This class is great, take it.

DS-GA 3001-004, Text as Data
Arthur Spirling

January 24, 2017

Professor and Lectures

Professor and Lectures

Prof Arthur Spirling

Professor and Lectures

Prof Arthur Spirling

arthur.spirling@nyu.edu

Professor and Lectures

Prof Arthur Spirling

`arthur.spirling@nyu.edu`

405, 19 West 4th St.

Professor and Lectures

Prof Arthur Spirling

`arthur.spirling@nyu.edu`

405, 19 West 4th St.

OH Tues, 11AM-12.

Professor and Lectures

Prof Arthur Spirling

arthur.spirling@nyu.edu

405, 19 West 4th St.

OH Tues, 11AM-12.

Lect Tuesdays 3:30 PM - 5:10 PM Lecture 60 Fifth Avenue, 150 (except Feb 21!)

TA and Sections

TA and Sections

Patrick Chester

TA and Sections

Patrick Chester

`pjc468@nyu.edu`

TA and Sections

Patrick Chester

pjc468@nyu.edu

421, 19 West 4th St.

TA and Sections

Patrick Chester

pjc468@nyu.edu

421, 19 West 4th St.

OH Mon, 1–2pm

TA and Sections

Patrick Chester

pjc468@nyu.edu

421, 19 West 4th St.

OH Mon, 1–2pm

Sec Tuesday, 7:30–820pm, 60 Fifth Avenue, 110 (start today!) (except Feb 21!)

What this class is about. . .

What this class is about...

What this class is about...

Text as the new frontier of...

What this class is about. . .

Text as the new frontier of. . .

data:

What this class is about...

Text as the new frontier of...

> data: lots of it (literally petabytes) on the web... not to mention archives.

What this class is about. . .

Text as the new frontier of. . .

> data: lots of it (literally petabytes) on the web. . . not to mention archives.
>
> methods:

What this class is about...

Text as the new frontier of...

data: lots of it (literally petabytes) on the web...not to mention archives.

methods: unstructured data needs to be harvested and modeled.

What this class is about. . .

Text as the new frontier of. . .

data: lots of it (literally petabytes) on the web. . . not to mention archives.

methods: unstructured data needs to be harvested and modeled.

social science:

What this class is about. . .

Text as the new frontier of. . .

data: lots of it (literally petabytes) on the web. . . not to mention archives.

methods: unstructured data needs to be harvested and modeled.

social science: politicians give speeches,

What this class is about...

Text as the new frontier of...

data: lots of it (literally petabytes) on the web... not to mention archives.

methods: unstructured data needs to be harvested and modeled.

social science: politicians give speeches, thinkers write articles,

What this class is about...

Text as the new frontier of...

data: lots of it (literally petabytes) on the web... not to mention archives.

methods: unstructured data needs to be harvested and modeled.

social science: politicians give speeches, thinkers write articles, nations sign treaties,

What this class is about. . .

Text as the new frontier of. . .

data: lots of it (literally petabytes) on the web. . . not to mention archives.

methods: unstructured data needs to be harvested and modeled.

social science: politicians give speeches, thinkers write articles, nations sign treaties, users connect on Facebook etc.

What this class is about...

Text as the new frontier of...

data: lots of it (literally petabytes) on the web... not to mention archives.

methods: unstructured data needs to be harvested and modeled.

social science: politicians give speeches, thinkers write articles, nations sign treaties, users connect on Facebook etc.

Introduction to quantitative 'text-as-data' approaches as strategies to learn more about social scientific phenomena of interest.

Overview

Overview


since
race
stand
responsibility
parents
no
live
law
together
matter
south
need

Overview

- Descriptive inference:

Overview

- Descriptive inference: how to characterize text,

Overview

- Descriptive inference: how to characterize text, vector space model,

Overview

- Descriptive inference: how to characterize text, vector space model, collocations,

Overview

- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words,

Overview

- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words, dissimilarity measures,

Overview

- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words, dissimilarity measures, diversity,

Overview

- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words, dissimilarity measures, diversity, complexity.

Overview

- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words, dissimilarity measures, diversity, complexity.

- Supervised techniques:

Overview

- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words, dissimilarity measures, diversity, complexity.

- Supervised techniques: dictionaries,

Overview

- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words, dissimilarity measures, diversity, complexity.
- Supervised techniques: dictionaries, classification,

Overview

- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words, dissimilarity measures, diversity, complexity.
- Supervised techniques: dictionaries, classification, scaling,

Overview

- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words, dissimilarity measures, diversity, complexity.

- Supervised techniques: dictionaries, classification, scaling, machine learning approaches.

Overview

- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words, dissimilarity measures, diversity, complexity.
- Supervised techniques: dictionaries, classification, scaling, machine learning approaches.
- Unsupervised techniques:

Overview

- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words, dissimilarity measures, diversity, complexity.
- Supervised techniques: dictionaries, classification, scaling, machine learning approaches.
- Unsupervised techniques: clustering,

Overview

- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words, dissimilarity measures, diversity, complexity.
- Supervised techniques: dictionaries, classification, scaling, machine learning approaches.
- Unsupervised techniques: clustering, principal components,

Overview

- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words, dissimilarity measures, diversity, complexity.
- Supervised techniques: dictionaries, classification, scaling, machine learning approaches.
- Unsupervised techniques: clustering, principal components, scaling,

Overview

- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words, dissimilarity measures, diversity, complexity.
- Supervised techniques: dictionaries, classification, scaling, machine learning approaches.
- Unsupervised techniques: clustering, principal components, scaling, topic models.

Overview

- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words, dissimilarity measures, diversity, complexity.
- Supervised techniques: dictionaries, classification, scaling, machine learning approaches.
- Unsupervised techniques: clustering, principal components, scaling, topic models.
- Special topics:

Overview

- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words, dissimilarity measures, diversity, complexity.
- Supervised techniques: dictionaries, classification, scaling, machine learning approaches.
- Unsupervised techniques: clustering, principal components, scaling, topic models.
- Special topics: modeling debate,

Overview

- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words, dissimilarity measures, diversity, complexity.
- Supervised techniques: dictionaries, classification, scaling, machine learning approaches.
- Unsupervised techniques: clustering, principal components, scaling, topic models.
- Special topics: modeling debate, text reuse,

Overview

- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words, dissimilarity measures, diversity, complexity.
- Supervised techniques: dictionaries, classification, scaling, machine learning approaches.
- Unsupervised techniques: clustering, principal components, scaling, topic models.
- Special topics: modeling debate, text reuse, networks,

Overview

- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words, dissimilarity measures, diversity, complexity.
- Supervised techniques: dictionaries, classification, scaling, machine learning approaches.
- Unsupervised techniques: clustering, principal components, scaling, topic models.
- Special topics: modeling debate, text reuse, networks, bursts.

Quantitative vs Qualitative

Quantitative vs Qualitative

Quantitative vs Qualitative

- For most of its history text analysis was qualitative.

Quantitative vs Qualitative

- For most of its history text analysis was qualitative.
- Partly still is:

Quantitative vs Qualitative

- For most of its history text analysis was qualitative.
- Partly still is: need to make qualitative judgements about what the text reveals,

Quantitative vs Qualitative

- For most of its history text analysis was qualitative.
- Partly still is: need to make qualitative judgements about what the text reveals, and validation requires substantive knowledge

Quantitative vs Qualitative

- For most of its history text analysis was qualitative.
- Partly still is: need to make qualitative judgements about what the text reveals, and validation requires substantive knowledge
- ‘Distant reading’ instead of ‘close reading’

Quantitative vs Qualitative

- For most of its history text analysis was qualitative.
- Partly still is: need to make qualitative judgements about what the text reveals, and validation requires substantive knowledge
- ‘Distant reading’ instead of ‘close reading’—not focussed on interpretation in light of norms or belief systems.

Quantitative vs Qualitative

- For most of its history text analysis was qualitative.
- Partly still is: need to make qualitative judgements about what the text reveals, and validation requires substantive knowledge
- 'Distant reading' instead of 'close reading'—not focussed on interpretation in light of norms or belief systems.
- Important: quantitative work is reliable and replicable (easily)

Quantitative vs Qualitative

- For most of its history text analysis was qualitative.
- Partly still is: need to make qualitative judgements about what the text reveals, and validation requires substantive knowledge
- 'Distant reading' instead of 'close reading'—not focussed on interpretation in light of norms or belief systems.
- Important: quantitative work is reliable and replicable (easily) and can cope with large volume of material.

What this class is not about. . .

What this class is not about. . .

What this class is not about...

- Data acquisition:

What this class is not about...

- Data acquisition: many source of text,

What this class is not about...

- Data acquisition: many source of text, but web-scraping better taught elsewhere.

What this class is not about...

- Data acquisition: many source of text, but web-scraping better taught elsewhere.

e.g. `http://www.crummy.com/software/BeautifulSoup/`

What this class is not about...

- Data acquisition: many source of text, but web-scraping better taught elsewhere.

e.g. `http://www.crummy.com/software/BeautifulSoup/`

- Regular expressions and basic text manipulation:

What this class is not about...

- Data acquisition: many source of text, but web-scraping better taught elsewhere.

e.g. `http://www.crummy.com/software/BeautifulSoup/`

- Regular expressions and basic text manipulation: won't generally be required,

What this class is not about. . .

- Data acquisition: many source of text, but web-scraping better taught elsewhere.

e.g. `http://www.crummy.com/software/BeautifulSoup/`

- Regular expressions and basic text manipulation: won't generally be required, though helpful if known.

What this class is not about...

- Data acquisition: many source of text, but web-scraping better taught elsewhere.

e.g. `http://www.crummy.com/software/BeautifulSoup/`

- Regular expressions and basic text manipulation: won't generally be required, though helpful if known.

- CS 'stuff' like machine translation, OCR, algorithm design etc.

What this class is not about. . .

- Data acquisition: many source of text, but web-scraping better taught elsewhere.

e.g. `http://www.crummy.com/software/BeautifulSoup/`

- Regular expressions and basic text manipulation: won't generally be required, though helpful if known.

- CS 'stuff' like machine translation, OCR, algorithm design etc.

→ excellent options elsewhere.

Requirements

Requirements

Requirements

- A first class in statistics and/or inference

Requirements

- A first class in statistics and/or inference
- Basic knowledge of calculus, probability, densities, distributions, statistical tests, hypothesis testing, the linear model, maximum likelihood and generalized linear models is assumed.

Requirements

- A first class in statistics and/or inference

- Basic knowledge of calculus, probability, densities, distributions, statistical tests, hypothesis testing, the linear model, maximum likelihood and generalized linear models is assumed.

- Familiarity with core language and software environment of this course: `R`.

Requirements

- A first class in statistics and/or inference
- Basic knowledge of calculus, probability, densities, distributions, statistical tests, hypothesis testing, the linear model, maximum likelihood and generalized linear models is assumed.
- Familiarity with core language and software environment of this course: `R`.

→ check in with me if unsure.

Terminology

Terminology

- sample is $x = \{3, 4, 20, 50, 68\}$.

Terminology

- sample is $x = \{3, 4, 20, 50, 68\}$. Problem requires calculation of maximum of discrete,

Terminology

- sample is $x = \{3, 4, 20, 50, 68\}$. Problem requires calculation of maximum of discrete, uniform

Terminology

- sample is $x = \{3, 4, 20, 50, 68\}$. Problem requires calculation of maximum of discrete, uniform distribution, $\hat{N}$.

Terminology

- sample is $x = \{3, 4, 20, 50, 68\}$. Problem requires calculation of maximum of discrete, uniform distribution, $\hat{N}$.
- MLE is obvious,

Terminology

- sample is $x = \{3, 4, 20, 50, 68\}$. Problem requires calculation of maximum of discrete, uniform distribution, $\hat{N}$.

- MLE is obvious, $\max(x) = m = \hat{N} = 68$,

Terminology

- sample is $x = \{3, 4, 20, 50, 68\}$. Problem requires calculation of maximum of discrete, uniform distribution, $\hat{N}$.

- MLE is obvious, $\max(x) = m = \hat{N} = 68$, but biased (though consistent).

Terminology

- sample is $x = \{3, 4, 20, 50, 68\}$. Problem requires calculation of maximum of discrete, uniform distribution, $\hat{N}$.
- MLE is obvious, $\max(x) = m = \hat{N} = 68$, but biased (though consistent). Note that max is sufficient statistic in this case.

Terminology

- sample is $x = \{3, 4, 20, 50, 68\}$. Problem requires calculation of maximum of discrete, uniform distribution, $\hat{N}$.
- MLE is obvious, $\max(x) = m = \hat{N} = 68$, but biased (though consistent). Note that max is sufficient statistic in this case.
- Bayesian estimator with uniform prior has closed form for the mean: $E(N|x) = \frac{k-1}{k-2}(m-1)$.

Terminology

- sample is $x = \{3, 4, 20, 50, 68\}$. Problem requires calculation of maximum of discrete, uniform distribution, $\hat{N}$.
- MLE is obvious, $\max(x) = m = \hat{N} = 68$, but biased (though consistent). Note that max is sufficient statistic in this case.
- Bayesian estimator with uniform prior has closed form for the mean: $E(N|x) = \frac{k-1}{k-2}(m-1)$. Thus, (average) $\hat{N} = 89$.

Terminology

- sample is $x = \{3, 4, 20, 50, 68\}$. Problem requires calculation of maximum of discrete, uniform distribution, $\hat{N}$.
- MLE is obvious, $\max(x) = m = \hat{N} = 68$, but biased (though consistent). Note that max is sufficient statistic in this case.
- Bayesian estimator with uniform prior has closed form for the mean: $E(N|x) = \frac{k-1}{k-2}(m-1)$. Thus, (average) $\hat{N} = 89$.

$\rightarrow$ no need for simulation.

Terminology

- sample is $x = \{3, 4, 20, 50, 68\}$. Problem requires calculation of maximum of discrete, uniform distribution, $\hat{N}$.
- MLE is obvious, $\max(x) = m = \hat{N} = 68$, but biased (though consistent). Note that max is sufficient statistic in this case.
- Bayesian estimator with uniform prior has closed form for the mean: $E(N|x) = \frac{k-1}{k-2}(m-1)$. Thus, (average) $\hat{N} = 89$.

$\rightarrow$ no need for simulation.

- Straightforward to implement via function writing in `R`.

sample, discrete, uniform, distribution, $\hat{N}$, MLE, biased, consistent, sufficient statistic, Bayesian, estimator, prior, closed form, simulation, function writing, R `sample`, `discrete`, `uniform`, `distribution`, $\hat{N}$, `MLE`, `biased`, `consistent`, `sufficient statistic`, `Bayesian`, `estimator`, `prior`, `closed form`, `simulation`, `function writing`, `R`

If the terms in blue were familiar to you (even if you can't recall their exact meaning), sample, discrete, uniform, distribution, $\hat{N}$, MLE, biased, consistent, sufficient statistic, Bayesian, estimator, prior, closed form, simulation, function writing, R

If the terms in blue were familiar to you (even if you can't recall their exact meaning), you are probably in good shape for this class.

Software

Software

Software

https://www.r-project.org/

Software

https://www.r-project.org/

Contrary to (un)popular opinion,

Software

https://www.r-project.org/

Contrary to (un)popular opinion, R has excellent text handling/modeling capabilities.

Software

https://www.r-project.org/

Contrary to (un)popular opinion, R has excellent text handling/modeling capabilities.

Free, and massive online community writing packages and extending modeling abilities.

Software

`https://www.r-project.org/`

Contrary to (un)popular opinion, `R` has excellent text handling/modeling capabilities.

Free, and massive online community writing packages and extending modeling abilities.

We will use `quanteda` and other packages.

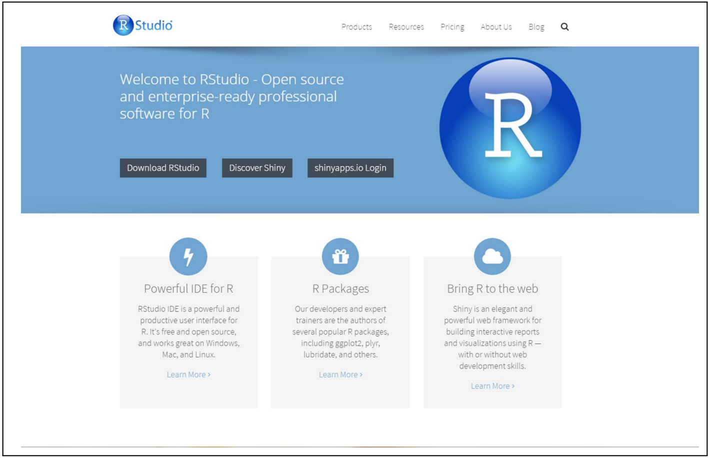
Writing R: RStudio

Writing R: RStudio

https://www.rstudio.com/

Writing R: RStudio

https://www.rstudio.com/



Readings

Readings


since
race
stand
responsibility
parents
no
live
law
together
matter
south
need

Readings

- No required text book(s),

Readings

- No required text book(s), though syllabus has some you might find interesting/helpful if you really want to part with $s.

Readings

- No required text book(s), though syllabus has some you might find interesting/helpful if you really want to part with \$s.
- Generally,

Readings

- No required text book(s), though syllabus has some you might find interesting/helpful if you really want to part with $s.

- Generally, we'll provide the readings on the class webpage,

Readings

- No required text book(s), though syllabus has some you might find interesting/helpful if you really want to part with $s.

- Generally, we'll provide the readings on the class webpage, or tell you where to get them.

Readings

- No required text book(s), though syllabus has some you might find interesting/helpful if you really want to part with $s.
- Generally, we'll provide the readings on the class webpage, or tell you where to get them.
- Substantive readings are especially important,

Readings

- No required text book(s), though syllabus has some you might find interesting/helpful if you really want to part with $s.
- Generally, we'll provide the readings on the class webpage, or tell you where to get them.
- Substantive readings are especially important, because they'll help you understand what an interesting question looks like (in social science).

Assessment

Assessment


since
race
stand
responsibility
parents
live
law
together
matter
south
need
research

Assessment

50%. There will be three homeworks that will enable you to practice your skills.

Assessment

50%. There will be three homeworks that will enable you to practice your skills. While you may consult with others when working, what you hand in must be your own work and no one else's work.

Assessment

50%. There will be three homeworks that will enable you to practice your skills. While you may consult with others when working, what you hand in must be your own work and no one else's work.

If you copy someone's work,

Assessment

50%. There will be three homeworks that will enable you to practice your skills. While you may consult with others when working, what you hand in must be your own work and no one else's work.

If you copy someone's work, or allow someone to copy yours,

Assessment

50%. There will be three homeworks that will enable you to practice your skills. While you may consult with others when working, what you hand in must be your own work and no one else's work.

If you copy someone's work, or allow someone to copy yours, you are breaking the rules of the class (and of NYU),

Assessment

50%. There will be three homeworks that will enable you to practice your skills. While you may consult with others when working, what you hand in must be your own work and no one else's work.

If you copy someone's work, or allow someone to copy yours, you are breaking the rules of the class (and of NYU), and bad things happen: we will report you for academic dishonesty.

Assessment

50%. There will be three homeworks that will enable you to practice your skills. While you may consult with others when working, what you hand in must be your own work and no one else's work.

If you copy someone's work, or allow someone to copy yours, you are breaking the rules of the class (and of NYU), and bad things happen: we will report you for academic dishonesty.

50%. Final paper.

Assessment

50%. There will be three homeworks that will enable you to practice your skills. While you may consult with others when working, what you hand in must be your own work and no one else's work.

If you copy someone's work, or allow someone to copy yours, you are breaking the rules of the class (and of NYU), and bad things happen: we will report you for academic dishonesty.

50%. Final paper. Must concern an original research question. Must involve text-as-data data and methods,

Assessment

50%. There will be three homeworks that will enable you to practice your skills. While you may consult with others when working, what you hand in must be your own work and no one else's work.

If you copy someone's work, or allow someone to copy yours, you are breaking the rules of the class (and of NYU), and bad things happen: we will report you for academic dishonesty.

50%. Final paper. Must concern an original research question. Must involve text-as-data data and methods, and receive approval from me.

Assessment

50%. There will be three homeworks that will enable you to practice your skills. While you may consult with others when working, what you hand in must be your own work and no one else's work.

If you copy someone's work, or allow someone to copy yours, you are breaking the rules of the class (and of NYU), and bad things happen: we will report you for academic dishonesty.

50%. Final paper. Must concern an original research question. Must involve text-as-data data and methods, and receive approval from me. May work in teams of up to two people. Deadline is May 9, 12 pages maximum.

Text as Data Speaker Series

Text as Data Speaker Series

Researchers from all over the country,

Text as Data Speaker Series

Researchers from all over the country, and industry (e.g. Facebook, Google) come to discuss their applied work and get feedback.

Text as Data Speaker Series

Researchers from all over the country, and industry (e.g. Facebook, Google) come to discuss their applied work and get feedback. Mostly social science, NLP and computer science focussed.

Text as Data Speaker Series

Researchers from all over the country, and industry (e.g. Facebook, Google) come to discuss their applied work and get feedback. Mostly social science, NLP and computer science focussed.

Thursdays, 4-530pm, 7th floor of CDS.

Text as Data Speaker Series

Researchers from all over the country, and industry (e.g. Facebook, Google) come to discuss their applied work and get feedback. Mostly social science, NLP and computer science focussed.

Thursdays, 4-530pm, 7th floor of CDS.

Sign up at: `http://cds.nyu.edu/text-data-speaker-series/`

Text as Data Speaker Series

Researchers from all over the country, and industry (e.g. Facebook, Google) come to discuss their applied work and get feedback. Mostly social science, NLP and computer science focussed.

Thursdays, 4-530pm, 7th floor of CDS.

Sign up at: `http://cds.nyu.edu/text-data-speaker-series/`