Corpus Analysis with Open Source Tools

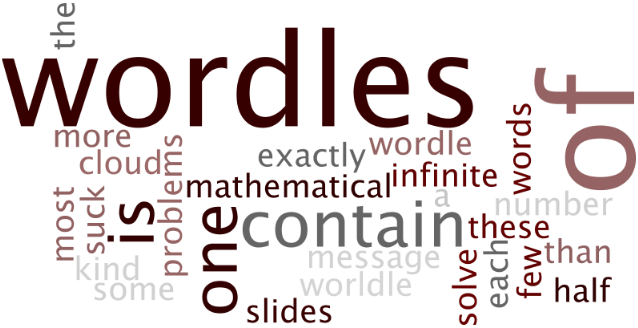
Camilo Thorne

Data and Web Science (DWS) Group
Universität Mannheim, Germany

NASSLLI 2016

Outline

Motivation - Corpora



Tutorial Materials

1. During the labs, we will be using
   - R: https://cran.rstudio.com/
   - Python: https://www.python.org/downloads/
   - a number of and tools and datasets
2. Please refer to the course's Git repository for detailed instructions
   - URI: https://github.com/camilothorne/nasslli2016

     (you can check it out without entering any credentials!)

Introduction

- Words and structures in English occur following some general laws or empirical hypothesis

- A distribution describes how often they occur/probable they are
- Many such distributions may hold:
  - power laws (Zipfian distributions)
  - normal distributions
  - binomial distributions
  - Poisson distributions
  - ...
- We can leverage on such distributions to infer or empirically validate such hypothesis

Introduction

- Words and structures in English occur following some general laws or empirical hypothesis
- A distribution describes how often they occur/probable they are
- Many such distributions may hold:
  - power laws (Zipfian distributions)
  - normal distributions
  - binomial distributions
  - Poisson distributions
  - ...
- We can leverage on such distributions to infer or empirically validate such hypothesis
- Methodology:
  1. Use corpora to reasonable approximate full English
  2. Use descriptive and inferential statistics to fit/estimate distributions and validate hypothesis

Population, Sample, Feature

- By population we mean the universe S of all possible observable events
  - e.g., the set W of all sentences ever written in English
- A sample is any representative subset $S' \subseteq S$ of the population
  - e.g., a given corpus $W \subset W'$, such as the Brown corpus
- A feature X is any property we may observe over S (or S'), viz., a random variable $X : S \to D$, where
  - if D is a number domain, X is a numeric feature
    (e.g., X_l: lengths $|w|$ of words w)
  - if $D = \{x_1, \ldots, x_n\}$, X is an (ordered) factor
    (e.g., X_s: syntactic categories *syn*(w) of words w) of words w)

Measures of Concentration and Dispersion

Consider a feature X, concentration measures how similar the values $x \in X$ are, and dispersion how much they differ

1. Measures of concentration:

Also known as parameters (many more!!!)

Measures of Concentration and Dispersion

Consider a feature X, concentration measures how similar the values $x \in X$ are, and dispersion how much they differ

1. Measures of concentration:
   - mean $\mu = E[X] = \begin{cases} \sum_{x \in X} x\, P(X = x), & \text{if } X \text{ is dicrete} \\ \int_{-\infty}^{\infty} x\, P(X = x)\, d(x), & \text{otherwise} \end{cases}$

Also known as parameters (many more!!!)

Measures of Concentration and Dispersion

Consider a feature X, concentration measures how similar the values $x \in X$ are, and dispersion how much they differ

1. Measures of concentration:
   - mean $\mu = E[X] = \begin{cases} \sum_{x \in X} x\, P(X = x), & \text{if } X \text{ is dicrete} \\ \int_{-\infty}^{\infty} x\, P(X = x)\, d(x), & \text{otherwise} \end{cases}$
   - mode *mod* $\in \{x \mid \text{ for all } x' \in X, x \geq x'\}$ (need not be unique)

Also known as parameters (many more!!!)

Measures of Concentration and Dispersion

Consider a feature X, concentration measures how similar the values $x \in X$ are, and dispersion how much they differ

1. Measures of concentration:

   - mean $\mu = E[X] = \begin{cases} \sum_{x \in X} x\, P(X = x), & \text{if } X \text{ is dicrete} \\ \int_{-\infty}^{\infty} x\, P(X = x)\, d(x), & \text{otherwise} \end{cases}$
   - mode *mod* $\in \{x \mid \text{ for all } x' \in X, x \geq x'\}$ (need not be unique)
   - median $m = \begin{cases} x_{\frac{|X|-1}{2}+1}, & \text{if } |X| \text{ is odd} \\ x_{\frac{|X|}{2}+1}, & \text{otherwise} \end{cases}$ (X sorted in increasing order)

Also known as parameters (many more!!!)

Measures of Concentration and Dispersion

Consider a feature X, concentration measures how similar the values $x \in X$ are, and dispersion how much they differ

1. Measures of concentration:
   - mean $\mu = E[X] = \begin{cases} \sum_{x \in X} x\, P(X = x), & \text{if } X \text{ is dicrete} \\ \int_{-\infty}^{\infty} x\, P(X = x)\, d(x), & \text{otherwise} \end{cases}$
   - mode *mod* $\in \{x \mid \text{ for all } x' \in X, x \geq x'\}$ (need not be unique)
   - median $m = \begin{cases} x_{\frac{|X|-1}{2}+1}, & \text{if } |X| \text{ is odd} \\ x_{\frac{|X|}{2}+1}, & \text{otherwise} \end{cases}$ (X sorted in increasing order)
2. Measures of dispersion:

Also known as parameters (many more!!!)

Measures of Concentration and Dispersion

Consider a feature X, concentration measures how similar the values $x \in X$ are, and dispersion how much they differ

1. Measures of concentration:
   - mean $\mu = E[X] = \begin{cases} \sum_{x \in X} x\, P(X = x), & \text{if } X \text{ is dicrete} \\ \int_{-\infty}^{\infty} x\, P(X = x)\, d(x), & \text{otherwise} \end{cases}$
   - mode *mod* $\in \{x \mid \text{ for all } x' \in X, x \geq x'\}$ (need not be unique)
   - median $m = \begin{cases} x_{\frac{|X|-1}{2}+1}, & \text{if } |X| \text{ is odd} \\ x_{\frac{|X|}{2}+1}, & \text{otherwise} \end{cases}$ (X sorted in increasing order)
2. Measures of dispersion:
   - variance $\sigma^2 = \mathit{Var}(X) = \sum_{x \in X} (\mu - x)^2$

Also known as parameters (many more!!!)

Measures of Concentration and Dispersion

Consider a feature X, concentration measures how similar the values $x \in X$ are, and dispersion how much they differ

1. Measures of concentration:
   - mean $\mu = E[X] = \begin{cases} \sum_{x \in X} x\, P(X = x), & \text{if } X \text{ is dicrete} \\ \int_{-\infty}^{\infty} x\, P(X = x)\, d(x), & \text{otherwise} \end{cases}$
   - mode $\mathit{mod} \in \{x \mid \text{ for all } x' \in X, x \geq x'\}$ (need not be unique)
   - median $m = \begin{cases} x_{\frac{|X|-1}{2}+1}, & \text{if } |X| \text{ is odd} \\ x_{\frac{|X|}{2}+1}, & \text{otherwise} \end{cases}$ (X sorted in increasing order)
2. Measures of dispersion:
   - variance $\sigma^2 = \mathit{Var}(X) = \sum_{x \in X} (\mu - x)^2$
   - standard deviation $\sigma = \sqrt{\mathit{Var}(X)}$

Also known as parameters (many more!!!)

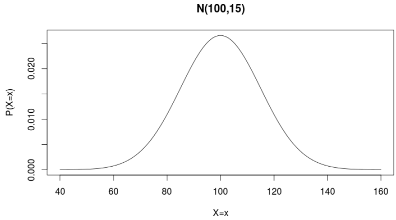
Parameters and Distributions

Definition (Distribution)

A distribution with parameters $\theta_1, \ldots, \theta_k$ is a function $F(\theta_1, \ldots, \theta_k)$ that describes the likelihood/probability of a feature (random variable) X taking value x, i.e., $P(X = x) = F(x; \theta_1, \ldots, \theta_k)$.

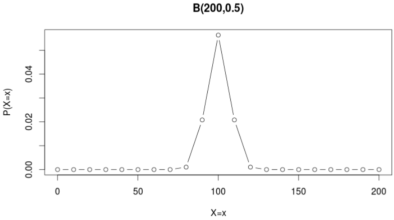
Normal distribution:

$$N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Binomial distribution:

$$B(x; n, p) = p^x C_n^x (1-p)^{n-x}$$



Measures of Correspondence

Question: Given two features X and Y, how to know if they relate two each other?

1. X and Y are numeric, in that case measure correlation

$$\mathit{Corr}(X, Y) = \frac{\mathit{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

2. X and Y are factors, in that case measure mutual information

$$MI(X;Y) = \sum_{x \in X} \sum_{y \in Y} P(X = x, Y = y) \log_2 \left(\frac{P(X = x, Y = y)}{P(X = x)P(Y = y)} \right)$$

English Corpora

- A corpus is a collection of written sentences
- It codifies information about a number of topics...

Format	Size (# words)	Source
Google NGram	$>$ 100 billion	link
Brown	$\sim$ 1 million	link
BNC	$\sim$ 100 million	link
WaCkY	$\sim$ 1 billion	link
$\vdots$	$\vdots$	$\vdots$

$\Rightarrow$ Rem: many small ones available with NLTK v2.0+

English Corpora

- A corpus is a collection of written sentences
- It codifies information about a number of topics...
- but also about language use!

Format	Size (# words)	Source
Google NGram	$>$ 100 billion	link
Brown	$\sim$ 1 million	link
BNC	$\sim$ 100 million	link
WaCkY	$\sim$ 1 billion	link
⋮	⋮	⋮

$\Rightarrow$ Rem: many small ones available with NLTK v2.0+

Annotation Standards [GB12]

- Sometimes, corpora are annotated to support NLP tasks
- This annotation is usually done manually, but sometimes also (semi)automatically
- The annotation labels typically respect a predefined format

Format	Task	Source
Brown/Penn tagging	POS tagging	POS tags
CoNLL NER	NER	entities
CoNNL chunk	chunk parsing	chunks
Penn treebank	constituency parsing	parse tags
Universal dependencies	dependency parsing	dependencies
⋮	⋮	⋮

Hypothesis Testing

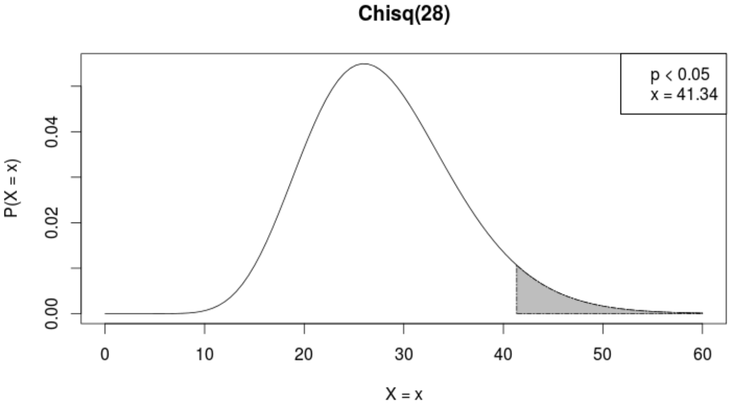
- A hypothesis h is a statement inferred from a sample S' that we would like generalize to the complete population S, for example:

  1. proper names of persons ("Jane Eyre") start with a capital letter and are not preceded by temporal prepositions ("during", "until")
  2. collocations/multiwords ("lone wolf") occur more frequently than the random combination of their constituent words
  3. ...

- Procedure:

  1. ensure S' is representative
  2. consider the alternative null hypothesis h_0
  3. try to *reject* h_0 via an statistical test
  4. if the test shows that h_0 doesn't fit $S' \implies h$ may hold over S

Test Statistics: The *Chisq*(k) Distribution



- Measure over H their goodness of fit X w.r.t. S
- $X \sim$ *Chisq*(k) with $k+1$ degrees of freedom $(k+1 = |S| \leq 100)$
- If $X(h)$ lies in region where $p = P(X > x) < 0.05 \implies$ accept
- $0 \leq p \leq 1$ is called the significance level

How to reason?

- Rejection of $h_0 \implies$ acceptance of alternative h

$$\frac{h_0 \vee h \qquad h_0 \text{ false}}{h \text{ true}}$$

- Minimize Type II error

	h_0 true	h_0 false
h_0 accept	true positive	(Type I error) false positive
h_0 reject	(Type II error) false negative	true negative

- Idea: we want as candidates as many true hypothesis as possible, even if the results are imprecise

- But: if H_0 clearly false, we can reasonably assume that H is a true positive

Linear Regression

Definition (Linear Regression)

A linear regression model has the form

$$E(Y \mid X_1, \ldots, X_n) = \theta_1 X_1 + \cdots + \theta_k X_k + \theta_{k+1}$$

with θ_{k+1} the intercept and $\theta_1, \ldots, \theta_k$ the slopes.

Linear Regression

Definition (Linear Regression)

A linear regression model has the form

$$E(Y \mid X_1, \dots, X_n) = \theta_1 X_1 + \cdots + \theta_k X_k + \theta_{k+1}$$

with θ_{k+1} the intercept and $\theta_1, \dots, \theta_k$ the slopes.

- Assumptions:
  1. The X_is are conditionally independent on Y
  2. The sample if it is normally distributed (i.e., whenever $(Y \mid X_1, \dots, X_n) \sim N(\mu, \sigma^2)$)

Linear Regression

Definition (Linear Regression)

A linear regression model has the form

$$E(Y \mid X_1, \ldots, X_n) = \theta_1 X_1 + \cdots + \theta_k X_k + \theta_{k+1}$$

with θ_{k+1} the intercept and $\theta_1, \ldots, \theta_k$ the slopes.

- Assumptions:
  1. The X_is are conditionally independent on Y
  2. The sample if it is normally distributed (i.e., whenever $(Y \mid X_1, \ldots, X_n) \sim N(\mu, \sigma^2)$)
- Linear models describe the average $E(Y \mid X_1, \ldots, X_n)$ a dependent feature Y and features $X_1, \ldots, X_n$

Linear Regression

Definition (Linear Regression)

A linear regression model has the form

$$E(Y \mid X_1, \ldots, X_n) = \theta_1 X_1 + \cdots + \theta_k X_k + \theta_{k+1}$$

with θ_{k+1} the intercept and $\theta_1, \ldots, \theta_k$ the slopes.

- Assumptions:
  1. The X_is are conditionally independent on Y
  2. The sample if it is normally distributed (i.e., whenever $(Y \mid X_1, \ldots, X_n) \sim N(\mu, \sigma^2)$)
- Linear models describe the average $E(Y \mid X_1, \ldots, X_n)$ a dependent feature Y and features $X_1, \ldots, X_n$
- The parameters/coefficients $\theta_1, \ldots, \theta_k, \theta_{k+1}$ can be estimated via different (equivalent methods): square error minimization, maximum likelihood, etc.

Linear Regression

Definition (Linear Regression)

A linear regression model has the form

$$E(Y \mid X_1, \ldots, X_n) = \theta_1 X_1 + \cdots + \theta_k X_k + \theta_{k+1}$$

with θ_{k+1} the intercept and $\theta_1, \ldots, \theta_k$ the slopes.

- Assumptions:
  1. The X_is are conditionally independent on Y
  2. The sample if it is normally distributed (i.e., whenever $(Y \mid X_1, \ldots, X_n) \sim N(\mu, \sigma^2)$)
- Linear models describe the average $E(Y \mid X_1, \ldots, X_n)$ a dependent feature Y and features $X_1, \ldots, X_n$
- The parameters/coefficients $\theta_1, \ldots, \theta_k, \theta_{k+1}$ can be estimated via different (equivalent methods): square error minimization, maximum likelihood, etc.
- We can quantify how well the model fits the data via a number of indexes
  - external: χ^2-goodness of fit, etc.
  - internal: R^2-goodness of fit, BIC, AIC, etc.

Analysis of Variance (ANOVA)

- Once we have fitted a regression model to a dataset we can use it to
  1. explore the impact of each single factor
  2. the impact of a factor is reflected by the variation it induces

- A technique to understand variation w.r.t. factors is analysis of variance (ANOVA)

- Procedure:
  1. consider factor X of k levels/groups
  2. for each level i, consider a linear model where levels/groups $j \neq i$ are fixed
  3. test (reject) if $\mu = \mu_i$, where
     - μ original conditional expectation
     - μ_i conditional expectation of linear model w.r.t. level/group i

Generalized Linear Models

- What do we do when a dependency is non-linear?
- What do we do when $(Y|X_1, \ldots, X_k)$ is not normally distributed?

Generalized Linear Models

- What do we do when a dependency is non-linear?
- What do we do when $(Y|X_1, \ldots, X_k)$ is not normally distributed?

- Answer: generalize linear models to arbitrary distributions, modulo some kind of transformation

Definition (GLM)

A generalized linear model has the form

$$f(E(Y|X_1, \ldots, X_k)) = \theta_1 X_1 + \cdots + \theta_k X_k + \theta_{k+1}$$

where

1. $f\colon \mathbb{R} \to \mathbb{R}$ is a link function
2. $(Y|X_1, \ldots, X_k) \sim D$, with D an arbitrary distribution

Mixed Effects Models

- Sometimes, when doing regression, some factors, while important, behave like error terms
  - example: when you repeatedly observe the same feature over the same individual over time
  - each time you make a measurement, some random noise gets mixed in the measurement!

Mixed Effects Models

- Sometimes, when doing regression, some factors, while important, behave like error terms
  - example: when you repeatedly observe the same feature over the same individual over time
  - each time you make a measurement, some random noise gets mixed in the measurement!
- One can refine linear models by letting such features to vary randomly

Mixed Effects Models

- Sometimes, when doing regression, some factors, while important, behave like error terms
  - example: when you repeatedly observe the same feature over the same individual over time
  - each time you make a measurement, some random noise gets mixed in the measurement!
- One can refine linear models by letting such features to vary randomly
- This results in a linear mixed model of the form

$$\begin{aligned} f(E(Y|X_1, \ldots, X_k)) &= \theta_1 X_1 + \cdots + \theta_n X_n + \\ &\quad \theta_{n+1} Z_1 + \cdots + \theta_{n+m} Z_m + \theta_{(n+m)+1} \end{aligned}$$

where the X_is are fixed and the Z_js are non-fixed

Power Laws [Bar09]

- The frequency of a word w in a sample $S' \subseteq S$ over vocabulary W is the number of times (count) we observe it in S', viz., *Freq*: $W \to \mathbb{N}$ s.t. *Freq*$(w) = P(W = w) \times |S'|$

- Words can be ordered by frequency rank *Rank*: $W \to |W|$ s.t. *Rank*$(w) <$ *R*(w) if
  1. *Freq*$(w) <$ *Freq*(w'), or
  2. *Freq*$(w) =$ *Freq*(w') and w comes before w' in lexicographic order

- We can use regression on the $\log$-$\log$ scale to model power laws among word frequency and rank

$$\log \mathit{Freq}(w) = \log(\theta) - \theta' \log(\mathit{Rank}(w))$$
$$\Longrightarrow \qquad \mathit{Freq}(w) = \frac{\theta}{\mathit{Rank}(w)^{\theta'}}$$
$$\Longrightarrow \qquad \mathit{Rank} \sim PL(\theta, \theta'), \text{ for some } \theta, \theta' \geq 0$$

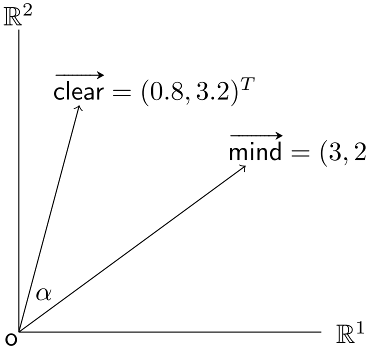
Semantic Relatedness [Nav09]

Distributional Hypothesis

1. The meaning of a word w is given by sentential context
2. Sentential context is the window of k words surrounding w
3. Two words w and w' have the same meaning if they occur always in the same context

- The semantic relatedness of two words can be estimated by computing context similarity over very large word corpora of vocabulary W
- Idea: compute a word space over the corpus
  1. a word space is a matrix $M_{|W|\times k}$ where each $m_{i,j}$ is the "F.IJF" of words $w_i, w_j \in W$: $m_{i,j} = \textit{freq}(w_i) \times \textit{ijfreq}(w_i, w_j)$
  2. $M_{|W|\times k}$ defines a k-dimensional real-valued vector space $\subseteq \mathbb{R}^k$
  3. each word w_i is mapped to a vector $\vec{w_i} = (m_{i,1}, \ldots, m_{i,k})^T$
  4. then $\textit{rel}(w, w') = \textit{sim}(\vec{w}, \vec{w}')$

Vector space $\mathbb{R}^k$ with dimensions $|k|$ (size of context window) (size of vocabulary)



Model semantic relatedness in terms of cosine similarity

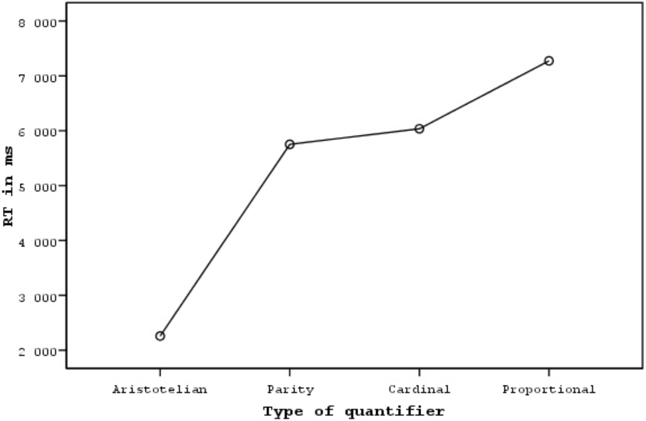
$$\textit{rel}(\text{clear}, \text{mind}) = \textit{cos}(\alpha) = \frac{\overrightarrow{\text{clear}} \bullet \overrightarrow{\text{mind}}}{||\overrightarrow{\text{clear}}|| \times ||\overrightarrow{\text{mind}}||}$$

The WaCkY Corpus [BBFZ09]

```
<s>
Flender Flender NP      1       3       VMOD
Werke   Werke   NP      2       3       SBJ
was     be      VBD     3       0       ROOT
a       a       DT      4       7       NMOD
German  German  JJ      5       7       NMOD
shipbuilding    shipbuilding    NN      6       7       NMOD
company company NN      7       3       PRD
,       ,       ,       8       7       P
located locate  VVN     9       7       NMOD
in      in      IN      10      9       ADV
Lubeck  Lubeck  NP      11      10      PMOD
.       .       SENT    12      0       ROOT
</s>
```

	Sentences	Tokens	Source
WaCkY (Eng)	$\sim$ 43 million	$\sim$ 800 million	Wikipedia (EN, 2008)

Answer Time and Complexity [Szy09]



Parity: "exactly 2" *Cardinal*: all the other counting quantifiers

Corpus Analysis [TS15]

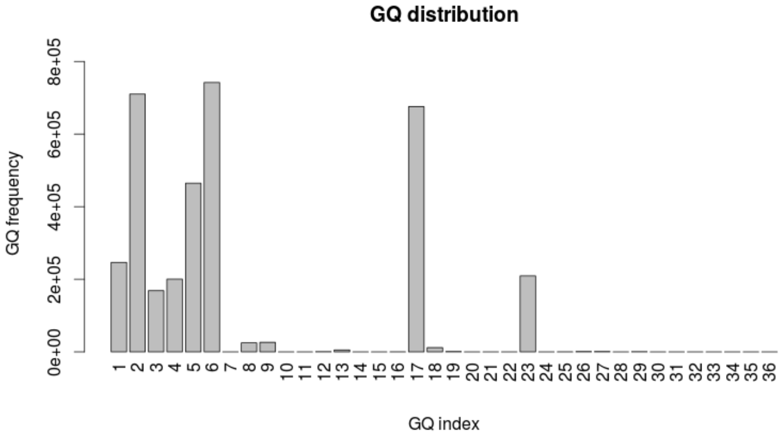
- We can build a list of simple patterns to identify and count quantifiers in corpora

  1. Aristotelian quantifiers: *all*, *some*
  2. counting quantifiers: k, *less (more) than* k
  3. proportional quantifiers: *most*, *few*, p/k, $k\%$, *less (more) than* p/k, *less (more) than* $k\%$

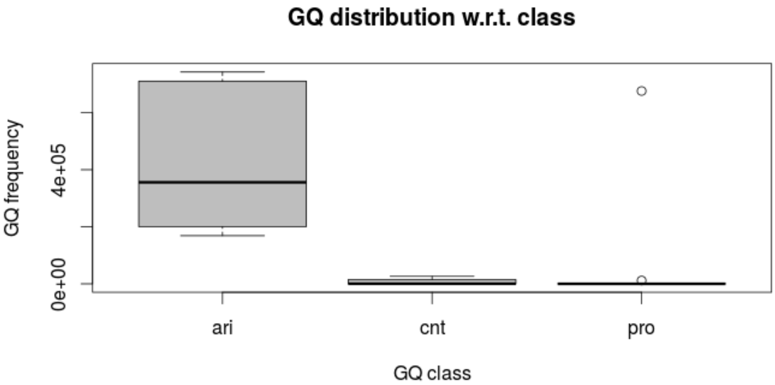
- Examples: $\begin{cases} \textit{most} = \texttt{most/dt}, \ \texttt{most/jjs [a-z]\{1,12\}/nns} \\ \textit{some} = \texttt{some/det} \end{cases}$

- Understand how much their frequency is influenced by

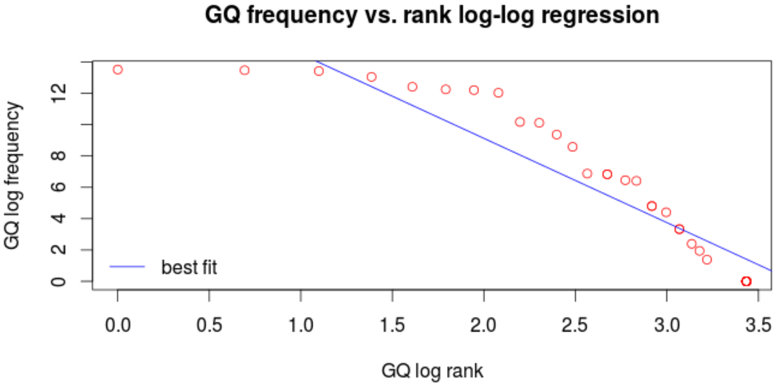
  1. length (characters, word units) ⇒ "syntactic complexity"
  2. quantifier class ⇒ "semantic complexity"
  3. other factors/features



- distribution not normal
- heavy (right) tailed
- most GQs in the sample are very rare (frequency ≤ 1)



- most quantifiers are Aristotelian
- Aristotelian quantifiers show a lot of variability
- proportional quantifiers contain some outliers ("most", "few")



- log-log (ln-ln) regression (to test for power law):

$$R^2 = 0.7581$$
$$p = 0.0001074$$

Larger is Better!

Theorem (Law of Big Numbers)

Let $S_1, S_2, \ldots, S_i$ be a family of increasingly large samples of some population S (i.e., $S_i \subseteq S_{i+1} \subseteq S$) distributed under $F(\theta_1, \ldots, \theta_k)$ (i.e., $S \sim F(\theta_1, \ldots, \theta_k)$). Then, for $i \geq 1$, the estimators $\hat{\theta}_{i,1}, \ldots, \hat{\theta}_{i,k}$ converge to the true parameters $\theta_1, \ldots, \theta_k$) of F (i.e., for $j \in [1, k]$, $\lim\limits_{i \mapsto \infty} \theta_{i,j} = \theta_j$).

- The larger the sample, the better we can observe or fit to it a distribution!
- Grain of salt:
  1. samples are assumed to be representative w.r.t. population and i.i.d.
  2. samples are assumed to contain minimal noise
  3. the distribution should fit the data
- Meaning in practice:
  1. avoid making inferences over very small corpora
  2. trade-off corpora size by corpora quality

Thank You!

References I

Marco Baroni.
Distributions in text.
In Mouton de Gruyter, editor, *Corpus linguistics: An International Handbook*, volume 2, pages 803–821. 2009.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta.
The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora.
Language Resources and Evaluation, 43(3):209–226, 2009.

Stefan Th. Gries and Andrea L. B.
Linguistic annotation in/for corpus linguistics.
Preprint: `http://www.linguistics.ucsb.edu/faculty/stgries/research/InProgr_STG_ALB_LingAnnotCorpLing_HbOfLingAnnot.pdf` (to appear in the *Handbook of Linguistic Annotation*)., 2012.

Roberto Navigli.
Word sense disambiguation: A survey.
ACM Computing Surveys, 41(2):10:1–10:69, 2009.

References II

Jakub Szymanik.
Quantifiers in Time and Space.
Institute for Logic, Language and Computation, 2009.

Camilo Thorne and Jakub Szymanik.
Semantic complexity of quantifiers and their distribution in corpora.
In *Proceedings of the 11th International Conference in Computational Semantics (IWCS 2015)*, 2015.

Appendix: Regression Inference and Validation

- The least squares method computes the linear model whose parameters $(\theta_1, \ldots, \theta_{k+1})^*$ minimize square error $J(\cdot)$

$$\begin{aligned}(\theta_1, \ldots, \theta_{k+1})^* &= \arg \min_{(\theta_1, \ldots, \theta_{k+1})} J(\theta_1, \ldots, \theta_{k+1}) \\ &= \arg \min_{(\theta_1, \ldots, \theta_{k+1})} \sum_i \sum_j (y_i - \theta_j(x_{i,j}))^2\end{aligned}$$

- The R^2 coefficient provides a measure of confidence in the inferred model and is defined as the ratio of square error variance and dependent variable variance, i.e.,

$$R^2 = 1 - \left[\frac{\sum_i \sum_j (\theta_j^*(x_{i,j}) - E[Y])^2}{\mathit{Var}(Y)} \right]$$

Appendix: Regression BIC and AIC

- The Bayesian information criterion (BIC) is defined by

$$\mathit{BIC}(k+1, S') = -([2 \times \ln L^*] + [(k+1) \times \ln(|S'|)])$$

where

1. $L^* = P(S \mid (\theta_1, \ldots, \theta_{k+1})^*)$ (maximum likelihood)
2. $(\theta_1, \ldots, \theta_{k+1})^*$ are the optimal model parameters
3. $k+1$ is the number of parameters/coefficients of the model
4. $S' \subseteq S$ is the sample to which we fit the linear model

- The Akaike information criterion (AIC) is defined by

$$\mathit{AIC}(k+1, S') = -([2 \times (k+1)] - [2 \times \ln L^*])$$

with parameters as for BIC

- Note: these measures are useful when evaluating regression models, where the output/dependent variable is numeric, they are not appropriate for classification or clustering

Appendix: Confidence Intervals

- The theoretical mean μ of feature X is the expected value $E[X]$ of X
- We can estimate the theoretical mean from a sample $S' \subseteq S$ via maximum likelihood, giving rise to the known empirical mean

$$\hat{\mu} = \frac{1}{|X|} \sum_{x \in X} x$$

- Confidence intervals allow to infer, from $\hat{\mu}$, $\hat{\sigma}$, X and S' the potential range of values of μ
- Let $p < 0.05$ be the significance level of test statistic T, then

$$\mu \in \left[\hat{\mu} - \frac{t^* \, \hat{\sigma}}{\sqrt{|X|}}, \;\; \hat{\mu} + \frac{t^* \, \hat{\sigma}}{\sqrt{|X|}} \right]$$

where t^* is the upper $(1 - p)/2$ critical value of test T distribution with $|X| - 1$ degrees of freedom