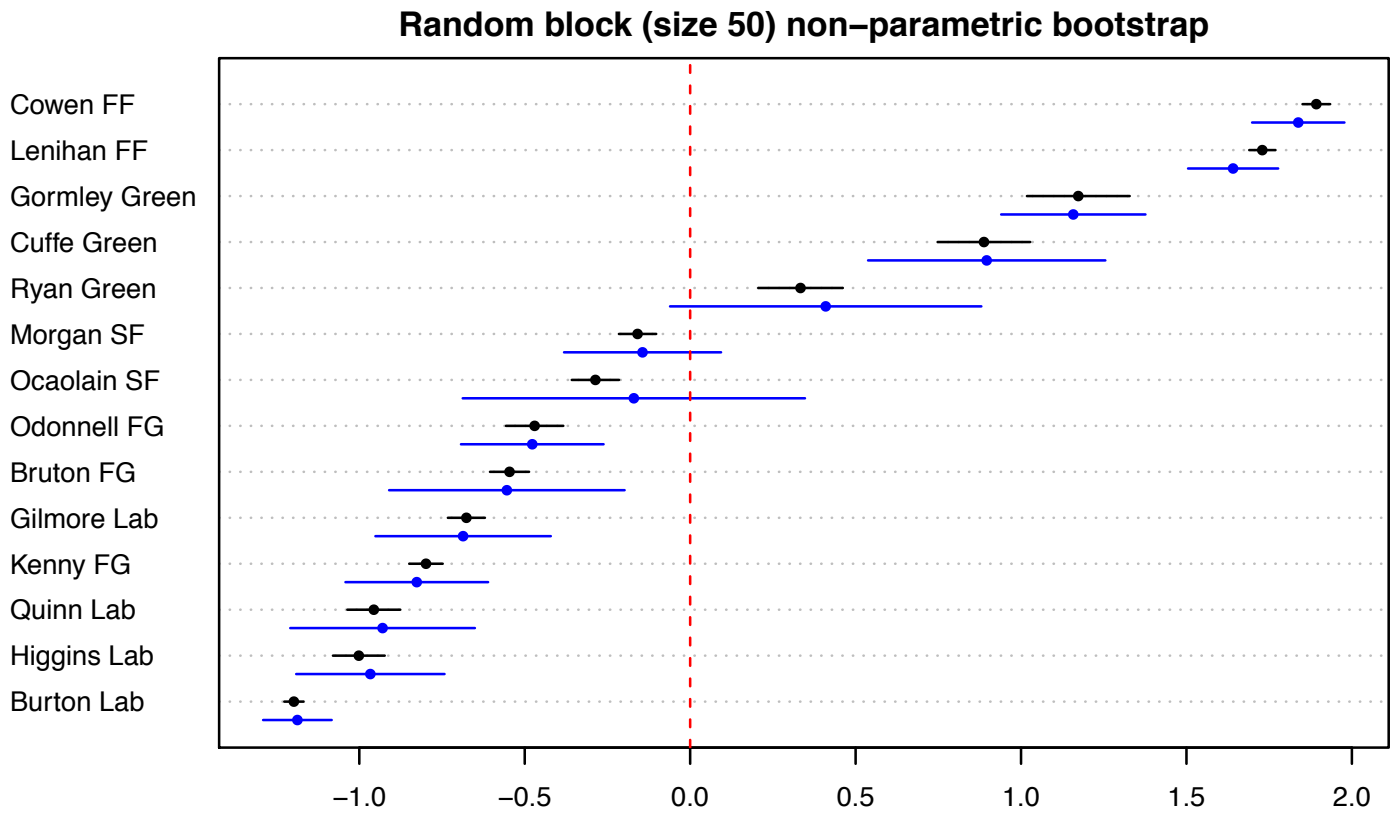


Figure 3: *Estimates of θ_i and 95% confidence intervals from the 2009 Irish Budget Debates using Block Bootstrapping.* Black points and lines are analytical SEs; blue point estimates and 95% CIs correspond to the labelled method.

Documents as vectors

- The idea is that (weighted) features form a vector for each document, and that these vectors can be judged using metrics of similarity
- A document's vector for us is simply (for us) the row of the document-feature matrix

Characteristics of similarity measures

Let A and B be any two documents in a set and $d(A, B)$ be the distance between A and B.

1. $d(x, y) \geq 0$ (the distance between any two points must be non-negative)
2. $d(A, B) = 0$ iff $A = B$ (the distance between two documents must be zero if and only if the two objects are identical)
3. $d(A, B) = d(B, A)$ (distance must be symmetric: A to B is the same distance as from B to A)
4. $d(A, C) \leq d(A, B) + d(B, C)$ (the measure must satisfy the triangle inequality)

Euclidean distance

Between document A and B where j indexes their features, where y_{ij} is the value for feature j of document i

- Euclidean distance is based on the Pythagorean theorem
- Formula

$$\sqrt{\sum_{j=1}^{j}(y_{Aj} - y_{Bj})^2} \tag{1}$$

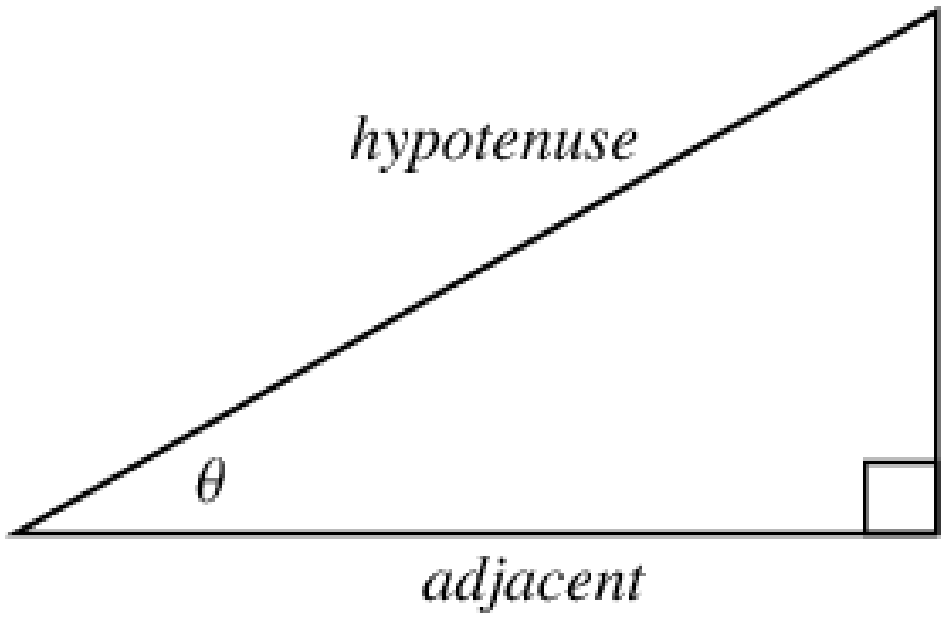
- In vector notation:

$$\|\mathbf{y}_A - \mathbf{y}_B\| \tag{2}$$

- Can be performed for any number of features J (or V as the vocabulary size is sometimes called – the number of columns in of the dfm, same as the number of feature types in the corpus)

A geometric interpretation of "distance"

In a right angled triangle, the cosine of an angle θ or $\cos(\theta)$ is the length of the adjacent side divided by the length of the hypotenuse



We can use the vectors to represent the text location in a V-dimensional vector space and compute the angles between them

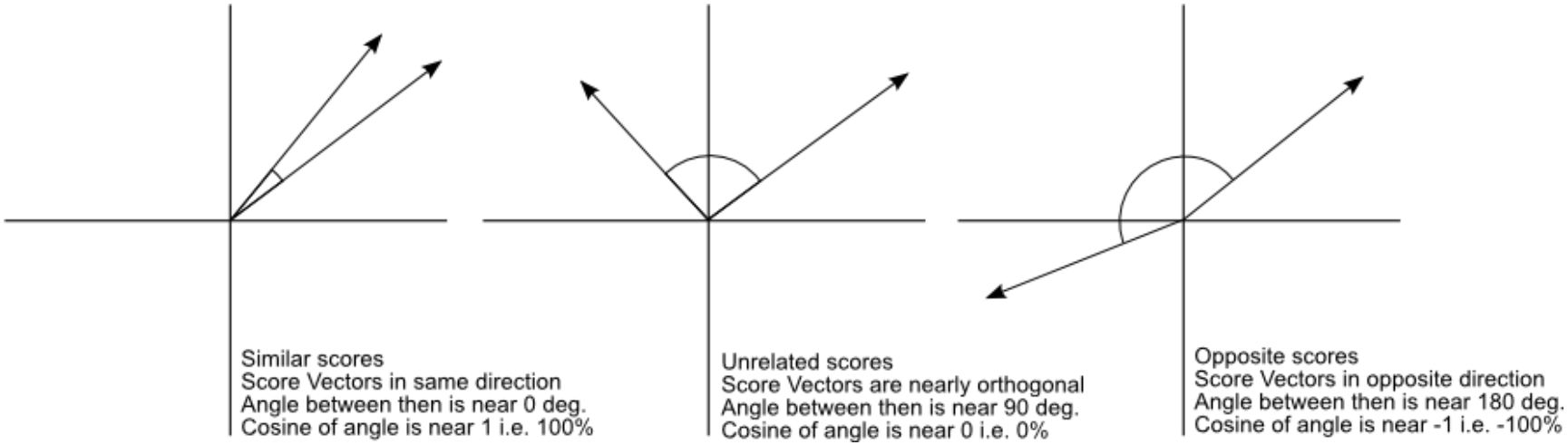
Cosine similarity

- Cosine distance is based on the size of the angle between the vectors
- Formula

$$\frac{\mathbf{y}_A \cdot \mathbf{y}_B}{\|\mathbf{y}_A\| \|\mathbf{y}_B\|} \tag{3}$$

- The $\cdot$ operator is the dot product, or $\sum_j y_{Aj} y_{Bj}$
- The $\|\mathbf{y}_A\|$ is the vector norm of the (vector of) features vector $\mathbf{y}$ for document A, such that $\|\mathbf{y}_A\| = \sqrt{\sum_j y_{Aj}^2}$
- Nice property: independent of document length, because it deals only with the angle of the vectors
- Ranges from -1.0 to 1.0 for term frequencies, or 0 to 1.0 for normalized term frequencies (or tf-idf)

Cosine similarity illustrated



Example text

Hurricane Gilbert swept toward the Dominican Republic Sunday , and the Civil Defense alerted its heavily populated south coast to prepare for high **winds**, heavy **rains** and high seas.

The **storm** was approaching from the southeast with sustained **winds** of 75 mph gusting to 92 mph .

"There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly before midnight Saturday .

Cabral said residents of the province of Barahona should closely follow **Gilbert** 's movement .

An estimated 100,000 people live in the province, including 70,000 in the city of Barahona , about 125 miles west of Santo Domingo .

Tropical **Storm Gilbert** formed in the eastern Caribbean and strengthened into a **hurricane** Saturday night

The National **Hurricane** Center in Miami reported its position at 2a.m. Sunday at latitude 16.1 north , longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.

The National Weather Service in San Juan , Puerto Rico , said **Gilbert** was moving westward at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the **storm**.

The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6p.m. Sunday.

Strong **winds** associated with the **Gilbert** brought coastal flooding , strong southeast **winds** and up to 12 feet to Puerto Rico 's south coast.

Example text: selected terms

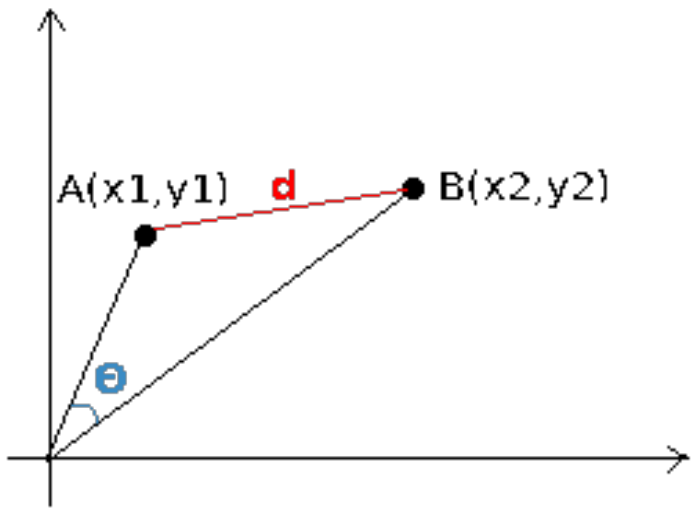
- Document 1
  Gilbert: 3, hurricane: 2, rains: 1, storm: 2, winds: 2
- Document 2
  Gilbert: 2, hurricane: 1, rains: 0, storm: 1, winds: 2

Example text: cosine similarity in R

```
> toyDfm <- matrix(c(3,2,1,2,2, 2,1,0,1,2), nrow=2, byrow=TRUE)
> colnames(toyDfm) <- c("Gilbert", "hurricane", "rain", "storm", "winds")
> rownames(toyDfm) <- c("doc1", "doc2")
> toyDfm
     Gilbert hurricane rain storm winds
doc1       3         2    1     2     2
doc2       2         1    0     1     2
> simil(toyDfm, "cosine")
          doc1
doc2 0.9438798
```

Relationship to Euclidean distance

- Cosine similarity measures the similarity of vectors with respect to the origin
- Euclidean distance measures the distance between particular points of interest along the vector



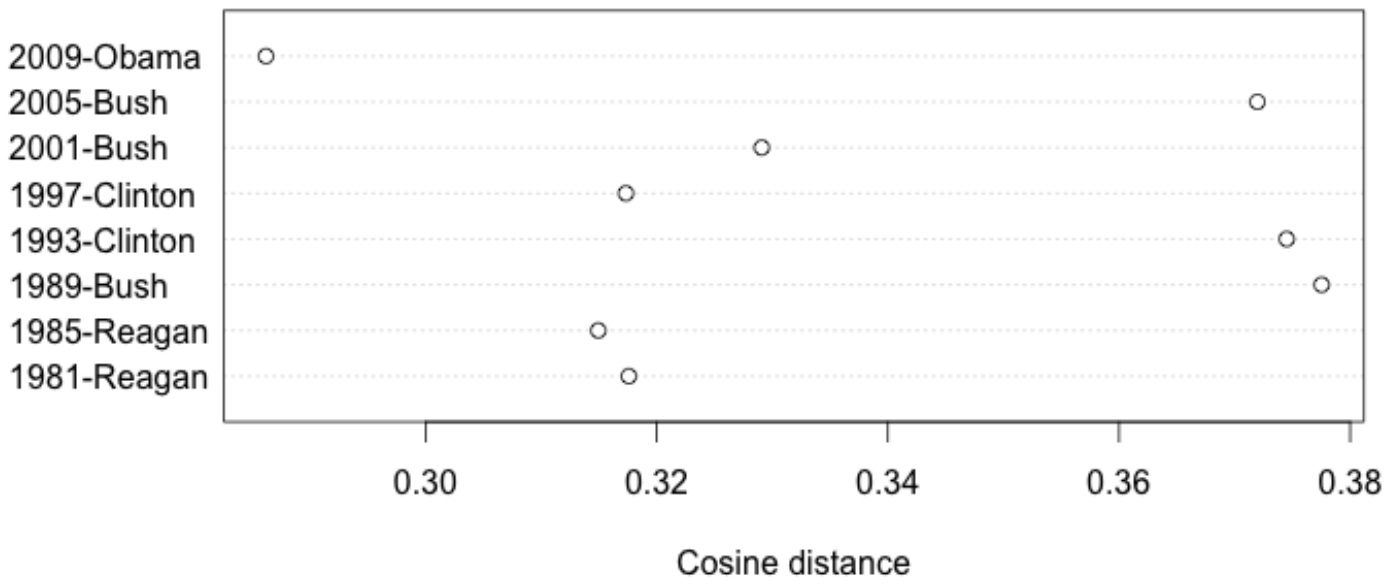
Jacquard coefficient

- Similar to the Cosine similarity
- Formula

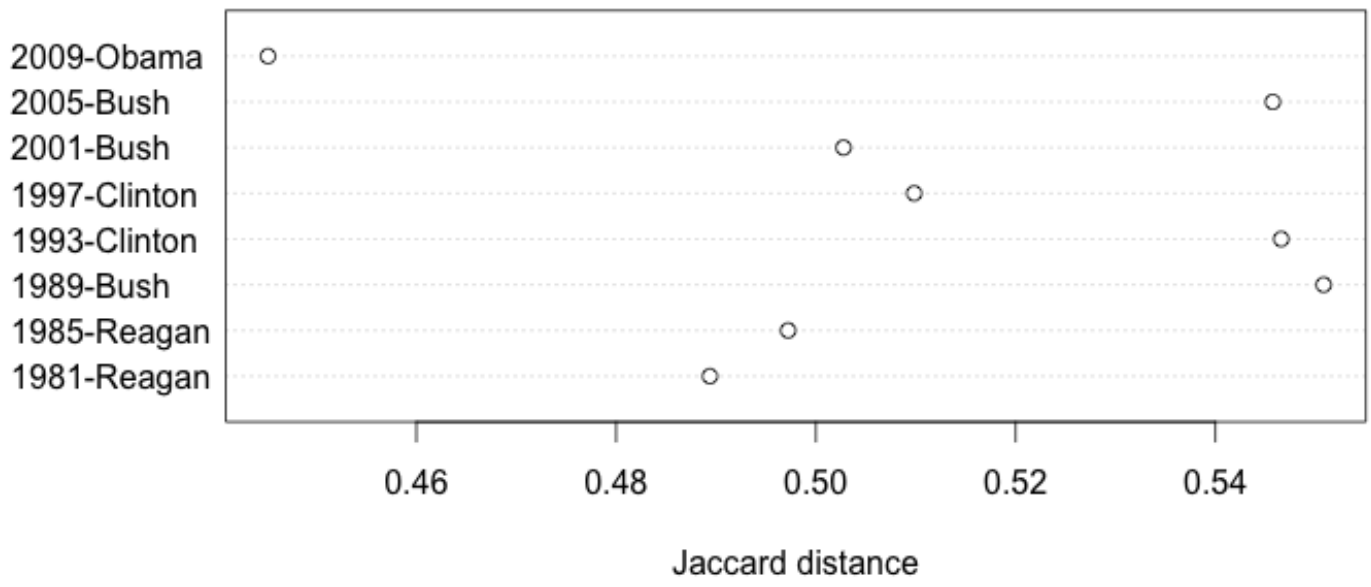
$$\frac{\mathbf{y}_A \cdot \mathbf{y}_B}{\|\mathbf{y}_A\| + \|\mathbf{y}_B\| - \mathbf{y}_A \cdot \mathbf{y}y_B} \tag{4}$$

- Ranges from 0 to 1.0

Example: Inaugural speeches



Example: Inaugural speeches



Can be made very general for binary features

Example: In the Choi et al paper, they compare vectors of features for (binary) absence or presence – called ("operational taxonomic units")

Table 1 OTUs Expression of Binary Instances i and j

j \ i	1 (Presence)	0 (Absence)	Sum
1 (Presence)	$a = i \bullet j$	$b = \bar{i} \bullet j$	$a+b$
0 (Absence)	$c = i \bullet \bar{j}$	$d = \bar{i} \bullet \bar{j}$	$c+d$
Sum	$a+c$	$b+d$	$n=a+b+c+d$

- Cosine similarity:

$$s_{\text{cosine}} = \frac{a}{\sqrt{(a+b)(a+c)}} \tag{5}$$

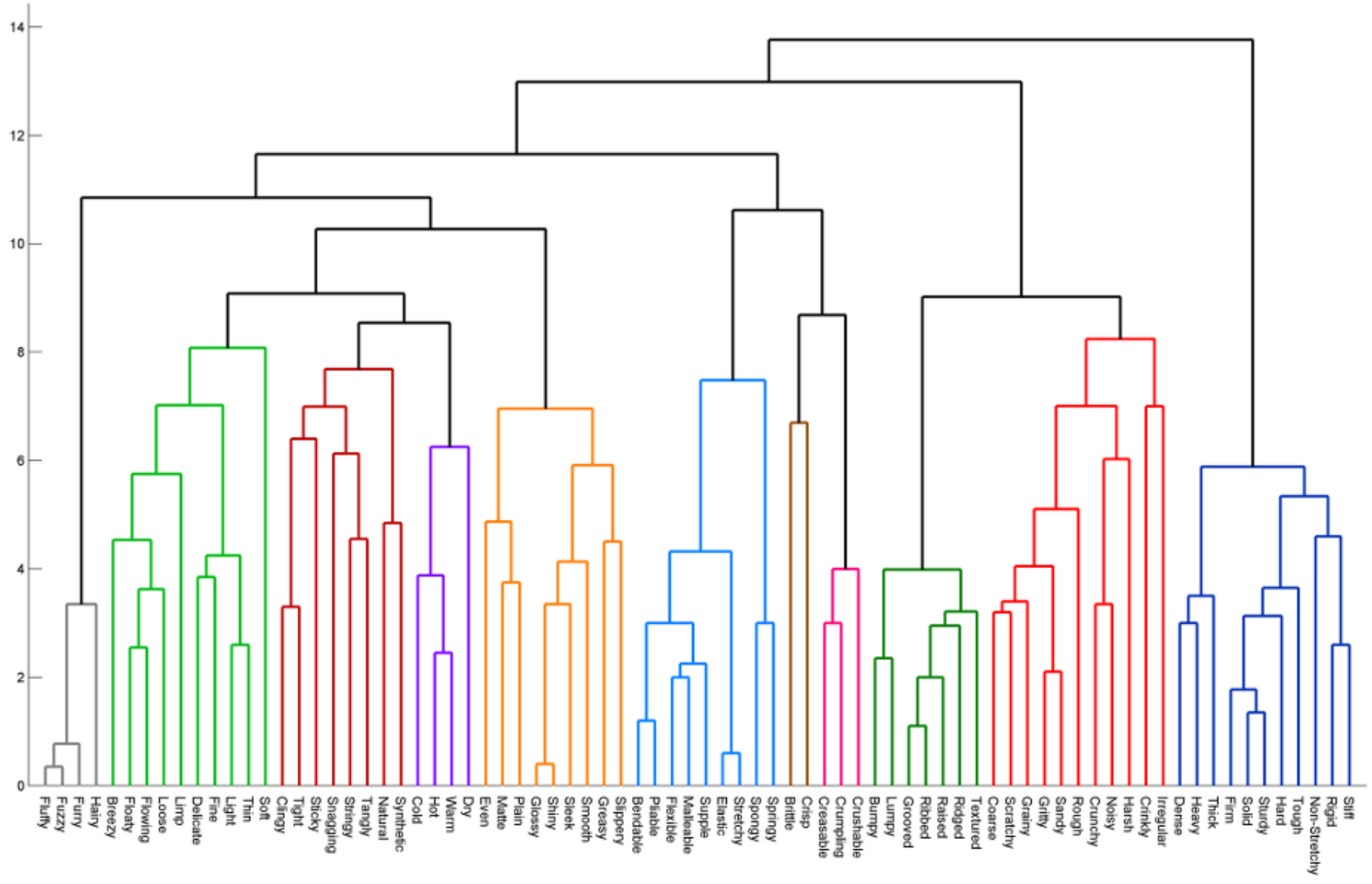
- Jaccard similarity:

$$s_{\text{Jaccard}} = \frac{a}{\sqrt{(a+b+c)}} \tag{6}$$

Typical features

- Normalized term frequency (almost certainly)
- Very common to use tf-idf – if not, similarity is boosted by common words (stop words)
- Not as common to use binary features

Uses for similarity measures: Clustering



Other uses, extensions

- Used extensively in information retrieval
- Summmary measures of how far apart two texts are – but be careful exactly how you define "features"
- Some but not many applications in social sciences to measure substantive similarity — scaling models are generally preferred
- Can be used to generalize or represent features in machine learning, by combining features using kernel methods to compute similarities between textual (sub)sequences without extracting the features explicitly (as we have done here)

Edit distances

- Edit distance refers to the number of operations required to transform one string into another for strings of equal length
- Common edit distance: the Levenshtein distance
- Example: the Levenshtein distance between "kitten" and "sitting" is 3
  - kitten $\rightarrow$ sitten (substitution of "s" for "k")
  - sitten $\rightarrow$ sittin (substitution of "i" for "e")
  - sittin $\rightarrow$ sitting (insertion of "g" at the end).
- Hamming distance: for two strings of equal length, the Hamming distance is the number of positions at which the corresponding characters are different
- Not common, as at a textual level this is hard to implement and possibly meaningless

Update on collocations

- **much** faster now

```
> time1 <- system.time(collocations(inaugTexts[50:57]))
> time2 <- system.time(collocations2(inaugTexts[50:57]))
> time1
   user  system elapsed
 62.928   1.796  65.176
> time2
   user  system elapsed
  0.128   0.002   0.129
> time1/time2
    user   system  elapsed
     Inf      Inf 505.2403
```

- trigrams

Sampling illustrated

- lexical diversity
- dictionaries (feature counts)
  Can construct multiple `dfm` objects, and count the "dictionary" features across texts. (Replicate the populism dictionary)