

Case Study: Jigsaw



CS 4460 – Intro. to Information Visualization
September 15, 2017
John Stasko

Learning Objectives



- Become familiar with investigative analysis process carried out by various types of analysts
- Learn about visualizations used to represent text and document data
- Gain familiarity with how Jigsaw system works and how to use it to explore document collections

An Illuminating Exercise



THE INTELLIGENCE REPORTS AT HAND

Here are the folders you now have that contain the intelligence reports that will be of concern to you in your analysis. You are evaluating this intelligence information during the first two weeks of December, 2004. The near future means after 16 December, 2004 which will be the date of the final reports you have.

FILE FOLDER:

1) Report Date: 28 April, 2003 From Canadian Security Intelligence Service. This report concerns a bulletin issued on 2 April, 2003 by the *Syrie*. On 12 April in Paris concerning a Moroccan named Abdallah Zinedine. Also Abdallah Zinedine CIA report for 17 April, 2003. The *Syrie* reported that Zinedine left Paris on a flight to Montreal, Canada on 1 April, 2003 and was traveling under a French passport in the name of Mehdi Rafiki. A Mehdi Rafiki did arrive in Montreal on 2 April, 2003 and said the purpose of his visit was to attend the funeral of an uncle. He telephoned Montreal contact address on 175 Rue Durocher. This address belongs to an Irish pub operated by a man named Patrick O'Malley. O'Malley says he never heard of any person named Mehdi Rafiki. But a man using a French passport in the name Abu Hafiz rented a car at the Canadian Car Rental Agency in Montreal on 2 April, 2003. This rented car was never returned to this agency. It was discovered abandoned on Walden Ave near Schiller Park in Buffalo, NY on 6 April, 2003.

2) Report Date: 14 May, 2003 From M.S. UK. Royal Yaseer, a UK citizen, was arrested on 14 May, 2003 following an accident on the M4 Motorway near the Weston Service Area outside of London. Yaseer has been an air traffic controller at Heathrow Airport for the past six years. Two kilos of Semtex were found in the trunk of his car. A videocassette of a sermon given by Omar Mahmoud Osman, formerly a Sultan (the preacher at the Baker St. Mosque in London, was found in Yaseer's apartment at # 44, Northumberland Circle, East Bedford, London. Also found in Yaseer's apartment was a note containing several addresses in Canada, the USA, and in Nassau in the Bahamas. The addresses are: 721 St. Clare St. Montreal, 455 11th Street, Miami Beach, FL, 17 12 Ferry Ave., Camden, NJ, and 11 Apple St. in Nassau, The Bahamas.

3) Report Date: 3 July, 2003 & Los Angeles Field office. In response to the CIA report (29 June, 2003) regarding Abu Sornad, 235 Buckthorn St. Inglewood CA, and Yusef Barfaba, 773 Fagon St., Culver City, CA, investigators visited these two persons at their residences. Both Sornad and Barfaba said they had never heard of persons named Robert Sornad or Omar Yusef. Each of Sornad and Barfaba was asked if they knew the other and they replied that they did not. However, it was later established that both Sornad and Barfaba appeared together in a surveillance photo taken at the California Bank in Culver City. Records of this bank indicated that Sornad and Barfaba hold a joint account at

CIA FOLDER:

1) Report Date: 12 February, 2003 From sources in Egypt. A Russian named Ignorichko was arrested in Cairo on 29 January, 2003 and charged with assault on an Egyptian police officer who had attempted to arrest him for being drunk in public. Kolobov sells medical supplies throughout the Middle East and represents a company in Moscow called Medskat. A background check on Kolobov reveals that he was formerly an administrator at the Soviet Institute Vector in St. Petersburg, near Kizny in Russia. When he was arrested, Kolobov was carrying a card with a note on it reading (in Russian): "Satygin for H. Q. Pushman".

2) Report Date: 18 February, 2003 (Transcript of names and information of interest taken from a captured computer hard drive in Afghanistan, just recently translated): (a) David Lobbajo, Canadian citizen from Toronto, Kanagi Training Camp, 2000-2001, wounded near Giza, March 2002. (b) Raed Baandall, American, Detroit MI. Mine warfare training, Afghanistan Training Camp, 2000-2001, wounded near Qalat, 26 November, 2002, tents, London, 12 December, 2002. (c) Fahid Khajaz, Canadian, Montreal, Light weapons training, Afghanistan Training Camp, 1999-2000, Sanborns 27 October, 2000.

3) Report Date: 22 February, 2003. Surveillance report on Cesar Azca, whose residence is 77 Avenue Francis, Santo Domingo, Dominican Republic. Azca, who moved from Havana, Cuba to Santo Domingo in 1992, works as a medical technician in Santo Domingo. Azca is under surveillance because of information that he is associated with Cuban intelligence services. Azca was photographed in company with a man identified as Hector, species Bogota, Colombia on 23 January, 2003. Lopez, a known representative of FARC, has conducted narcotic distribution activities throughout South and Central America and the Caribbean.

4) Report Date: 17 March, 2003. Report from a source in Calamar, Colombia. This source reports observing, near Calamar on 3 March, 2003, three men he knew were Cubans discussing the transfer of cocaine with members of the National Liberation Army (ELN) in Colombia. The source recognized one of these Cubans as Jose Escalante. This source says that he knew Escalante when they were students in 1991 at a school for medical laboratory technicians in Havana. This source said he believes that Escalante works in some capacity for the Cuban military. Our source also stated that Cuban representatives frequently visit in the border area of Colombia, Ecuador and Peru's, arranging for the exchange of weapons of various sorts needed by ELN for cocaine. That the ELN supplies to the Cubans who sell and distribute this narcotic throughout the Caribbean and Central American region.

Fall 2017

CS 4460

3

HW 3

- 50 synthetic intelligence documents
- Identify and describe the hidden threat



Homework 3: Investigative Analysis: Putting the Pieces Together

This assignment will give you experience working with text documents and conducting an investigation of the events described within them. In essence, you get to play the role of an intelligence analyst or police detective and "put the pieces together" to discover a hidden plot. In terms of our course topics, the assignment also will provide you with experience working with text and document visualizations, as well as a visual analysis exercise.

On the class 3 report site you will find a document containing a number of synthetic intelligence reports. You should examine these reports closely. Each one contains facts and events, some of which may be relevant to the hidden plot and some of which may be irrelevant. The hidden plot is a complex scenario being planned that you, as an intelligence analyst, must identify and name the appropriate message for it to be successful.

Your challenge is to make sense of the documents and "put the pieces together." The information you need to do so will be scattered throughout the hidden events and discover the plan. To assist you with this task, you will be able to use the *Splice* visual analysis system. A *Splice* display corresponding to the document collection will be available on *Splice* too. The exercise is an open task, so don't get frustrated if you struggle a little at first. Stick with it.

What has to be done in one part of your assignment. On a, you should write a brief (less than 200 words) summary of the plot of the investigation exercise that you report will occur. You should also list the top 1-4 messages. Explain what your task will happen, where, when, who is responsible, and how did they get the exercise to carry out such as success. You should write a report form if you were to present to your supervisor to work them about their insight happens. In a second section, as if you had completed the investigation process and discovered the hidden plot, you should write a report to your supervisor. What were the key challenges you faced in this exercise? What would have helped you in doing the investigation and why? How did *Splice* help or not help? What did you think about performing the kind of analysis? Additionally, if you created any auxiliary materials (text or diagrams) to help you work on the assignment, please bring those in so that we can see them.

Finally, this task involves both accurate and descriptive of the plot. We know that this is a challenging task, however, so we are mostly looking that you make a good effort at the problem. We also will reward thoughtful descriptions of your investigative process and assessing, reflective comments about it.

Fall 2017

CS 4460

4

Problem Addressed



Help “investigators” explore, analyze and understand large document collections

Articles & reports

XML documents

Spreadsheets

Blogs

Fall 2017 CS 4460 5

Example Documents



1) Report Date: 12 July, 2004. A routine customs inspection was performed on a package that was sent by a person named Pieter Doppie, 22 Hoveniersstraat, Antwerp, Belgium. This package was addressed to A. Hijazi, 1212 Lyons Ave., Newark, NJ. The customs form stated that this package contained two decorative clocks having a commercial value of \$250.00. The package instead contained 111 polished diamonds, whose value is estimated to be \$47,000. Discussions with the FBI resulted in a decision not to pursue a customs violation charge against Hijazi. The reason is that Hijazi is currently under FBI surveillance. This package was resealed and delivered to Hijazi by USPS.

2) Report Date: 22 February, 2003. Surveillance report on Cesar Arze, whose residence is 77 Avenue Francis, Santo Domingo, Dominican Republic. Arze, who moved from Havana, Cuba to Santo Domingo in 1992, works as a medical technician in Santo Domingo. Arze is under surveillance because of information that he is associated with Cuban intelligence services. Arze was photographed in company with a man identified as Hector Lopez in Bogota, Columbia on 23 January, 2003. Lopez, a known representative of FARC, has conducted narcotics distribution activities throughout South and Central America and the Caribbean.

Challenge

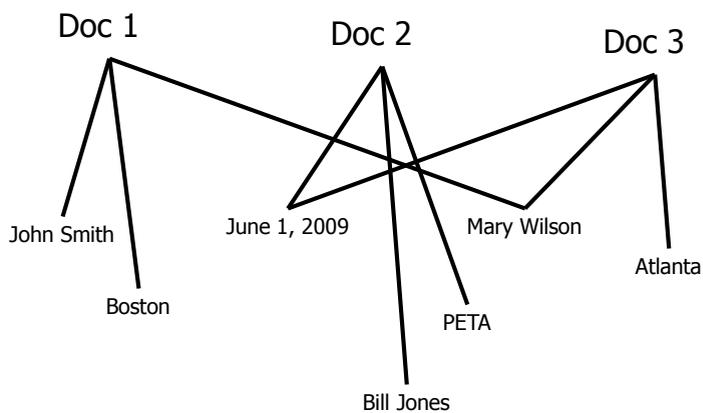


- Suppose you are an intelligence analyst and you have hundreds or thousands of such documents
- How would you explore for hidden plots and threats?

One Emphasis



- Entities within the documents
 - Person, place, organization, phone number, date, license plate, etc.
- Thesis: A story/narrative/plot/threat within the documents will involve a set of entities in coordination



Multiple Domains



Visualization for Investigative Analysis across Document Collections

- Law enforcement & intelligence community
- Fraud (finance, accounting, banking)
- Academic research
- Product reviews
- Blogs, stories
- Journalism & reporting
- Consumer research

"Putting the pieces together"



Jigsaw

Computational analysis of documents' text

Entity identification, document similarity, clustering,
summarization, sentiment

Multiple visualizations of documents, analysis results,
entities, and their connections

Views are highly coordinated



Analysis Process

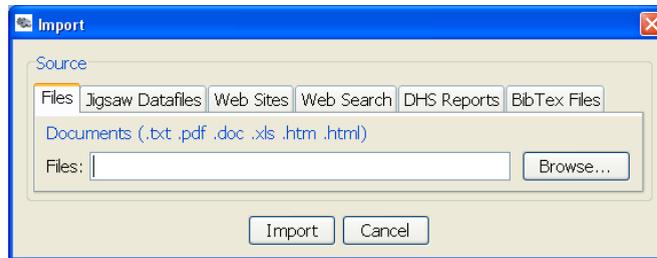


1. Document import
2. Entity identification
3. Computational text mining
4. Exploration of visualizations

1. Document Import



- Text, csv, pdf, Word, html
- Jigsaw data file format
 - Our own xml



Fall 2017

CS 4460

13

2. Entity Identification



- Must identify and extract entities from plain text documents
 - Crucial for our work
- Not our main research focus – We use tools from others
 - GATE, Lingpipe, OpenCalais, Illinois (Roth)

Fall 2017

CS 4460

14

Sample Document 1



Report: 20040510-4_16
May 14 2004

VANCOUVER, British Columbia - A Canadian immigration panel is considering whether accused environmental saboteur Tre Arrow can apply for refugee status in Canada.

Arrow, 30, who is wanted for fire bombing logging and cement trucks in Oregon, asked the Canadian authorities to remain in Canada as a political refugee at a hearing in Vancouver on Tuesday.

A key issue will be whether Arrow is affiliated with a terrorist group, which would immediately disqualify him from receiving refugee status in Canada, authorities said.

The Immigration and Refugee Board is scheduled to decide by May 31 whether Arrow is affiliated with the Earth Liberation Front, a group the FBI considers a terrorist organization responsible for scores of attacks on property over the past dozen years.

Fall 2017

CS 4460

15

Entities Identified



Source:

Date: May 14, 2004

VANCOUVER, British Columbia - A Canadian immigration panel is considering whether accused environmental saboteur Tre Arrow can apply for refugee status in Canada.

Arrow, 30, who is wanted for fire bombing logging and cement trucks in Oregon, asked the Canadian authorities to remain in Canada as a political refugee at a hearing in Vancouver on Tuesday.

A key issue will be whether Arrow is affiliated with a terrorist group, which would immediately disqualify him from receiving refugee status in Canada, authorities said.

The Immigration and Refugee Board is scheduled to decide by May 31 whether Arrow is affiliated with the Earth Liberation Front, a group the FBI considers a terrorist organization responsible for scores of attacks on property over the past dozen years.

Fall 2017

CS 4460

16

Sample Document 2



Title: Proving Columbus was Wrong
Abstract: In this work, we show the world is really flat. To do this, we build a bunch of ships. Then we...
PI: Amerigo Vespucci
Co-PI: Vasco de Gama, Ponce de Leon
Organization: Northwest Central Univ.
Amount: 123,456
Program Mgr: Ephraim Glinert
Division: IIS
ProgramElementCode: 2860

Fall 2017

CS 4460

17

Entities Already Identified



Title: Proving Columbus was Wrong **Text**
Abstract: In this work, we show the world is really flat. To do this, we build a bunch of ships. Then we...

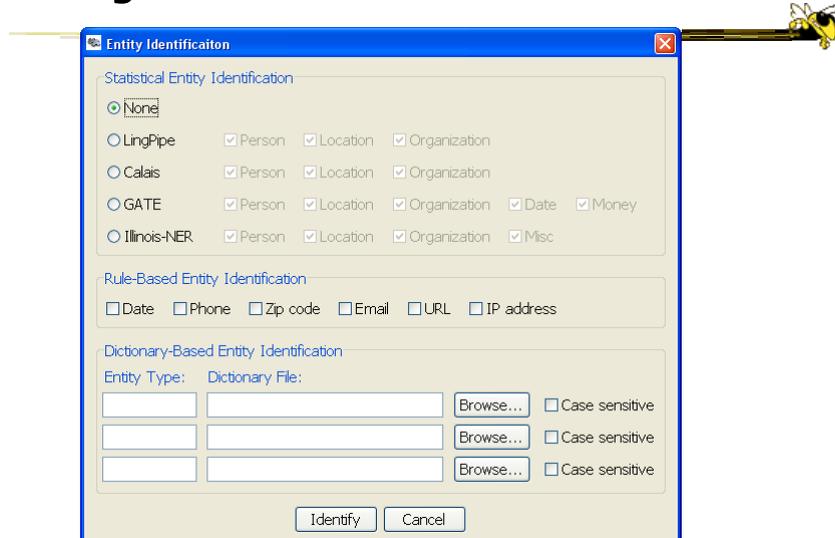
Entities
PI: Amerigo Vespucci
Co-PI: Vasco de Gama, Ponce de Leon
Organization: Northwest Central Univ.
Amount: 123,456
Program Mgr: Ephraim Glinert
Division: IIS
ProgramElementCode: 2860

Fall 2017

CS 4460

18

Entity Identification



Fall 2017

CS 4460

19

3. Computational Text Analyses

- Document summarization
- Document similarity
 - Text or entities
- Document clustering by content
 - Text or entities
- Sentiment analysis

Fall 2017

CS 4460

20

4. Visualization Exploration



Fall 2017

CS 4460

21

Entity Connections

- Core Principle
 - Entities relate/connect to each other to make a larger “story”
- Connection definition:
 - Two entities are **connected** if they appear in a document together
 - The more documents they appear in together, the stronger the connection

Fall 2017

CS 4460

22

View Coordination



- Actions (selections, expansions) in one view are propagated to others that update appropriately
- Views can be “frozen”

Fall 2017

CS 4460

23

System Views

The image displays a collection of system views from a software interface. At the top left is a small navigation pane. Below it is a large hierarchical tree view showing a complex structure of nodes and edges. To the right of the tree is a network graph with nodes and connecting lines. Below the tree is a search results window with a list of items and a table of data. At the bottom left is another hierarchical tree view. At the bottom center is a network graph with nodes colored in various colors. At the bottom right is a circular network graph with many nodes and edges.

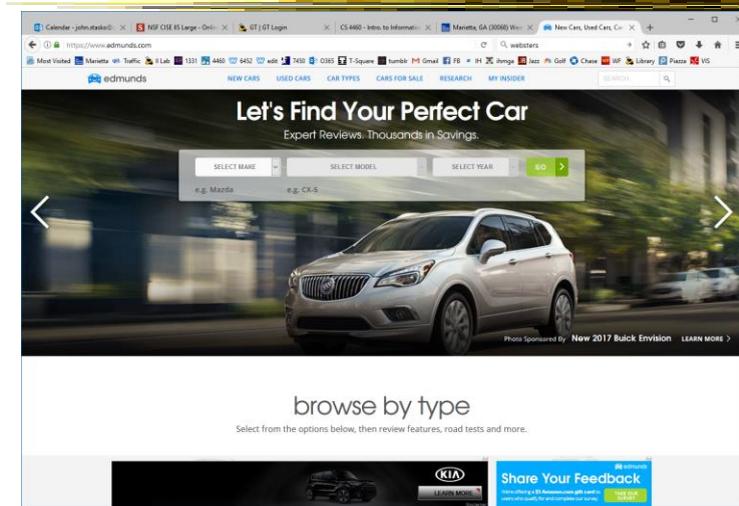
Fall 2017

CS 4460

24

Demo

www.edmunds.com



Fall 2017

CS 4460

25

Documents



231 reviews of 2009 Hyundai Genesis

Fall 2017

CS 4460

26

A Review



Rating from 1 to 10 on multiple categories:
Reliability, Overall

Free narrative text

Entities

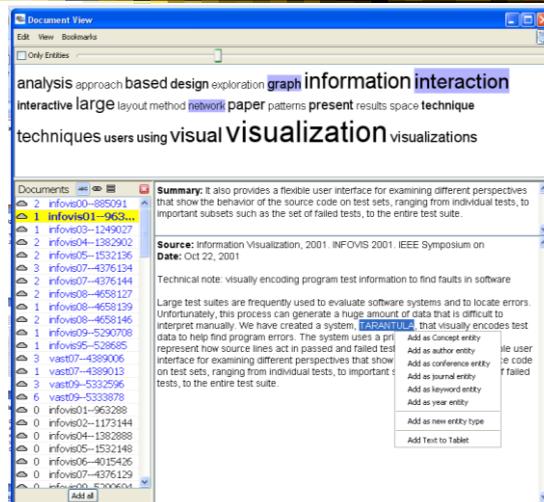
- Overall rating, value is integer score
- Car makes and models mentioned in narrative text (Ford, Audi, BMW, Corvette, Sonata, ES300)
- Car "features" mentioned in narrative text (brakes, transmission, stereo, seats)

Fall 2017

CS 4460

27

EI Correction



Fall 2017

CS 4460

28

Tablet



Fall 2017

CS 4460

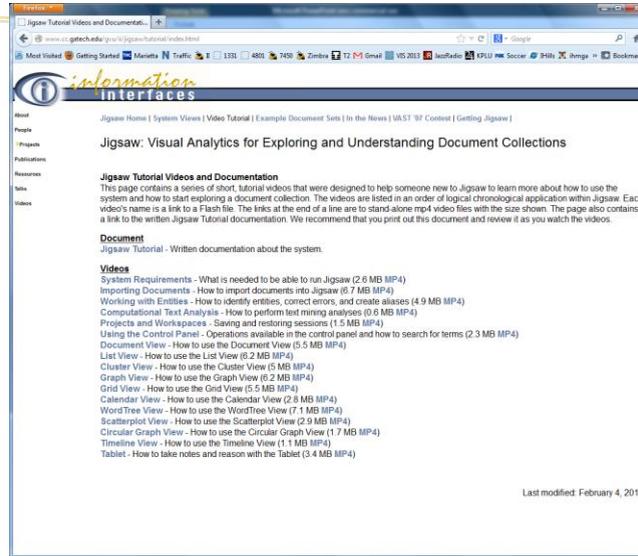
31

Website Items



- Tutorial videos
- Sample data sets
- Scenario videos
- ...

Video Tutorials

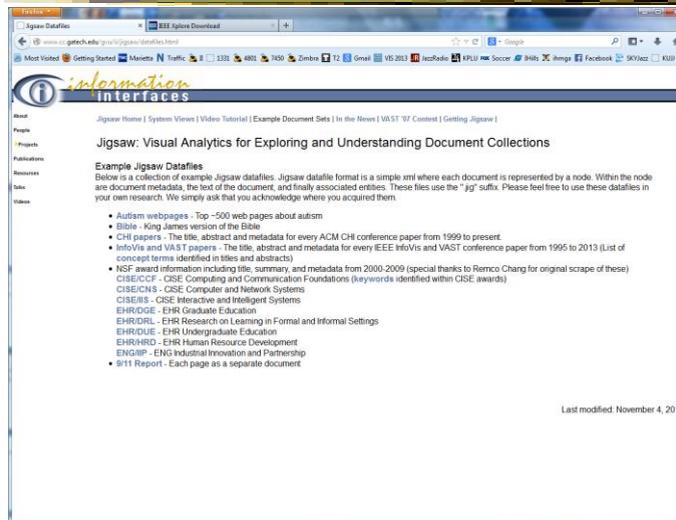


Fall 2017

CS 4460

33

Sample Data Sets



Fall 2017

CS 4460

34

Application Domains



- Intelligence & law enforcement
 - Police cases
 - Won 2007 VAST Contest
 - Stasko et al, *Information Visualization* '08
- Academic papers, PubMed
 - All InfoVis & VAST papers
 - CHI papers
 - Görg et al, KES '10
- Investigative reporting
- Fraud
 - Finance, accounting, banking
- Grants
 - NSF CISE awards from 2000
- Topics on the web (medical condition)
 - Autism
- Consumer reviews
 - Amazon product reviews, edmunds.com, wine reviews
 - Görg et al, HCIR '10
- Business Intelligence
 - Patents, press releases, corporate agreements, ...
- Emails
 - White House logs
- Software
 - Source code repositories
 - Ruan et al, SoftVis '10

Fall 2017

CS 4460

35

Learning Objectives



- Become familiar with investigative analysis process carried out by various types of analysts
- Learn about visualizations used to represent text and document data
- Gain familiarity with how Jigsaw system works and how to use it to explore document collections

Fall 2017

CS 4460

36

HW 3



- 50 synthetic intelligence documents
 - In T-square
 - As one big document, as 50 documents
 - As a Jigsaw data file
- Identify and describe the hidden threat
 - Write a 3-4 sentence hypothesis
- Write a few paragraphs about your analysis process
- Due next Friday -Don't wait til last minute

Fall 2017

CS 4460

37

Upcoming



- Multivariate visual representations 1
 - Prep: Watch system demo videos
- Multivariate visual representations 2

Fall 2017

CS 4460

38