Lesson 3 - Association Measures

Erin M. Buchanan

01/24/2019

Language Topics Discussed

- Collocations: words that occur together more frequently than expected due to chance (peanut-butter)
- *n*-grams: n words that occur together, so a bigram is two words occurring together
- https://books.google.com/ngrams
- https://xkcd.com/ngram-charts/
- https://www.ted.com/talks/what_we_learned_from_5_million_books?language=en

Culturomics

- A term coined by the folks who used the Google Dataset to glean interesting information about humans based on the language that they used
- Looked at 4% of all printed books, digitized by Google
- Corpus of over 500 billion words across seven or more languages

Culturomics

- Estimated that English is around 1 million different words that are at least one per billion
- Showed that the dictionary only covers a small portion of these words
- Showed another proof for Zipf's law

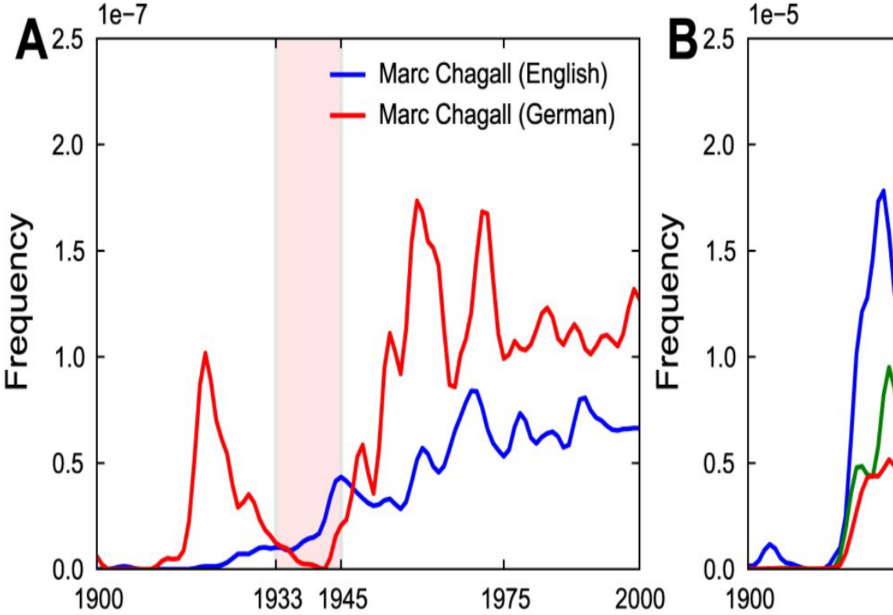
Culturomics

- Examined the competition of irregular and regular verbs (burnt, burned; found, finded; dwelt, dwelled)
- Looked at the frequency of naming for famous people - showed their rapid "fame rise", then the peak, followed by a half-life (decline in their listings)
  - The rise to fame was affected by job choice though - actors show the earliest peaks, followed by writers and politicians

Culturomics

- Censorship and Suppression: we can see when cultures (or those in charge) are suppressing certain instances of words across time
- You see across lots of countries:
  - Russia: Trotsky
  - Germany: Marc Chagall
  - China: Tiananmen Square
  - US: The Hollywood Ten

Culturomics



Some Considerations

- Optical Character Recognition (OCR) isn't perfect (s versus f)
- The meta-data is not perfect, so dates may be incorrect
- Synonymy: multiple meanings over the years can be difficult to interpret (tweet)

Association Measures

	Y variable	Not Y
X variable	A	B
Not X	C	D

- We can take any two variables we are interested in and calculate the relation between them using a basic contingency table.

Association Measures

- Unidirectional/asymmetric: Association measures that change based on if you switch rows/columns in our frequency table
  - Conditional probabilities: $P(X|Y)$ is not always equal to $P(Y|X)$
- Bidirectional/symmetric: Association measures that do not change based on the layout of the table

Conditional Probability

```
#collocate table for cellar (Y) and door (X)
#common to put collexeme on X, lexeme on Y
a = 146
b = 18828
c = 2282
d = 560000000-a-b-c
#P(Y|X) probability of cellar given door (door to cellar)
a/(b+a) * 100
```

```
## [1] 0.769474
```

```
#P(X|Y) probability of door given cellar (cellar to door)
a/(c+a) * 100
```

```
## [1] 6.01318
```

Conditional Probability

- Attraction: conditional probability of lexeme given construction
- Reliance/Faith: conditional probability of construction given lexeme
- As noted, these are not necessarily going to be the same

Another consideration

- Contingency based measures: measures of associative strength that account for the other possible co-occurrences
- For example, category learning shows a distinct hierarchy of features that are important for categories (i.e., wings to bird versus eyes to bird)

Example: We Can Do It!

```
he = c(33582, 866416, 2916576, (560000000 - 33582 - 866416
she = c(14180, 866416, 1533454, (560000000 - 14180 - 86641
he_she = as.data.frame(rbind(he,she))
colnames(he_she) = c("a", "b", "c", "d")
he_she
```

```
##            a      b       c         d
## he  33582 866416 2916576 556183426
## she 14180 866416 1533454 557585950
```

Attraction

- Attraction: probability of X given Y
- X here is can, Y is he or she

```
attraction = he_she$a/(he_she$a+he_she$c)*100
attraction
```

```
## [1] 1.1383119 0.9162373
```

```
rownames(he_she)
```

```
## [1] "he"  "she"
```

Reliance

- Reliance: probability of Y given X
- X here is can, Y is he or she

```
reliance = he_she$a/(he_she$a+he_she$b)*100
reliance
```

```
## [1] 3.731342 1.610273
```

```
rownames(he_she)
```

```
## [1] "he"  "she"
```

Delta-P

```
#treats it as Y to X (lexeme to collexeme) so he-can, she-c
dp_YX = he_she$a / (he_she$a + he_she$c) - he_she$b / (he_s
dp_YX
```

```
## [1] 0.009827754 0.007610914
```

```
#treats as X to Y so can-he, can-she similar to reliance
dp_XY = he_she$a / (he_she$a + he_she$b) - he_she$c / (he_s
dp_XY
```

```
## [1] 0.03209686 0.01336011
```

Probability based on Fisher's Test

- Fisher's Exact is a form of chi-square analysis that determines if there are associations in categorical variables
- You can take these *p*-values and log transform them
- Interpretation is:
  - Positive numbers = mutual attraction
  - Negative numbers = no attraction, "repelling"
  - Close to zero = no relation
- Good for low frequency variables

LogP.Fisher

```
library(Rling)
#expected frequency
aExp = (he_she$a + he_she$b)*(he_she$a + he_she$c)/(he_she$
#p values
pvF = pv.Fisher.collostr(he_she$a, he_she$b, he_she$c, he_s
#log based on expected frequency
logpvF = ifelse(he_she$a < aExp, log10(pvF), -log10(pvF))
logpvF
```

```
## [1] Inf Inf
```

Log Likelihood

- Ratio of probabilities of the likelihood of your lexeme-collexeme combination to not
- Positive indicates attraction type value
- Negative indicates repelling

```
LL = LL.collostr(he_she$a, he_she$b, he_she$c, he_she$d)
LL1 = ifelse(he_she$a < aExp, -LL, LL)
LL1
```

```
## [1] 75028.03 26737.77
```

Pointwise Mutual Information

- Ratio of the probability of Y given X divided by the probability of Y

```
PMI = log(he_she$a / aExp)^2
PMI
```

```
## [1] 3.832494 3.106205
```

Log Odds Ratio

- Ratio of the likelihood of Y given the presence of X to Y given not the presence of X

```
logOR = log(he_she$a*he_she$d/(he_she$b*he_she$c))
logOR
```

```
## [1] 2.000313 1.783561
```

Which one?

- If these all give me the same basic answer, which one should I use?
  - What is typical in your field?
  - Small sample sizes: Fisher's Test, Log Likelihood
  - Larger sample sizes: PMI, others
  - Compare across datasets: Odds Ratios

Summary

- Culturomics or ways that we can study culture through language
- Different types of ways to calculate association based on X and Y frequencies