Campus Irapuato-Salamanca | División de Ingenierías

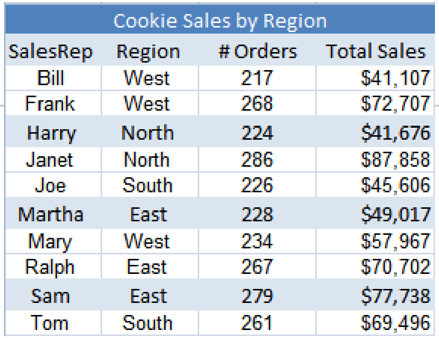
Data Mining

Topic 2: Data/information modalities

© blog.tail.digital

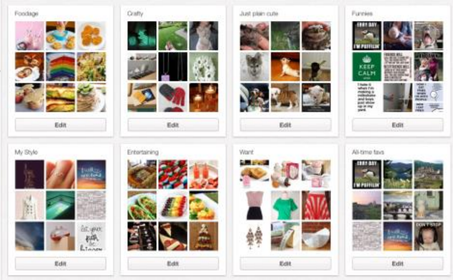
August – December 2021

Juan Carlos Gómez
PhD in Computer Science
jc.gomez@ugto.mx
http://jcgcarranza.wix.com/juancarlosgomez
Office 314

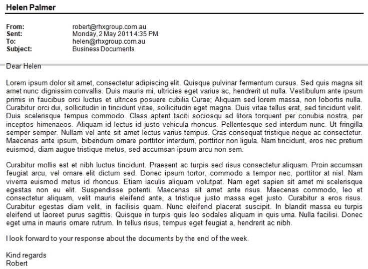


Cookie Sales by Region			
SalesRep	Region	# Orders	Total Sales
Bill	West	217	$41,107
Frank	West	268	$72,707
Harry	North	224	$41,676
Janet	North	286	$87,858
Joe	South	226	$45,606
Martha	East	228	$49,017
Mary	West	234	$57,967
Ralph	East	267	$70,702
Sam	East	279	$77,738
Tom	South	261	$69,496

Structured



Images



Helen Palmer

From: robert@rhxgroup.com.au
Sent: Monday, 2 May 2011 4:35 PM
To: helen@rhxgroup.com.au
Subject: Business Documents

Dear Helen

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Quisque pulvinar fermentum cursus. Sed quis magna sit amet nunc dignissim convallis. Duis mauris mi, ultricies eget varius ac, hendrerit ut nulla. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam sed lorem massa, non lobortis nulla. Curabitur orci dui, sollicitudin in tincidunt vitae, sollicitudin eget magna. Duis vitae tellus erat, sed tincidunt velit. Duis scelerisque tempus commodo. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Aliquam id lectus id justo vehicula rhoncus. Pellentesque sed interdum nunc. Ut fringilla semper semper. Nullam vel ante sit amet lectus varius tempus. Cras consequat tristique neque ac consectetur. Maecenas ante ipsum, bibendum ornare porttitor interdum, porttitor non ligula. Nam tincidunt, eros nec pretium euismod, diam augue tristique metus, sed accumsan ipsum arcu non sem.

Curabitur mollis est et nibh luctus tincidunt. Praesent ac turpis sed risus consectetur aliquam. Proin accumsan feugiat arcu, vel ornare elit dictum sed. Donec ipsum tortor, commodo a tempor nec, porttitor at nisl. Nam viverra euismod metus id rhoncus. Etiam iaculis aliquam volutpat. Nam eget sapien sit amet mi scelerisque egestas non eu elit. Suspendisse potenti. Maecenas sit amet ante risus. Maecenas commodo, leo et consectetur aliquam, velit mauris eleifend ante, a tristique justo massa eget justo. Curabitur a eros risus. Curabitur egestas diam velit, in facilisis quam. Nunc eleifend placerat suscipit. In blandit massa eu turpis eleifend ut laoreet purus sagittis. Quisque in turpis quis leo sodales aliquam in quis urna. Nulla facilisi. Donec eget urna in mauris ornare rutrum. In tellus risus, tempus eget feugiat a, hendrerit ac nibh.

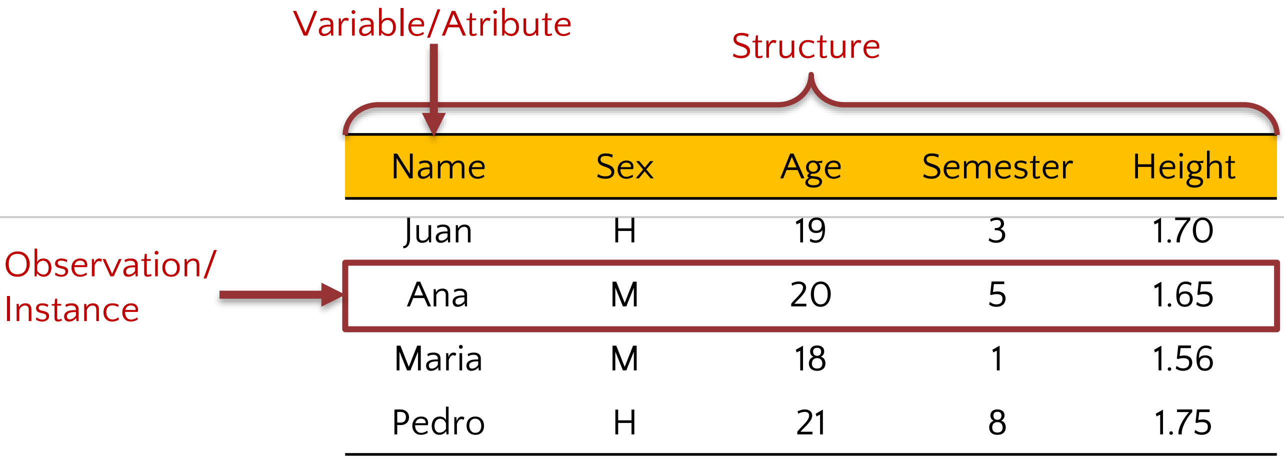
I look forward to your response about the documents by the end of the week.

Kind regards
Robert

Text

Information Modality

The **way**/form in which the information is **presented**.



Variable/Atribute

Structure

Observation/
Instance

Name	Sex	Age	Semester	Height
Juan	H	19	3	1.70
Ana	M	20	5	1.65
Maria	M	18	1	1.56
Pedro	H	21	8	1.75

Structured Data

The **structure** is separated from the content. That means, each observation is an **instance** of a previously defined structure

Structured data

- Each **column** could represent a different type of **variable** (nominal, ordinal or interval)

- (Generally) **Easy** to enter, manipulate, store, read, query and process in a computer

Structured data: processes

Statistical description

- Max $\max\{x_i \ldots n\}$
- Min $\min\{x_i \ldots n\}$
- Sum $\sum_{i=1}^{n} x_i$
- Mode $\mathrm{mod}\{x_i \ldots n\}$
- Median $median\{x_i \ldots n\}$
- Mean $\frac{1}{n} * \sum_{i=1}^{n} x_i$

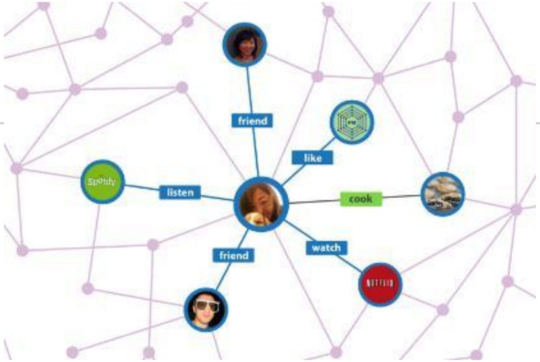
- Variance $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{\mu})^2$
- STD $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$
- Unique $unique\{x_i \ldots n\}$

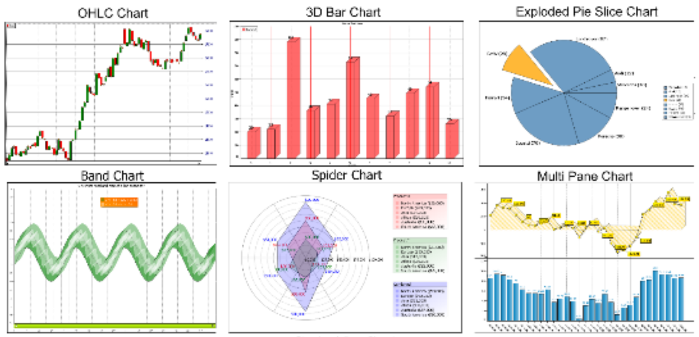
$$P(G_i) = \left(\frac{cell\ count(G_i)}{n}\right)$$

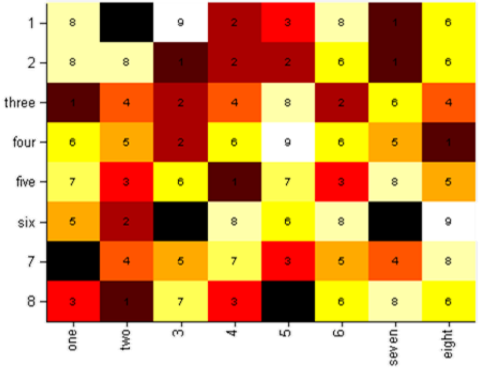
- Entropy $\sum_{i=1}^{J} P(G_i)(-\ln(P(G_i)))$
- Probability $\sum_{i=1}^{J}(P(G_i))^2$

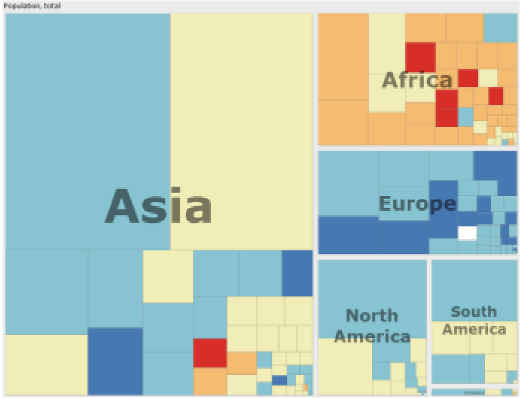
Structured data: processes

- **Plots**/charts (data visualization)





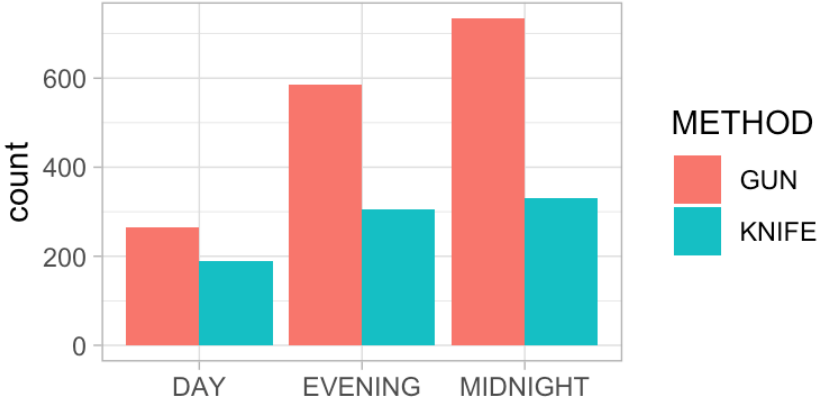




Structured data: task

- Variable associations

Crime incidents in 2017 in Washington, D.C.

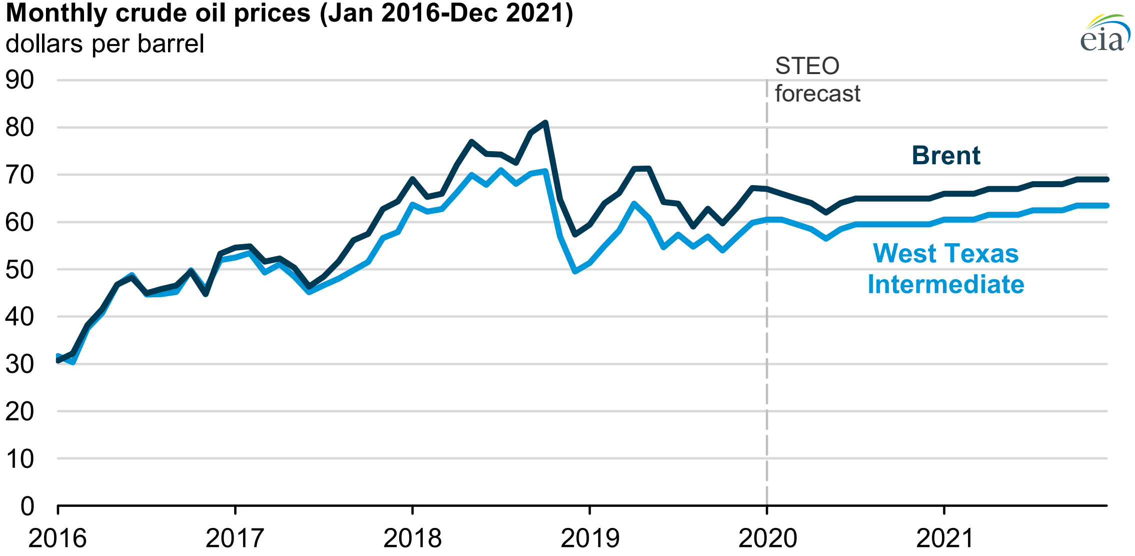


© bookdown.org

Structured data: task

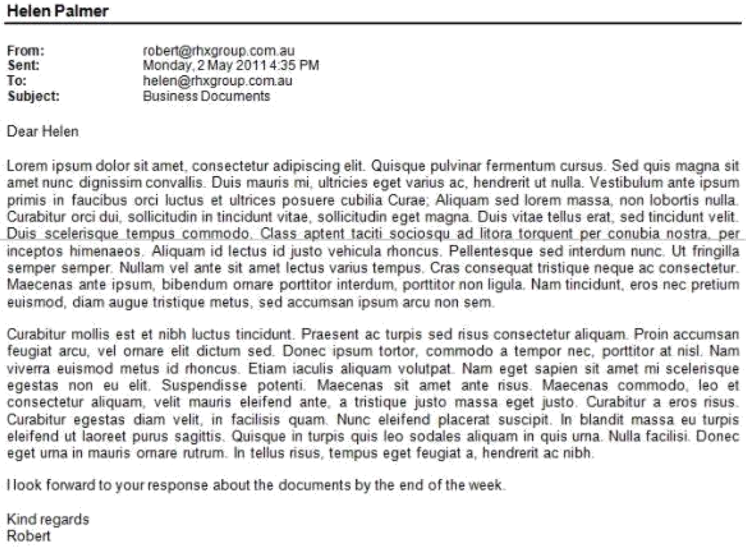
- Prediction

Monthly crude oil prices (Jan 2016-Dec 2021)
dollars per barrel



Structured data: Issues

- If there are **many variables**, there will be difficulties to process them
- **Incomplete** data
- **Noise** (spurious data)
- **Outliers**



Helen Palmer

From: robert@rhxgroup.com.au
Sent: Monday, 2 May 2011 4:35 PM
To: helen@rhxgroup.com.au
Subject: Business Documents

Dear Helen

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Quisque pulvinar fermentum cursus. Sed quis magna sit amet nunc dignissim convallis. Duis mauris mi, ultricies eget varius ac, hendrerit ut nulla. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam sed lorem massa, non lobortis nulla. Curabitur orci dui, sollicitudin in tincidunt vitae, sollicitudin eget magna. Duis vitae tellus erat, sed tincidunt velit. Duis scelerisque tempus commodo. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Aliquam id lectus id justo vehicula rhoncus. Pellentesque sed interdum nunc. Ut fringilla semper semper. Nullam vel ante sit amet lectus varius tempus. Cras consequat tristique neque ac consectetur. Maecenas ante ipsum, bibendum ornare porttitor interdum, porttitor non ligula. Nam tincidunt, eros nec pretium euismod, diam augue tristique metus, sed accumsan ipsum arcu non sem.

Curabitur mollis est et nibh luctus tincidunt. Praesent ac turpis sed risus consectetur aliquam. Proin accumsan feugiat arcu, vel ornare elit dictum sed. Donec ipsum tortor, commodo a tempor nec, porttitor at nisl. Nam viverra euismod metus id rhoncus. Etiam iaculis aliquam volutpat. Nam eget sapien sit amet mi scelerisque egestas non eu elit. Suspendisse potenti. Maecenas sit amet ante risus. Maecenas commodo, leo et consectetur aliquam, velit mauris eleifend ante, a tristique justo massa eget justo. Curabitur a eros risus. Curabitur egestas diam velit, in facilisis quam. Nunc eleifend placerat suscipit. In blandit massa eu turpis eleifend ut laoreet purus sagittis. Quisque in turpis quis leo sodales aliquam in quis urna. Nulla facilisi. Donec eget urna in mauris ornare rutrum. In tellus risus, tempus eget feugiat a, hendrerit ac nibh.

I look forward to your response about the documents by the end of the week.

Kind regards
Robert

Textual Data

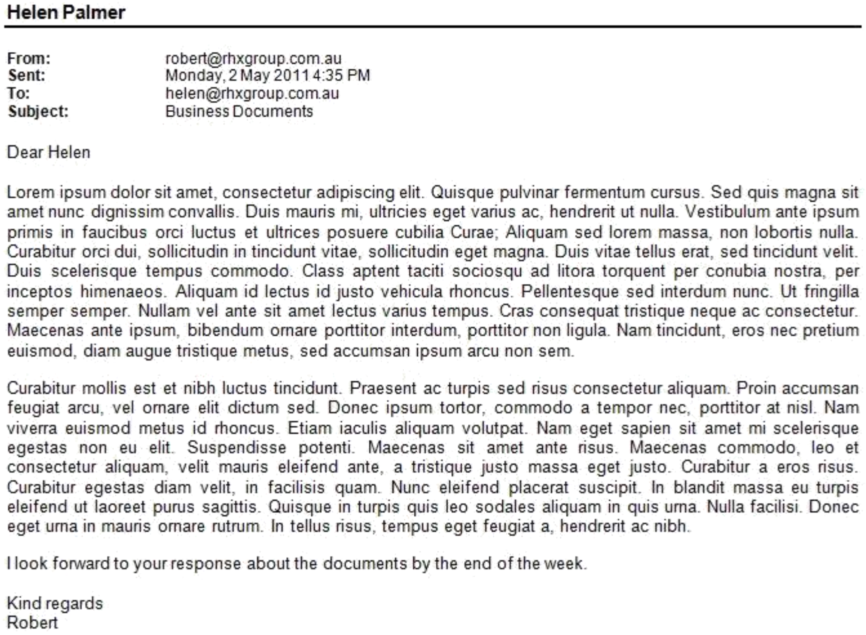
A coherent set of written **signs** that **transmits** an informative message

Textual data

- **Raw** meaning: Set of symbols (letters/logograms).
- **Semantic** meaning: Informative message content (what is the reader understanding from the text)

Textual data

- Emails
- Social networks posts
- General documents
- Web pages
- Legal documents
- Medical reports
- Research papers
- Books
- …



Helen Palmer

From: robert@rhxgroup.com.au
Sent: Monday, 2 May 2011 4:35 PM
To: helen@rhxgroup.com.au
Subject: Business Documents

Dear Helen

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Quisque pulvinar fermentum cursus. Sed quis magna sit amet nunc dignissim convallis. Duis mauris mi, ultricies eget varius ac, hendrerit ut nulla. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam sed lorem massa, non lobortis nulla. Curabitur orci dui, sollicitudin in tincidunt vitae, sollicitudin eget magna. Duis vitae tellus erat, sed tincidunt velit. Duis scelerisque tempus commodo. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Aliquam id lectus id justo vehicula rhoncus. Pellentesque sed interdum nunc. Ut fringilla semper semper. Nullam vel ante sit amet lectus varius tempus. Cras consequat tristique neque ac consectetur. Maecenas ante ipsum, bibendum ornare porttitor interdum, porttitor non ligula. Nam tincidunt, eros nec pretium euismod, diam augue tristique metus, sed accumsan ipsum arcu non sem.

Curabitur mollis est et nibh luctus tincidunt. Praesent ac turpis sed risus consectetur aliquam. Proin accumsan feugiat arcu, vel ornare elit dictum sed. Donec ipsum tortor, commodo a tempor nec, porttitor at nisl. Nam viverra euismod metus id rhoncus. Etiam iaculis aliquam volutpat. Nam eget sapien sit amet mi scelerisque egestas non eu elit. Suspendisse potenti. Maecenas sit amet ante risus. Maecenas commodo, leo et consectetur aliquam, velit mauris eleifend ante, a tristique justo massa eget justo. Curabitur a eros risus. Curabitur egestas diam velit, in facilisis quam. Nunc eleifend placerat suscipit. In blandit massa eu turpis eleifend ut laoreet purus sagittis. Quisque in turpis quis leo sodales aliquam in quis urna. Nulla facilisi. Donec eget urna in mauris ornare rutrum. In tellus risus, tempus eget feugiat a, hendrerit ac nibh.

I look forward to your response about the documents by the end of the week.

Kind regards
Robert

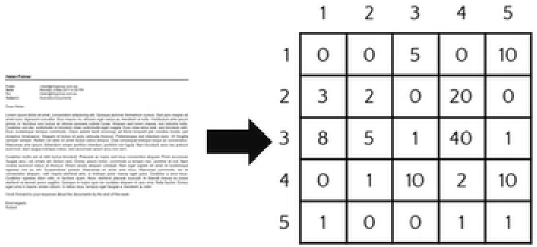
Textual data: processes

Word filtering

Frequency analysis

knowledge
education
learning
student
development
wisdom
value
proficiency
school
comprehension
motivation
faculty
awareness
notion
doctrine
science
semantics
academics
discernment
reading
theory
principles
potential
facts
consciousness
accomplishments
book
sense
brilliance
global
attitude
discipline
attention
dope
insight
power
observation
enlightenment
graduate
substance
apprehension
practice
intelligence
scholarship
coaching
vision
dogma
scoop
familiarity
genius
university
experience
schooling
concept
erudition
expertise
goods
instruction
perception
know-how
picture
ability
teacher
curriculum
tolerance

Transformation

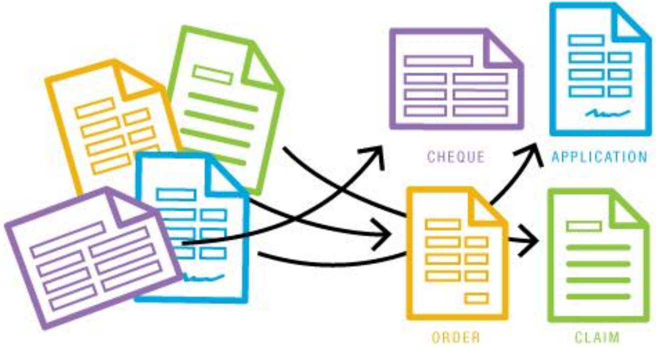


	1	2	3	4	5
1	0	0	5	0	10
2	3	2	0	20	0
3	8	5	1	40	1
4	0	1	10	2	10
5	1	0	0	1	1

Textual data: tasks

- **Classification**: Spam/phishing/legal email, news, advertises, web pages, social media posts, sentiment detection, etc.

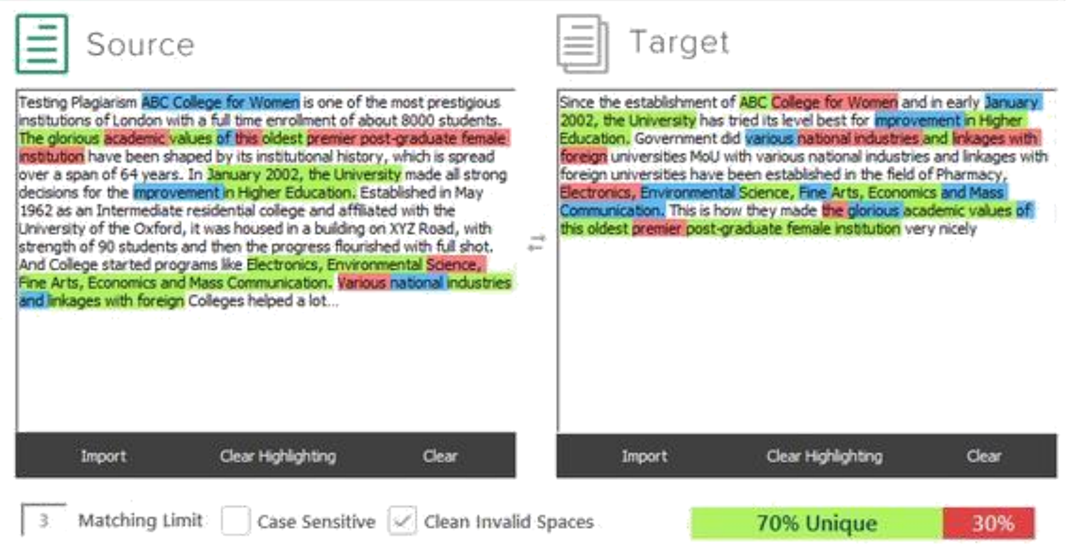
© opentext.com



'Infinity war', the last Avengers movie was great!	→	**Good**
'Infinity war' stars Robert Downey Jr as Iron Man.	→	**Neutral**
'Infinity war' was a horrible movie.	→	**Bad**

Textual data: tasks

- **Analysis**: Plagiarism detection



Source

Testing Plagiarism ABC College for Women is one of the most prestigious institutions of London with a full time enrollment of about 8000 students. The glorious academic values of this oldest premier post-graduate female institution have been shaped by its institutional history, which is spread over a span of 64 years. In January 2002, the University made all strong decisions for the improvement in Higher Education. Established in May 1962 as an Intermediate residential college and affiliated with the University of the Oxford, it was housed in a building on XYZ Road, with strength of 90 students and then the progress flourished with full shot. And College started programs like Electronics, Environmental Science, Fine Arts, Economics and Mass Communication. Various national industries and linkages with foreign Colleges helped a lot...

Target

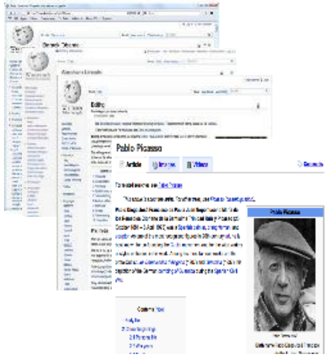
Since the establishment of ABC College for Women and in early January 2002, the University has tried its level best for improvement in Higher Education. Government did various national industries and linkages with foreign universities MoU with various national industries and linkages with foreign universities have been established in the field of Pharmacy, Electronics, Environmental Science, Fine Arts, Economics and Mass Communication. This is how they made the glorious academic values of this oldest premier post-graduate female institution very nicely

3 Matching Limit
□ Case Sensitive
☑ Clean Invalid Spaces

70% Unique 30%

Textual data: issues

- **Large** number of words



5.3 million of articles in Wikipedia (English) → **3.2 billion words!**

Textual data: issues

- **Polysemy**

All New 2018 **Jaguar** XF
The **jaguar** is an endangered species
Mac OS X 10.2 “**Jaguar**”

- **Synonymy**

The **big** dog has a **large** house
It makes me **sad** to see you
so **unhappy**

Textual data: issues

- Language

Different modalities, same observation

Brand	Model	Year	HP	Cylinders
BMW	435i	2018	248	4

Coupe and convertible variant of BMW’s compact 3-series sedan, the 435i doesn’t disappoint. It has a turbocharged engine of 248-hp 2.0-liter inline-four and can be equipped with either a six-speed manual or an eight-speed automatic and rear- or all-wheel drive.

Each modality provides with specific (sometimes complementary) details about the observation

End topic 2

Next **topics**

- Analysis of structured data
- Analysis of text data
- Unsupervised learning

Credits

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival