



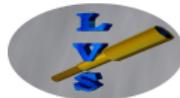
Introduction
Data Preparation
LVS
Working with Data
Visual Analytics
Inferential Analysis
References

Interactive Visual Data Analysis

Part One

Language Variation Suite

Olga Scrivner



Indiana University



Workshop in Methods

Introduction

Data Preparation

LVS

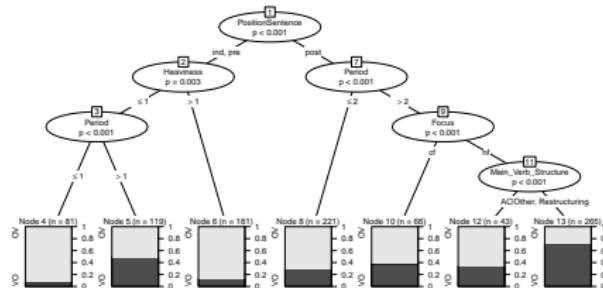
Working with Data

Visual Analytics

Inferential Analysis

References

- ① Introduce a web application for quantitative analysis: LVS
- ② Develop practical skills
- ③ Understand and interpret advanced statistical models





What is LVS?

Introduction

Data Preparation

LVS

Working with Data

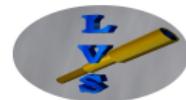
Visual Analytics

Inferential Analysis

References

Language Variation Suite

It is a Shiny web application originally designed for data analysis in sociolinguistic research.



It can be used for:

- Processing spreadsheet data
- Reporting in tables and graphs
- Analyzing means, regression, conditional trees ...
(and much more)



Background

Introduction

Data
Preparation

LVS

Working with
Data

Visual
Analytics

Inferential
Analysis

References

LVS is built in R using Shiny package:

- ① **R** - a free programming language for statistical computing and graphics
- ② **Shiny App** - a web application framework for R



Computational power of R + Web interactivity

Background

Introduction

Data Preparation

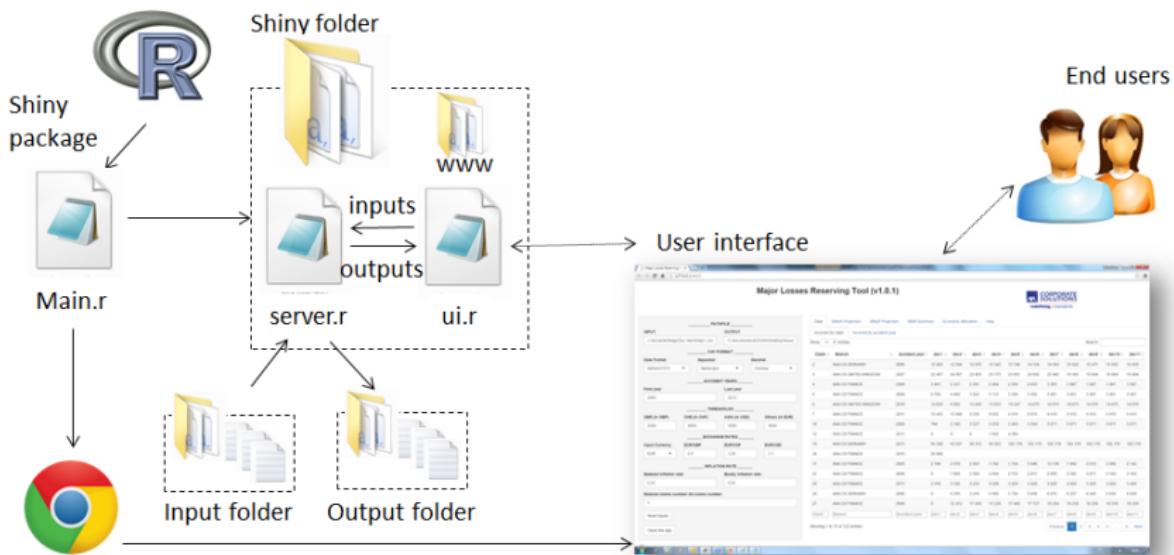
LVS

Working with Data

Visual Analytics

Inferential Analysis

References



<http://littleactuary.github.io/blog/Web-application-framework-with-Shiny/>



Workspace

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

Browser

- Chrome, Firefox, Safari - recommendable
- Explorer may cause instability issues



Accessibility

- PC, Mac, Linux
 - Data files will be uploaded from any location on your computer
- Smart Phone
 - Data files must be on a cloud platform connected to your phone account (e.g. dropbox)



Server

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

Since LVS is hosted on a server, Shiny idle time-out settings may stop application when it is left inactive (it will grey out).

Solution: Click **reload** and re-upload your csv file



Data Preparation

Introduction

Data
Preparation

LVS

Working with
Data

Visual
Analytics

Inferential
Analysis

References

Important things to consider before data entry:

- File format:
 - Comma separated value (CSV) - faster processing
 - Excel format will slow processing
- Column names should not contain spaces
 - Permitted: non-accented characters, numbers, underscore, hyphen, and period
- One column must contain your **dependent** variable
- The rest of the columns contain **independent** variables

A	B	C	D	E	F
Case	Number	R.Use	Lexical.Item	Style	Store
1	1	retention	Fourth	normal	Saks
1	2	retention	Fourth	normal	Saks
1	3	retention	Fourth	normal	Saks
1	4	retention	Fourth	normal	Saks
1	5	retention	Fourth	normal	Saks
1	6	retention	Fourth	normal	Saks
1	7	retention	Fourth	normal	Saks
1	8	retention	Fourth	normal	Saks



Terminology Review

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

- a. **Categorical** - non-numerical data with **two** values
 - yes - no; male - female
- b. **Continuous** - numerical data
 - duration, age, year
- c. **Multinomial** - non-numerical data with **three or more** values
 - regions, nationalities
- d. **Ordinal** - scale: currently not supported



Terminology Review

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

a. **Categorical** - non-numerical data with **two** values

- yes - no; male - female

b. **Continuous** - numerical data

- duration, age, year

c. **Multinomial** - non-numerical data with **three or more** values

- regions, nationalities

d. **Ordinal** - scale: currently not supported



Workshop Files

Introduction

Data
Preparation

LVS

Working with
Data

Visual
Analytics

Inferential
Analysis

References

<http://ssrc.indiana.edu/seminars/wim.shtml>

① **movie_metadata.csv**

Simplified set from <https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset>

② **LVS web site:** <https://languagevariationsuite.com>



Movie Data

Introduction

Data
Preparation

LVS

Working with
Data

Visual
Analytics

Inferential
Analysis

References

- Budget
- Director
- Actor 1
- Director facebook **likes**
- Actor 1 facebook **likes**
- Genre
- Year

director_name	director_facebook_likes	actor_1_facebook_likes	genres
Joel Schumacher	541	920	Action
Tim Burton	13000	920	Action
Michael Winnick	155	981	Action
Alec Asten	5	472	Action
Jon Hess	29	683	Action
John Stockwell	134	260000	Action
Quentin Tarantino	16000	926	Action



Language Variation Suite - Structure

Language Variation Suite (LVS)

About Data Visualization Inferential Statistics

Introduction
Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

① Data

- Upload file, data summary, adjust data, cross tabulation

② Visual Analysis

- Plotting, cluster classification

③ Inferential Statistics

- Modeling, regression, conditional trees, random forest



Language Variation Suite - Structure

Language Variation Suite (LVS)

About Data Visualization Inferential Statistics

Introduction
Data Preparation
LVS

Working with Data

Visual Analytics

Inferential Analysis

References

① Data

- Upload file, data summary, adjust data, cross tabulation

② Visual Analysis

- Plotting, cluster classification

③ Inferential Statistics

- Modeling, regression, conditional trees, random forest



Upload File

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

Language Variation Suite (LVS)

About Demo Data Visual Analysis RBRUL Inferential Statistics

Upload **movie_metadata.csv**

File Upload

Uploaded Dataset

Summary

Data Structure

Cross Tabulation

Frequency

Adjust Data

Step1: Upload CSV File

Choose CSV File

Browse...

movie_metadata.csv

Upload complete



Uploaded Dataset

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

The data content is imported as a table and allows for sorting columns.

The screenshot shows a user interface for data analysis. On the left, a sidebar menu lists several options: 'File Upload' (highlighted with a red box), 'Summary', 'Data Structure', 'Cross Tabulation', 'Frequency', and 'Adjust Data'. To the right of the sidebar is a table with data. At the top of the table, there is a 'Show 25 entries' dropdown. The table has three columns: 'director_name', 'director_facebook_likes', and 'actor_1_facebook_likes'. The data rows are as follows:

director_name	director_facebook_likes	actor_1_facebook_likes
Joel Schumacher	541	920
Tim Burton	13000	920
Michael Winnick	155	981
Alec Asten	5	472
Jon Hess	29	683
John Stockwell	134	260000



Summary

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

Summary provides a quantitative summary for each variable,
e.g. frequency count, mean, median.

File Upload

Uploaded Dataset

Summary

Data Structure

Cross Tabulation

Frequency

Adjust Data

	director_name	director_facebook_likes	actor_1_facebook_likes
Steven Spielberg	: 22	Min. : 0.0	Min. : 0
Clint Eastwood	: 18	1st Qu.: 10.0	1st Qu.: 767
Woody Allen	: 18	Median : 58.0	Median : 1000
Martin Scorsese	: 16	Mean : 890.9	Mean : 8016
Spike Lee	: 16	3rd Qu.: 228.0	3rd Qu.: 13000
Tim Burton	: 14	Max. : 23000.0	Max. : 640000
Barry Levinson	: 13		



Data Structure

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

File Upload

Uploaded Dataset

Summary

Data Structure

Cross Tabulation

Frequency

Adjust Data

```
'data.frame': 3086 obs. of 10 variables:  
 $ director_name      : Factor w/ 1501 levels "Aaron Schneider",...  
 $ director_facebook_likes : int 541 13000 155 5 29 134 16000 561 13000  
 $ actor_1_facebook_likes : int 920 920 981 472 683 260000 926 746 920  
 $ genres              : Factor w/ 5 levels "Action","Animation",...  
 $ actor_1_name        : Factor w/ 1250 levels "50 Cent","Aaliyah",...  
 $ movie_title         : Factor w/ 3039 levels "102 Dalmatians",...  
 $ cast_total_facebook_likes: int 2699 2899 2741 1752 1139 261818 3983 23  
 $ budget               : int 125000000 80000000 8000000 500000 30000  
 $ title_year           : int 1997 1992 2016 2015 1993 2016 2003 2012  
 $ movie_facebook_likes : int 0 0 689 62 107 0 13000 29000 12000 1500
```

① Total number of **observations** (rows)

② Number of **variables** (columns)

③ Variable **types**

- **Factor** - categorical values
- **Num** - numeric values (0.95, 1.05)
- **Int** - integer values (1, 2, 3)



Cross Tabulation

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

Cross-tabulation examines the relationship between variables.

The screenshot shows a software interface for statistical analysis. On the left, a sidebar lists several options: File Upload, Uploaded Dataset, Summary, Data Structure, Cross Tabulation (which is highlighted with a red box), Frequency, and Adjust Data. The main area has three tabs at the top: Instructions, Two-by-Two Cross Tabulation (also highlighted with a red box), and Multiple-Cross Tabulation. The 'Instructions' tab contains two dropdown menus: 'Select Dependent Variable (Rows)' and 'Which column contains your dependent variable?'. The first dropdown is set to 'NULL'. The second dropdown is also set to 'NULL' and has a small arrow indicating it can be changed. To the right of these dropdowns, there are two more sections: 'Select One Independent Variable (Columns)' and 'Variable for Column', both currently set to 'NULL'.

Cross Tabulation Plot

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

Select Dependent Variable (Rows)

Which column contains your dependent variable?

NULL

- actor_1_facebook_likes
- genres
- actor_1_name
- movie_title
- cast_total_facebook_likes
- budget** ←
- title_year
- movie_facebook_likes

Select One Independent Variable (Columns)

Variable for Column

NULL

- NULL
- director_name
- director_facebook_likes
- actor_1_facebook_likes
- genres** ←
- actor_1_name
- movie_title
- cast_total_facebook_likes

Cross Tabulation Plot

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

Select Dependent Variable (Rows)

Which column contains your dependent variable?

NULL

actor_1_facebook_likes
genres
actor_1_name
movie_title
cast_total_facebook_likes
budget
title_year
movie_facebook_likes



Select One Independent Variable (Columns)

Variable for Column

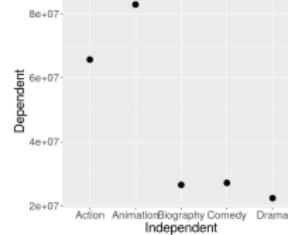
NULL

NULL
director_name
director_facebook_likes
actor_1_facebook_likes
genres
actor_1_name
movie_title
cast_total_facebook_likes



Action	65673657.30
Animation	82867225.43
Biography	26536984.86
Comedy	27161439.39
Drama	22394283.83

Continuous Dependent Variable (n)



Introduction

Data
Preparation

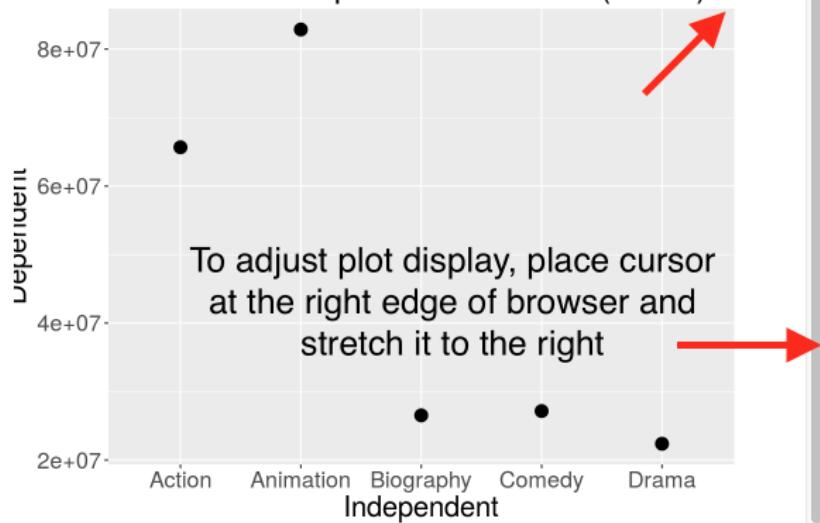
LVS

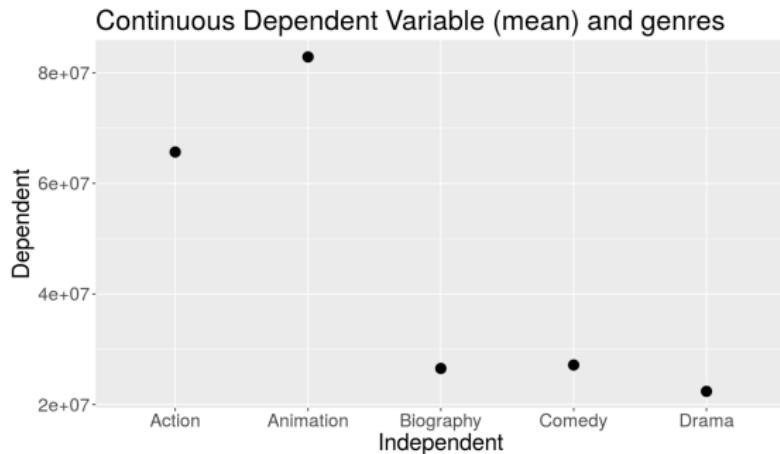
Working with
DataVisual
AnalyticsInferential
Analysis

References

Shiny pages are fluid and reactive.

Continuous Dependent Variable (mean) at



[Introduction](#)[Data Preparation](#)[LVS](#)[Working with Data](#)[Visual Analytics](#)[Inferential Analysis](#)[References](#)



Language Variation Suite - Structure

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

Language Variation Suite (LVS)

About Demo Data Visual Analysis RBRUL Inferential Statistics

① Data

- Upload file, data summary, adjust data, cross tabulation

② Visual Analysis

- Plotting, cluster classification

③ Inferential statistics

- Modeling, regression, varbrul analysis, conditional trees, random forest



Visual Analytics

Introduction

Data
Preparation

LVS

Working with
Data

Visual
Analytics

Inferential
Analysis

References

Visual Analytics: “The science of analytical reasoning
facilitated by visual interactive interfaces”
(Thomas et al. 2005)





One Variable Plot

Introduction

Data
Preparation

LVS

Working with
Data

Visual
Analytics

Inferential
Analysis

References

Language Variation Suite (LVS)

About Demo Data **Visual Analysis** RBRUL Inferential Statistics

One Variable Plot

Two Variables Plot

Three Variables
Plot

Cluster Plot

Frequency Plot



One Variable Plot

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

Language Variation Suite (LVS)

About Demo Data Visual Analysis RBRUL Inferential Statistics

One Variable Plot

Two Variables Plot

Three Variables Plot

Cluster Plot

Frequency Plot

Select one variable

NULL

NULL

director_name

director_facebook_likes

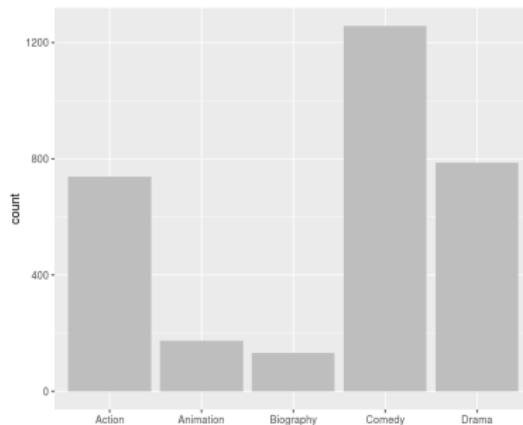
actor_1_facebook_likes

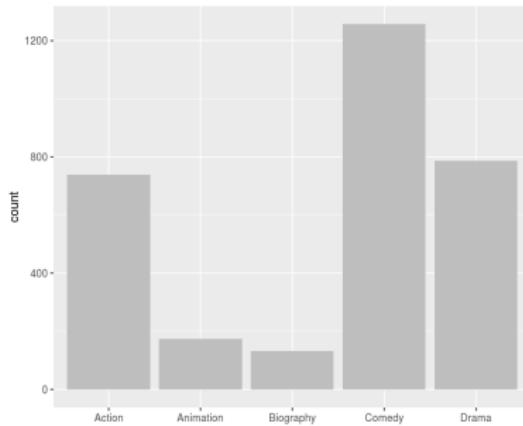
genres

actor_1_name

movie_title

cost_total_facebook_likes

[Introduction](#)[Data Preparation](#)[LVS](#)[Working with Data](#)[Visual Analytics](#)[Inferential Analysis](#)[References](#)

[Introduction](#)[Data Preparation](#)[LVS](#)[Working with Data](#)[Visual Analytics](#)[Inferential Analysis](#)[References](#)**Choose colour**

- blue
- red
- green
- grey

Enter the title for your plot

My plot ←

Name for your x-axis

Genre ←



Saving Plot

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References





Cluster Plot

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

- Classification of data into **sub-groups** is based on **pairwise similarities**
- Groups are clustered in the form of a **tree-like dendrogram**

One Variable Plot

Two Variables Plot

Three Variables Plot

Cluster Plot

Frequency Plot



Cluster Plot

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

Variable must contain at least three values to be clustered.

Your dependent variable

NULL

One independent variable for cluster

NULL

Your dependent variable

NULL

actor_1_facebook_likes
genres
actor_1_name
movie_title
cast_total_facebook_likes
budget ←
title_year
movie_facebook_likes

One independent variable for cluster

NULL

NULL
director_name
director_facebook_likes
actor_1_facebook_likes
genres ←
actor_1_name
movie_title
cast_total_facebook_likes

Cluster Plot

Introduction

Data Preparation

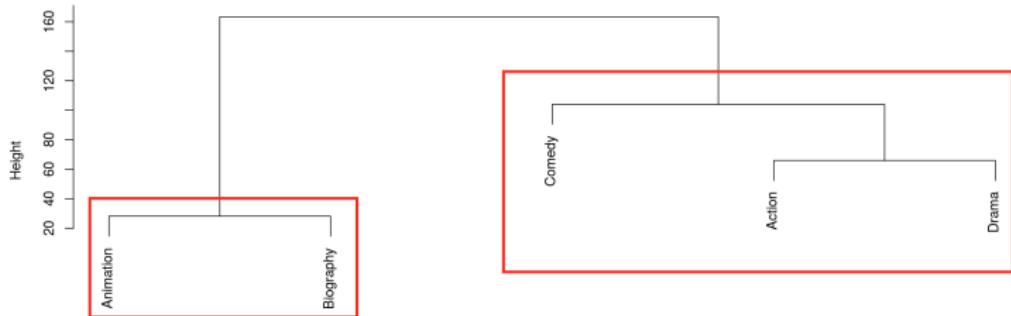
LVS

Working with Data

Visual Analytics

Inferential Analysis

References



Group 1 **Animation, Biography** and Group 2 **Action, Drama, Comedy**



Inferential Statistics

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References





Language Variation Suite - Structure

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

Language Variation Suite (LVS)

About Demo Data Visual Analysis RBRUL Inferential Statistics

① Data

- Upload file, data summary, adjust data, cross tabulation

② Visual Analysis

- Plotting, cluster classification

③ Inferential statistics

- Modeling, regression, varbrul analysis, conditional trees, random forest



How to Create a Regression Model

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

Modeling

Regression

Stepwise Regression

Varbrul Analysis

Conditional Trees

Random Forest

Step 1 Modeling - create a model with dependent and independent variables

Step 2 Regression - specify the type of regression (fixed, mixed) and type of dependent variable (binary, continuous, multinomial)

Step 3 Stepwise Regression - compare models
(Log-likelihood, AIC, BIC)

Step 4 Conditional Trees - apply non-parametric tests to the model



Modeling

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

Modeling

Regression

Stepwise Regression

Varbrul Analysis

Conditional Trees

Random Forest

Select one dependent variable

Choose one column:

NULL

director_facebook_likes
actor_1_facebook_likes
genres
actor_1_name
movie_title
cast_total_facebook_likes
budget
title_year

Select one or more independent variables

Choose columns:

|

director_name
director_facebook_likes
actor_1_facebook_likes
genres
actor_1_name
movie_title
cast_total_facebook_likes
budget



Regression Types

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

● Model

- a.) Fixed effect
- b.) Mixed effect - individual speaker/token variation (within group)

● Type of Dependent Variable

- a.) Binary/categorical (only two values)
- b.) Continuous (numeric)
- c.) Multinomial - categorical with more than two values



Regression

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

Modeling

Regression

Stepwise Regression

Varbrul Analysis

Conditional Trees

Random Forest

Type of Regression Model

Models

NULL

NULL

Fixed Effect Model

Mixed Effect Model

Type of Dependent Variable

Dependent Variable

NULL

NULL

binary

continuous

multinomial



Model Output

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

```
Call:  
lm(formula = as.formula(paste(y, paste(listfactors, collapse = "+"),  
sep = "~")), data = plotData(), na.action = na.omit)  
  
Residuals:  
      Min       1Q   Median       3Q      Max  
-82717225 -21661439  -7880755  12838561 234326343  
  
Coefficients:  
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) 65673657  1421800  46.19 < 2e-16 ***  
genresAnimation 17193568  3262681   5.27 1.46e-07 ***  
genresBiography -39136672  3650154  -10.72 < 2e-16 ***  
genresComedy    -38512218  1791456  -21.50 < 2e-16 ***  
genresDrama     -43279374  1979184  -21.87 < 2e-16 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 38620000 on 3081 degrees of freedom  
Multiple R-squared:  0.2165,   Adjusted R-squared:  0.2154  
F-statistic: 212.8 on 4 and 3081 DF, p-value: < 2.2e-16
```

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

```
Call:  
lm(formula = as.formula(paste(y, paste(listfactors, collapse = "+"),  
sep = "~")), data = plotData(), na.action = na.omit)  
  
Residuals:  
      Min        1Q    Median        3Q       Max  
-82717225 -21661439  -7880755   12838561 234326343  
  
Coefficients:  
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) 65673657  1421800  46.19 < 2e-16 ***  
genresAnimation 17193568  3262681   5.27 1.46e-07 ***  
genresBiography -39136672  3650154  -10.72 < 2e-16 ***  
genresComedy   -38512218  1791456  -21.50 < 2e-16 ***  
genresDrama     -43279374  1979184  -21.87 < 2e-16 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 38620000 on 3081 degrees of freedom  
Multiple R-squared:  0.2165,   Adjusted R-squared:  0.2154  
F-statistic: 212.8 on 4 and 3081 DF,  p-value: < 2.2e-16
```





Interpretation: Budget and Genre

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

Coefficients:						
		Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	65673657	1421800	46.19	< 2e-16	***	
genresAnimation	17193568	3262681	5.27	1.46e-07	***	
genresBiography	-39136672	3650154	-10.72	< 2e-16	***	
genresComedy	-38512218	1791456	-21.50	< 2e-16	***	
genresDrama	-43279374	1979184	-21.87	< 2e-16	***	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

- Genre **Action** is the reference value
- Positive coefficient - positive effect
- Negative coefficient - negative effect

<http://www.free-online-calculator-use.com/scientific-notation-converter.html>

Interpretation: Budget and Genre

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

exponential notation:

Coefficients:						
		Estimate	Std. Error	t value	Pr(> t)	1.46e-7
	(Intercept)	65673657	1421800	46.19	< 2e-16	*** .0000000146
	genresAnimation	17193568	3262681	5.27	1.46e-07	*** 87654321
	genresBiography	-39136672	3650154	-10.72	< 2e-16	*** 0.000000146
	genresComedy	-38512218	1791456	-21.50	< 2e-16	***
	genresDrama	-43279374	1979184	-21.87	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

- Genre **Action** is the reference value
- Positive coefficient - positive effect
- Negative coefficient - negative effect

<http://www.free-online-calculator-use.com/scientific-notation-converter.html>

Conditional Tree

Introduction

Data Preparation

LVS

Working with Data

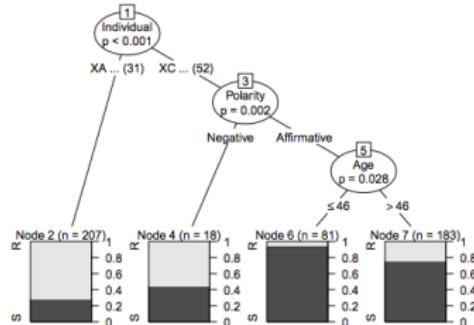
Visual Analytics

Inferential Analysis

References

Conditional tree: a simple non-parametric regression analysis, commonly used in social and psychological studies

- Linear regression: all information is combined linearly
- Conditional tree regression: visual splitting to capture interaction between variables



Recursive splitting (tree branches)



Conditional Tree

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

Modeling Regression Stepwise Regression Varbrul Analysis **Conditional Trees** Random Forest

Select Apply

- none
 apply

[1] "Dependent Variable: budget Independent Variables: genres"

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

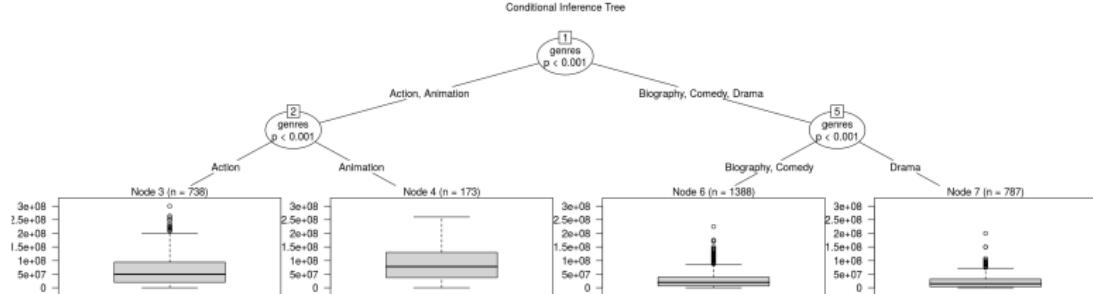
References

[Modeling](#) [Regression](#) [Stepwise Regression](#) [Varbrul Analysis](#) [Conditional Trees](#) [Random Forest](#)

Select Apply

 none apply

[1] "Dependent Variable: budget Independent Variables: genres"



Introduction

Data Preparation

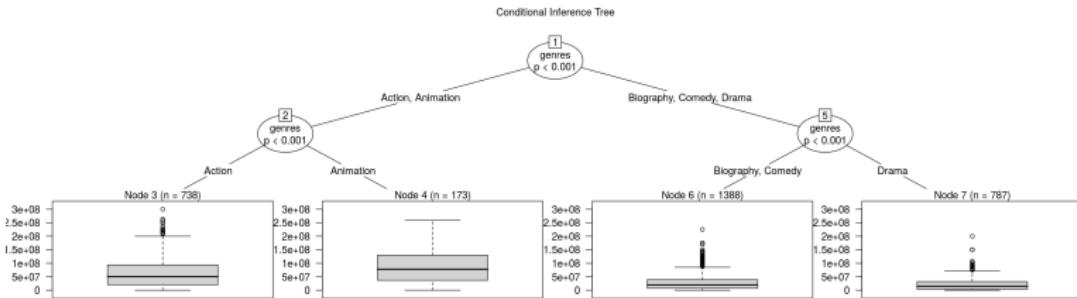
LVS

Working with Data

Visual Analytics

Inferential Analysis

References



- ① **Genre** is the significant factor for budget
- ② Budget distribution is split in two groups:
 - Action and Animation
 - Biography, Comedy and Drama
- ③ Budget is significantly higher for Animation and Action



Random Forest

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

- ① Variable importance for predictors
- ② Robust technique with *small n large p* data
- ③ All predictors considered jointly (allows for inclusion of correlated factors)





Random Forest

Introduction

Data Preparation

LVS

Working with Data

Visual Analytics

Inferential Analysis

References

Let's add more factors!

- Return to **Modeling**

[Modeling](#) [Regression](#) [Stepwise Regression](#) [Varbrul Analysis](#) [Conditional Trees](#) [Random Forest](#)

- Add independent factors: **director facebook likes, actor 1 facebook likes, title year**

Choose columns:

genres director_facebook_likes
actor_1_facebook_likes title_year |

director_name

actor_1_name

movie_title

cast_total_facebook_likes

budget

movie_facebook_likes

Select one dependent variable

Choose one column:

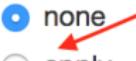
budget

The same dependent variable



Random Forest

Select Apply

- none
 apply
- 

Select Apply

- none
 apply

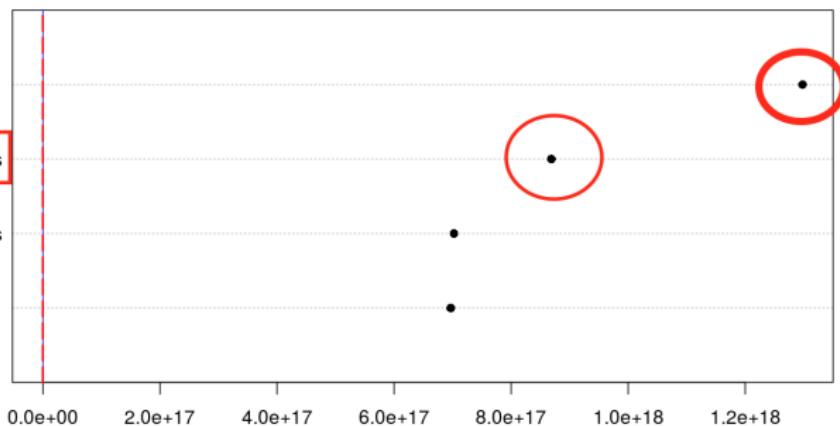
genres

actor_1_facebook_likes

director_facebook_likes

title_year

Variable Importance for budget



- Genre is the most important predictor for this model.
- Close to zero or red-dotted line (cut off values) - irrelevant for this model



Let's Have a Short Break

Introduction

Data
Preparation

LVS

Working with
Data

Visual
Analytics

Inferential
Analysis

References



References I

Introduction

Data
Preparation

LVS

Working with
Data

Visual
Analytics

Inferential
Analysis

References

- [1] Baayen, Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics*. Cambridge: Cambridge University Press
- [2] Gries, Stefan Th. 2015. *Quantitative designs and statistical techniques*. In Douglas Biber Randi Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press
- [3] Schnapp, Jeffrey, and Peter Presner. 2009. Digital Humanities Manifesto 2.0.
- [4] http://gifsanimados.espaciolatino.com/x_bob_esponja_8.gif
- [5] <https://daniellestolt.files.wordpress.com/2013/01/are-you-ready1.gif>
- [6] <http://www.martijnwieling.nl/R/sheets.pdf>

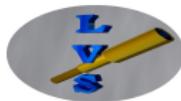


Introduction
ITMS
Preprocessing Data
Data Visualization
Cluster Analysis
Topic Modeling
Google Book API
Future Directions
References

Interactive Visual Data Analysis Part Two Interactive Text Mining Suite

Olga Scrivner

Indiana University



Workshop in Methods



Introduction

ITMS

Preprocessing
DataData
VisualizationCluster
AnalysisTopic
ModelingGoogle Book
APIFuture
Directions

References

- ① Introduce a web application for text processing and mining
- ② Learn about natural language processing techniques
- ③ Develop practical skills





Data Mining

Introduction

ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

Topic
Modeling

Google Book
API

Future
Directions

References

“As our collective knowledge continues to be digitized and stored (...) it becomes more difficult to find and discover what we are looking for.” (Blei 2012)

New Ways of Exploring Data Collections

Introduction

ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

Topic
Modeling

Google Book
API

Future
Directions

References

Word clouds (Vuillemot et al., 2009)



Fig. 4: Chapter 6: the word "dead" highlighted in DotsCloud visualization

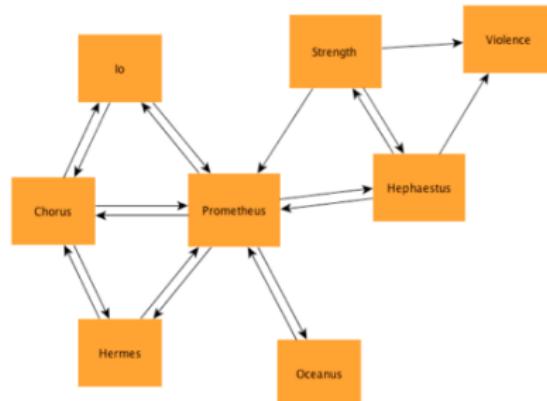
Introduction

ITMS

Preprocessing
DataData
VisualizationCluster
AnalysisTopic
ModelingGoogle Book
APIFuture
Directions

References

● Social network graphs (Rydberg-Cox, 2011)



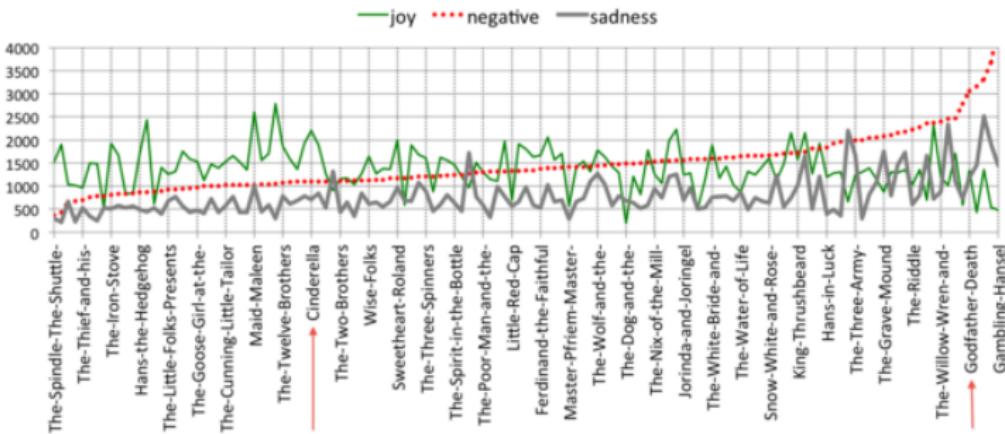
Introduction

ITMS

Preprocessing
DataData
VisualizationCluster
AnalysisTopic
ModelingGoogle Book
APIFuture
Directions

References

- Tracking emotion and sentiment in fairy tales
(Mohammad, 2012)



Topic Modeling

Introduction

ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

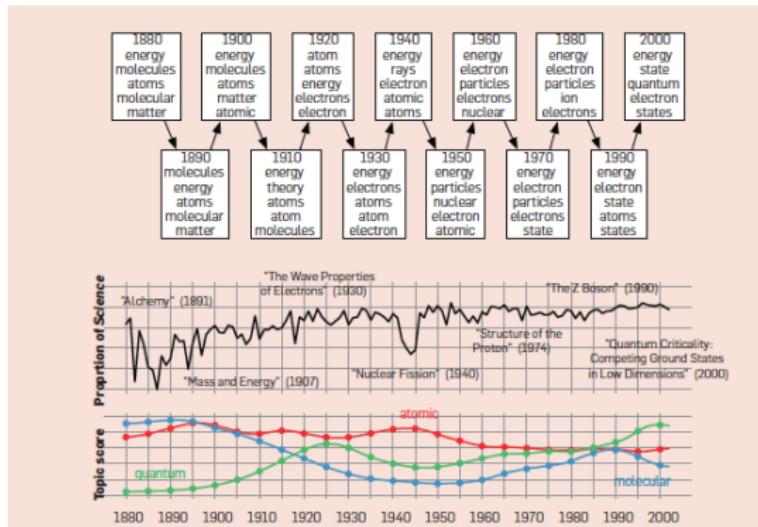
Topic
Modeling

Google Book
API

Future
Directions

References

Discovering underlying theme of collection from *Science* magazine 1990-2000 (Blei 2012)



Technological and Methodological Obstacles

Introduction

ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

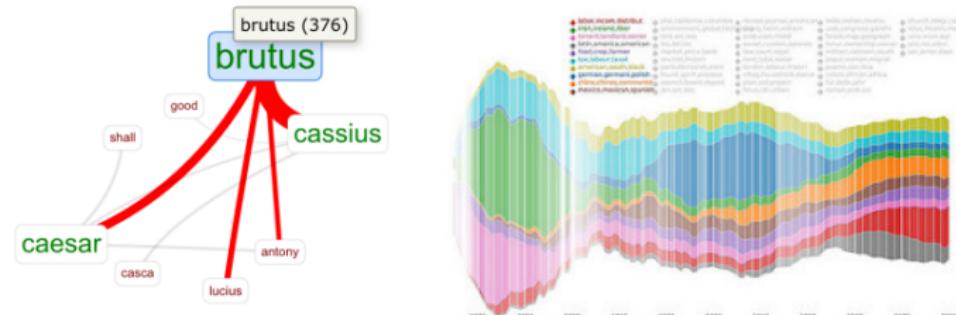
Topic
Modeling

Google Book
API

Future
Directions

References

- Many tools require some programming skills (Mallet, Meta, R and Python libraries)
- GUI tools are limited to certain formats and functions (Voyant, PaperMachine)
- Lack of active control by users





Interactive Text Mining Suite

Introduction

ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

Topic
Modeling

Google Book
API

Future
Directions

References

- A user-friendly tool for quantitative analysis and visualization of unstructured data
- Platform-independent
- Interactive





ITMS Structure

The screenshot shows a top navigation bar with tabs: About, File Uploads (which is active), Data Preparation, and Data Visualization (with a dropdown menu). The dropdown menu for Data Visualization contains three items: Word Frequency, Cluster Analysis, and Topic Analysis.

① File Uploads

- Upload files (txt, pdf, rdf and Google books API)

② Data Preparation

- Data preprocessing (stopwords, stemming, metadata)

③ Data Visualization

- Word frequencies, Cluster analysis and topic modeling



ITMS Structure

The screenshot shows a top navigation bar with tabs: About, File Uploads (which is active), Data Preparation, and Data Visualization. The Data Visualization tab has a dropdown menu with three options: Word Frequency, Cluster Analysis, and Topic Analysis.

① File Uploads

- Upload files (txt, pdf, rdf and Google books API)

② Data Preparation

- Data preprocessing (stopwords, stemming, metadata)

③ Data Visualization

- Word frequencies, Cluster analysis and topic modeling



Workshop Files

Introduction

ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

Topic
Modeling

Google Book
API

Future
Directions

References

- Download 3 text files

<http://ssrc.indiana.edu/seminars/wim.shtml>

- NY Times articles (3 documents in a plain text format)

- ITMS Web site:

<http://www.interactivetextminingsuite.com>





Upload File

Introduction
ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

Topic
Modeling

Google Book
API

Future
Directions

References

About

File Uploads

Data Preparation

Data Visualization ▾

Choose File(s) in TEXT format

Browse...

No file selected

Text Files

PDF Files

ZOTERO

Structured Data

POS-Tagged
Text



Upload File

Introduction
ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

Topic
Modeling

Google Book
API

Future
Directions

References

About

File Uploads

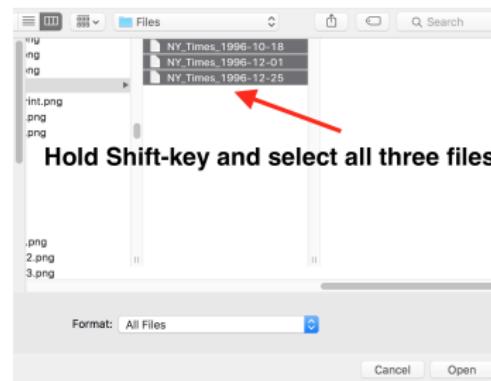
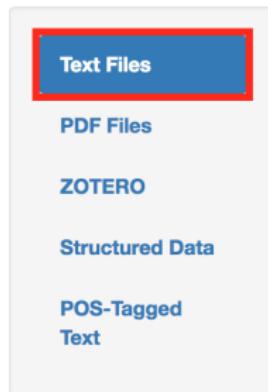
Data Preparation

Data Visualization ▾

Choose File(s) in TEXT format

Browse...

No file selected





Upload File

Introduction
ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

Topic
Modeling

Google Book
API

Future
Directions

References

About

File Uploads

Data Preparation

Data Visualization ▾

Choose File(s) in TEXT format

Browse...

No file selected

Text Files

PDF Files

ZOTERO

Structured Data

POS-Tagged
Text

Choose File(s) in TEXT format

Browse...

3 files

Upload complete

[1] "NY_Times_1996-10-18.txt"
"NY_Times_1996-12-01.txt" [3]
"NY_Times_1996-12-25.txt"
Corpus Size Total: 1611





Preprocessing Data

Introduction
ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

Topic
Modeling

Google Book
API

Future
Directions

References

About

File Uploads

Data Preparation

Data Visualization ▾

Before performing data analysis we should preprocess data.

Data Cleaning

Stopwords

Stemming

Metadata

Document:

The New York Times October 18, 1996, Friday, Late Edition Final Two Different Pleas for Change: Excerpts From a Second Senate Debate SECTION: Section B; Page 22; Column 1; Metropolitan Desk LENGTH: 1145 words Following are excerpts from last night's debate in Trenton between Robert G. Torricelli, a Democrat, and Richard A. Zimmer, a Republican, as transcribed by The New York Times. Opening Statements TORRICELLI For several months Dick Zimmer and I have been campaigning around New Jersey, asking for your help to get elected to the United States Senate. I know you're disappointed in the campaign. So am I. It's deteriorated into personal accusations and acrimony. But in truth, this campaign isn't about me. And it's not about Dick Zimmer. It's about you, your families and your future. It's also not about taxes or spending. I voted for a tax cut last year. So did Dick Zimmer. I voted for the

Introduction

ITMS

Preprocessing
DataData
VisualizationCluster
AnalysisTopic
ModelingGoogle Book
APIFuture
Directions

References

Select preprocessing options and click **apply**.

Select Preprocessing Steps

Remove Punctuation



Exceptions (keep hyphen or apostrophe)

none

both

hyphen

Preprocessing Viewer

Apply Steps or Default (no preprocessing)

apply

default

Lower Case



Remove Numbers



Stopwords (e.g. **the**, **and**): select **Default** for English

The screenshot shows a software interface with a sidebar on the left containing links: Introduction, ITMS, Preprocessing Data, Data Visualization, Cluster Analysis, Topic Modeling, Google Book API, Future Directions, and References. The main area has a title "Select Default or Upload". Under "Data Cleaning", the "Stopwords" button is highlighted with a red box and arrow. To its right are three radio buttons: "None", "Default" (which is selected), and "Upload". Below the radio buttons is a note: "Default is the list from tm package: stopwords('SMART')".

```
[1] "a"           "a's"  
[5] "above"       "accordin  
[9] "actually"    "after"  
[13] "against"    "ain't"  
[17] "allows"      "almost"  
[21] "already"     "also"  
[25] "am"          "among"  
[29] "and"         "another"  
[33] "anyhow"      "anyone"  
[37] "anyways"     "anywhere  
[41] "appreciate"   "appropri  
[45] "around"       "as"  
[49] "asking"        "associat  
[53] "away"         "awfully"
```

Manual Removal of Stopwords

Introduction

ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

Topic
Modeling

Google Book
API

Future
Directions

References

Manual Removal

Select one or multiple words (hold shift key down)

Select words to be removed

made written |

subject
supported
systems
textbooks
training
ultimately
union
voluntary

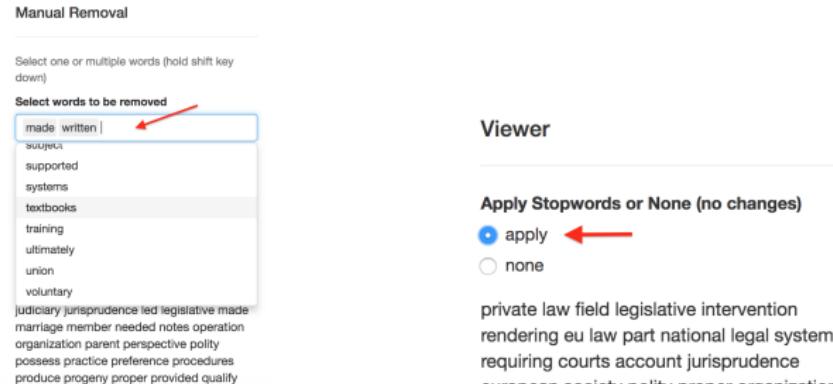
Judiciary jurisprudence led legislative made
marriage member needed notes operation
organization parent perspective polity
possess practice preference procedures
produce progeny proper provided quality

Viewer

Apply Stopwords or None (no changes)

apply
 none

private law field legislative intervention
rendering eu law part national legal system
requiring courts account jurisprudence
protection society policy human organization

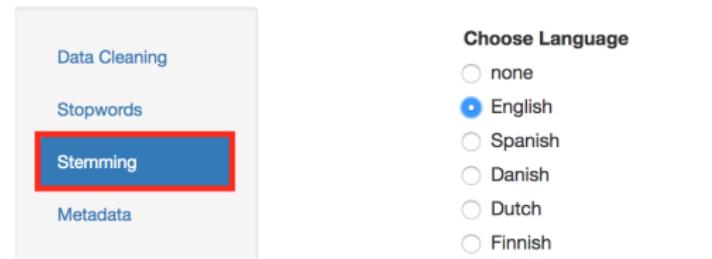


Select **apply**

[Introduction](#)[ITMS](#)[Preprocessing
Data](#)[Data
Visualization](#)[Cluster
Analysis](#)[Topic
Modeling](#)[Google Book
API](#)[Future
Directions](#)[References](#)

To improve analytics, you can stem all your tokens, ex. instead of **worked**, **works**, **working**, you will have only one relevant stem **work**

Stems - tm package



Stem Viewer

privat law field legisl intervient render eu law part nation legal system requir court account jurisprud european societi politi proper organ oper legal system law appli studi sourc led reconfigur common law legal famili parent legal system enter marriag give rise progeni privat european communiti act european union law legal effect nation legal system basi nation court requir appli eu law remain conceptu strike heart domest legal system hold state court subject eu law requir note prefer australia canada zealand legal system analysi support earlier studi year train need qualifi practic favour legal system possess subconsci bias system procedur vis vis member state embedd legal system conceiv state complianc eu law voluntari act reli extent echr case law part uk legal system dealt textbook uk academ general access english ultim produc judgment higher qualiti give judiciari perspect legal system court benefit fulli insight provid compar law



Metadata Extraction

Introduction

ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

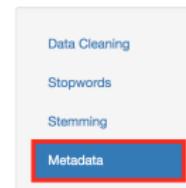
Topic
Modeling

Google Book
API

Future
Directions

References

You can extract or upload metadata. You will need datestamp (year) information for chronological topic modeling.



The screenshot shows a user interface for extracting metadata. At the top, there are buttons for 'Show 25 entries' and 'Search'. Below this is a table with columns: date, title, and author. One entry is listed: '2015 Gilliker - 2015 - The Influence of Eu and European Human Rights Law .pdf'. Below the table are input fields for 'date', 'title', and 'author', followed by a 'Previous' button, a page number '1', and a 'Next' button. At the bottom left, there is a section titled 'Choose metadata source' with the following options:

- None
- From metadata of each uploaded PDF
- From separate CSV file
- From separate JSON file
- From separate XML file
- From zotero files metadata

A red arrow points to the last option, 'From zotero files metadata'.

date	title	author
2015	Gilliker - 2015 - The Influence of Eu and European Human Rights Law .pdf	Gilliker

Show 25 entries Search: _____

date title author

2015 Gilliker - 2015 - The Influence of Eu and European Human Rights Law .pdf Gilliker

date title author

Showing 1 to 1 of 1 entries Previous 1 Next

Choose metadata source

- None
- From metadata of each uploaded PDF
- From separate CSV file
- From separate JSON file
- From separate XML file
- From zotero files metadata



Visualization

Introduction
ITMS
Preprocessing Data
Data Visualization
Cluster Analysis
Topic Modeling
Google Book API
Future Directions
References

File Uploads

Data Preparation

Data Visualization

Frequency Table

Word Clouds

Length

KWIC

Punctuation

Data Visualization ▾

Word Frequency

Cluster Analysis

Topic Analysis



Customization

Introduction

ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

Topic
Modeling

Google Book
API

Future
Directions

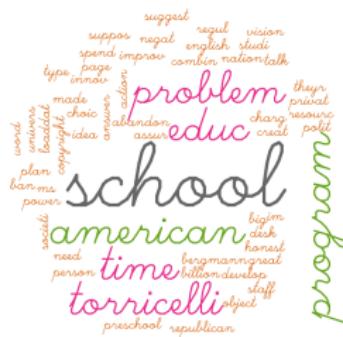
References

Select Font

- Sans Serif
- Script
- Gothic

Select Color Palette

- black
- green
- multi





Cluster Analysis

- Introduction
- ITMS
- Preprocessing Data
- Data Visualization
- Cluster Analysis
- Topic Modeling
- Google Book API
- Future Directions
- References

File Uploads

Data Preparation

Data Visualization

Word Frequency

Cluster Analysis

Topic Analysis

You need to have at least **three** documents

Documents will be grouped based on their term similarity measures

Agglomeration Methods

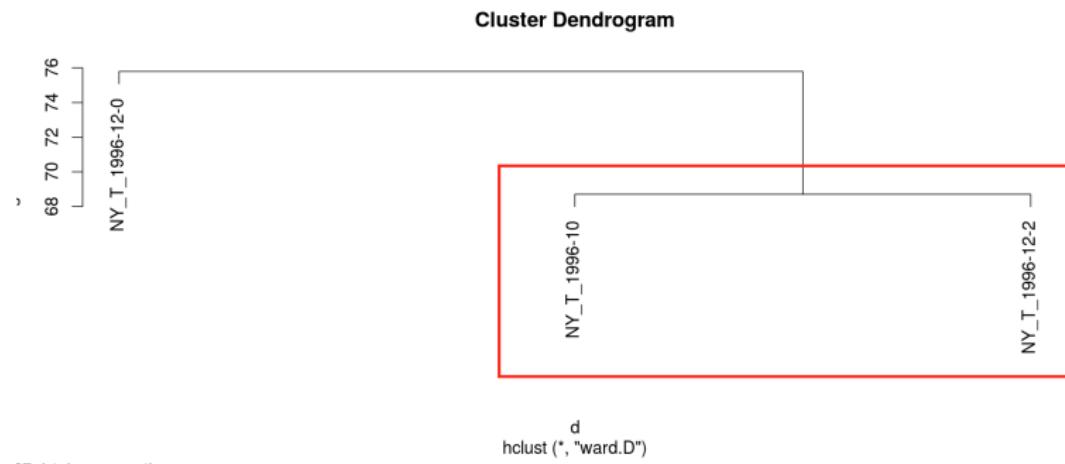
Select method for cluster groups

- ward.D
- single
- complete
- average
- median
- centroid

Distance Measure

Select measure type

- euclidean
- maximum
- manhattan
- minkowski
- canberra
- binary

[Introduction](#)[ITMS](#)[Preprocessing
Data](#)[Data
Visualization](#)[Cluster
Analysis](#)[Topic
Modeling](#)[Google Book
API](#)[Future
Directions](#)[References](#)



Topic Modeling

Introduction

ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

Topic
Modeling

Google Book
API

Future
Directions

References

- **LDA** (Latent Dirichlet allocation)
- **STM** (Structural Topic model)
- Chronological topic visualization (lda): requires metadata



Topic Modeling Tuning

Introduction

ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

Topic
Modeling

Google Book
API

Future
Directions

References

- Selection of topics (how many different themes)
- Selection of words per theme (how many words per topic)
- Selection of iteration



Topic Model Selection

Introduction

ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

Topic
Modeling

Google Book
API

Future
Directions

References

File Uploads

Data Preparation

Data Visualization

Word Frequency

Cluster Analysis

Topic Analysis

Model Creation

LDA Visualization

STM Visualization

Metadata Topic
Visualization

Topic selection

Select number of topics - an integer representing the number of topics in the model. Default is 3.

Select or Type Number of Topics

3

Select the top number of words associated with a given topic. Default is 3.

Select or Type Number of Words per Topic

3



LDA Topic Model

Introduction

ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

Topic
Modeling

Google Book
API

Future
Directions

References

Model Creation

LDA Visualization

STM Visualization

Metadata Topic
Visualization

Run LDA Analysis

none

run

Selected Topics LDA (`lda.collapsed.gibbs`
package)

V1	V2	V3
policy	children	public
evidence	care	zimmer
president	vouchers	schools



STM Topic Model

Introduction

ITMS

Preprocessing
Data

Data
Visualization

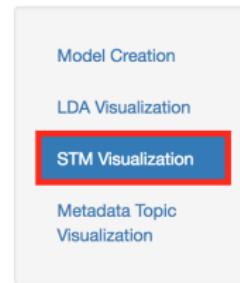
Cluster
Analysis

Topic
Modeling

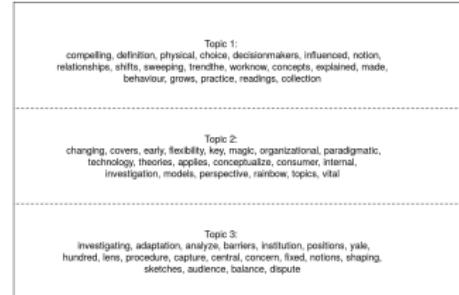
Google Book
API

Future
Directions

References



Structural Topics STM





Other Formats - Google Books

Introduction

ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

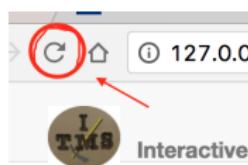
Topic
Modeling

Google Book
API

Future
Directions

References

Before switching to other data formats, refresh your local browser.



Start with **File Uploads** and select **Structured Data**

Text Files

PDF Files

ZOTERO

Structured Data

POS-Tagged Text



Other Formats - Google Books

Introduction

ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

Topic
Modeling

Google Book
API

Future
Directions

References

Select your search terms and submit

Choose file format

XML

JSON

Google Books Search

Enter your search terms for Google Books,
separated by spaces

social science

Submit

Current limitation is 40 books

Show 25 entries

Search:

titles	authors	dates	corpus
Readings in the Philosophy of Social Science	Michael Martin, Lee C. McIntyre	1994	Readings in the Philosophy of Social Science the first comprehensive anthology in the



Future Options

Introduction

ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

Topic
Modeling

Google Book
API

Future
Directions

References

Shiny Web Application is highly customizable

- ① Part-of-speech tagging (tm package)
- ② Network analysis (igraph package)
- ③ Name Entity Recognition (NLP package)
- ④ Twitter Streaming (twitterR package) - will require user's twitter set-up for streaming but information will be provided how to set it up

Open for other suggestions and collaboration - contact
obscrivn@indiana.edu



Acknowledgements

Introduction

ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

Topic
Modeling

Google Book
API

Future
Directions

References

I would like to thank WIM for providing this opportunity.

Contributors: Jefferson Davis, Irina Trapido, Jay Lee



References I

Introduction

ITMS

Preprocessing
Data

Data
Visualization

Cluster
Analysis

Topic
Modeling

Google Book
API

Future
Directions

References

- [1] Many open source R packages: tm, shiny, NLP, stringi, stringr, topicmodels, lda and many more
 - [2] Baayen, Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics*. Cambridge: Cambridge University Press
 - [3] Gries, Stefan Th. 2015. *Quantitative designs and statistical techniques*. In Douglas Biber Randi Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press
 - [4] Jockers, Matthew. 2014. Text Analysis with R for Students of Literature. Quantitative Methods in the Humanities and Social Sciences. Springer International Publishing, Cham
 - [5] Moretti, Franco. 2005. Graphs, Maps, Trees: Abstract Models for a Literary History. Verso
 - [6] Oelke, Daniella, Dimitrios Kokkinakis, and Mats Malm. 2012. Advanced visual analytics methods for literature analysis. Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social 561Sciences, and Humanities, pages 35–44
- image credits: <https://media.giphy.com/media/10zsjaH4g0GgmY/giphy.gif>