Informatics 1B
Data and Analysis: Tutorial 6
Introduction to Unstructured Data

Manuel Marques Pita and Helen Pain
Edited by Gaya Nadarajan

For Week 11 beginning March 19, 2007

1 Introduction

In the first section of this course you learnt how to work with data which have a clear structure. In many domains it is natural to think about *entities, attributes* and *relationships*. Methods for querying highly structured data are very useful for a vast number of applications. At present, thousands of internet based applications rely on databases. IMDB, Easyjet, Ryan-air, BBC news and Tesco are just a few examples.

During the second part of the course you were introduced to methods which are useful for imposing (and utilising) predetermined semi-structural patterns on data to be exploited, but which cannot be naturally framed into a fully predefined structure. These methods are extremely useful for another vast domain of applications such as text summarisation, cataloging and querying dynamically generated data (such as data from human dialogues) and Bioinformatics.

In the context of working with data, still one general question remains unanswered: What happens when you have data which you cannot frame in a set structure, or when you do not even know what the *features* of this data are? Or even when the data is not there?

A good way of start thinking about this is to imagine that you are an explorer, a pioneer going into a distant, unknown forest. The *structured* forest will be one that is, in many ways, known by you, you will be expecting to find trees of certain types of species. Each of them will have a number of salient, identifying attributes you will record in your notebook, you will count the trees you found and record their corresponding characteristics in tables that are easy to read and implement in a database.

The *semi-structured* forest will be, for example, one which you can *see* through the writings of an author (Mr. X) who writes books about the forests that exist in the area of Scotland where he grew up. You will need to impose some structure on that text annotating things such as *references to*

named trees, size, medicinal uses, interactions with other species, and maybe even *cultural references* and *poetic licenses* that might exist describing a tree's relationship to the humans in its environment.

When in the *unstructured* forest, you could be surrounded by people from the Cassini-Huygens mission to Saturn who are trying to make sense of the data sent from one of its moons, Titan, in order to understand the dynamics of that distant place which seemingly resembles the Earth in many ways. You could also be surrounded by Historians who have just found all the records kept by the Guipuzcoana Company (remember this from the first tutorial?) Trying to find out what we can learn from these documents. You wont know what structure is in the data, and even sometimes the data will not be available, you will need to collect it.

The real world which we deal with everyday is the best example of unstructured data. We are constantly bombarded by information which we use to make sense of the world and to survive in it. When determining whether or not global warming is happening, for example, scientists don't use solely a database, or some annotation scheme in some data they gathered (although they might use data in these two levels too). They will be mostly deciding what to measure, where, when, for how long and at what intervals. Then they will use all the collected data to decide if there are correlations between the variables they have decided to study.

Back to our less extreme example, the Guipuzcoana, the analysis of the data found can help us to find answers about historical facts and the evolution of countries and societies previously unknown to us.

The thing with unstructured data is that we don't know yet what we are looking for, exactly at the outset...

2 What happened when the Guipuzcoana books and records were found?

You have been hired now as a consultant working for the historiographical division of the Guggenheim Museum in Bilbao, Spain. This museum is now the keeper of all the newly found documents belonging to the defunct Guipuzcoana Company. There are, at least, a hundred boxes containing documents of all sorts in a special room where you will be working during the next 45 minutes or so.

The Guggenheim's representative, Dr. Echeverri, tells you that they need to use this data to inform the design of a multimedia tool to be used in the museum. The users will be people interested in learning more about different aspects of this colonial company.

Dr. Echeverri starts by saying that there is a large number of documents describing in a lot of detail most of the **ships** owned by the company, their type, all measures, history, all is recorded and available. There are also drawings and construction planes for almost all the vessels!

There is also a lot of information about each of the **journeys** the Guipuzcoana made for trade

purposes over almost 40 years. Journeys are documented in detail with logs kept every six hours specifying environmental conditions, route adjustments, estimated position, etc. Dr. Echeverri tells you that there have been doubts about the accuracy of certain logs though. Historians think that some captains might have "altered" the logs to hide the fact that they stopped in other places on their way to or from South America. According to Dr. Echeverri, some of the documented journeys were seemingly impossible to make given the naval technology of the time and adverse environmental conditions in the North Atlantic ocean. She mentions that this will be an area some specific users will be very interested in.

Dr. Echeverri, going a little bit ahead of her duty as always, has also contacted people who can provide you with the average environmental conditions of the North Atlantic ocean during the period the company was in operation. This information basically includes average values for winds and currents (direction and force) vectors in each 20-metre square area in the whole North Atlantic. She thinks you could maybe use this information to build a simulation if that was decided to be appropriate.

Dr. Echeverri also explains that a large collection of **maps** of the two main destination ports were found too. These two ports (located in the coast of Venezuela) are called *La Guaira* and *Puerto Cabello*. Cartographers hired by the Guipuzcoana kept documenting year after year all the changes these two places went through as a direct effect of the company's operation. They documented urbanistic development, production, civil confrontations, population numbers and population distribution according to race.

Again, going ahead of her duty, Dr. Echeverri says there is a bunch of computer applications available to generate reconstructions of geographical spaces. The basic idea is that you provide the system with a map of the area to be modelled. This map can only have areas ranging from white to black. The lighter the area, the higher the terrain will be elevated and vice-versa. Once the surface has been modelled, it is possible to add reconstructions of houses or any other item the designer decides to add. This is, just in case you are interested in using them.

The company also kept a detailed record of **taxes** it charged people living in these two (richest) areas of the country.

What Dr. Echeverri expects you to do is to come up with a *concrete* list of requirements that could be potentially made by users of the multimedia system (to be developed) which can be satisfied by the data available. In other words, what concrete, educationally valuable features can this system have provided the available data (and other tools)?

3 Your Tasks

3.1 Data Requirements

In this session, focus on two types of user that will be interacting with the multimedia application. One of them will be a *historian* and another one a *school pupil*. Try to come up with two sets of

questions these two types of user could ask the system (devise at least five questions for each of the two types).

Once you have determined the questions, concentrate on one for each set, and specify what would need to be *represented* in the system (and how) in order for it to be able to provide the correct answers. Is there any question you formulated previously which would be hard for the system to answer? If so, why is that the case? What would the system need to do to get an answer?

3.2 Functional Requirements

For this task, imagine that you have to design the user interface for the multimedia system to enable users (such as the historian and the school child) to get access to the information in the application. Write a *story-board* that shows a short interaction between an user and the system. Indicate what *options* are available in the interface for the user, what *actions* the user performs and what *responses* the system would make. (Do this for at least two questions in your initial sets)

3.3 User Requirements

How would the user interface differ for the two types of user we have considered? Think for example that when displaying texts, you can make assumptions about what the historian knows, or make the text very accessible for the school child. What other differences what you determine to be relevant?
[Hint: Using good user interface design principles as your basis, see how these should be utilised in different ways for your users]