Informatics 1B
Data and Analysis: Tutorial 3
Introduction to Semi-structured Data

Manuel Marques Pita and Frank Keller

February 7, 2007

1 Introduction

Have a look at the following text. As its title suggests, this is a list of plot summaries for different films. Only the summaries for two films are included here (from our original file containing a few hundred thousand of them).

```
PLOT SUMMARIES LIST
===================

MV: "$1,000,000 Chance of a Lifetime" (1986)

PL: A short-lived quiz show hosted by TV veteran Jim Lange. Husband-and-wife
PL: teams would compete against each other in answering a tough series of
PL: trivia questions. The winning couple would then go on to the championship
PL: round, where they would have a chance to win one million dollars.

BY: Jean-Marc Rocher <rocher@tss.yh.nec.co.jp>

PL: Two married couples, sometimes including a returning champion, competed in
PL: one of the first game shows to offer $1 million as a grand prize. In
PL: "$1,000,000 Chance of a Lifetime," which combined elements of "Wheel of
PL: Fortune" and "Scrabble," the couples competed to solve word puzzles. Up to
PL: five clues were given for the puzzle's solution; one member of each team
PL: was selected to try to guess each clue word, with letters inserted at
PL: random except for the last one. Each correct guess was worth $25 and
PL: allowed the contestant to go to the keyboard. The keyboard indicated which
PL: letters were in the puzzle, plus one "stinger" (an extra letter not in the
PL: puzzle); selecting the "stinger" meant the contestant was eliminated from
PL: further play in that round and the spouse took his/her place. Each selected
```

PL: letter revealed in the puzzle added $25 to the pot; the team that solves
PL: the puzzle won the pot. Three rounds were played, with the second round
PL: values of $50 per clue/puzzle letter and third round values of $100 each.
PL: The team with the most money after three rounds was champion, kept their
PL: winnings and advanced to the bonus round. In the bonus round, the couple
PL: selected a category from a choice of three and then are locked in an
PL: isolation booth. The couple has 60 seconds to guess six words associated
PL: with that category. For each successful bonus round win, the couple was
PL: given a cash prize ($5,000 on day 1, $10,000 on day 2), or they could give
PL: it back and continue, always at risk of losing a subsequent bonus round and
PL: thus having to leave the show. Only by winning three bonus rounds in a row
PL: did the couple win the $1 million grand prize. The $1 million was awarded
PL: as all cash during the spring 1986 season, and $900,000 plus a $100,000
PL: luxury prize package (of a car, furs, jewelry, a grand piano, furniture and
PL: a trip of up to five stops anywhere in the world) during the 1986-1987
PL: season.

BY: Brian Rathjen <briguy_52732@yahoo.com>

MV: Un long dimanche de fiancailles (2004)

PL: Five desperate men shoot themselves in order to be relieved from the
PL: horrifying frontline at the Somme, in WWI. A court-martial decides to
PL: punish them by leaving them alone in no-man's land, to be killed in the
PL: crossfire. Then all hell breaks loose and they all die. Or not? One of
PL: these men's fiancee, a young girl who can't walk since age 3, receives
PL: information that makes her suspect his boyfriend might have gotten away
PL: alive. So she embarks in a painful, long and often frustrating ordeal to
PL: find out the truth.

BY: Erwin van Moll <max404@hotmail.com>

PL: From the director and star of "Amelie" (Jean-Pierre Jeunet and Audrey
PL: Tautou) comes a very different love story, "A Very Long Engagement," based
PL: on the acclaimed novel by Sebastien Japrisot. The film is set in France
PL: near the end of World War I in the deadly trenches of the Somme, in the
PL: gilded Parisian halls of power, and in the modest home of an indomitable
PL: provincial girl. It tells the story of this young woman's relentless,
PL: moving and sometimes comic search for her fiancee, who has disappeared. He
PL: is one of five French soldiers believed to have been court-martialed under
PL: mysterious circumstances and pushed out of an allied trench into an
PL: almost-certain death in no-man's land. What follows is an investigation
PL: into the arbitrary nature of secrecy, the absurdity of war, and the
PL: enduring passion, intuition and tenacity of the human heart.

BY: bondish

2 Levels of description: From structured to semi-structured data

(1) Is it possible to include the information contained in the plot summaries above in a relational database?

(2) What could be the structure of the data required to do this?

(3) What kind of *knowledge* can we extract from this data at this structural level?

(4) Why would we ever want to read one of these summaries?

(5) If we had a few Gigabytes of plot summaries, and we were looking for films with a *certain kind of plot* surely we would not want to read through the whole file. Our highly structured databases, even if so useful for so many tasks, would start having problems at this level of description. Why do you think this is actually the case?

Look at the provided text again and focus on how each line starts. Notice that every single line is *identified* with the kind of information it contains `MV:` identifies a line containing a movie title, `PL:` identifies a line in the plot-summary and `BY:` identifies the author of the summary. We are interested in knowing more about the actual plots, which up to this point are represented just as text (which, as we have just seen, could be an attribute of some entity). In the previous chapter, everything we wanted to know was represented as *entities, attributes* and *relationships*. Now we go to another level, and our focus is on the *information* provided by the text in the plot-summaries.

The text in these plot-summaries is just a sequence of words. It would be quite hard and it would not make much sense to put these words in a relational database: finding entities, attributes and relationships would very hard and in many cases impossible. However, the text can be enriched with some structure which would be defined in the context of the application in which we want to use this data. As you will see in the lectures, there are many different kinds of applications we can mention here: recording DNA sequencing of some species, building dictionary for a newly found language and identifying changes in the human body exposed to some new medicine are just some examples of them.

In the kinds of applications just mentioned, we normally start with a large amount of data which is very often available in a *textual* form – which will be referred to as *Corpus,* and the task is to *label* portions of the text in such a way that some form of structure, relevant to the application being developed, is imposed on the available data. In the case of a dictionary, we would be interested in identifying elements such *verbs, nouns, adjectives* and so on. In the DNA data we would want to identity things such as *bases* and *amino acids,* for example. The process of assigning these labels to the available text is called *annotation* as you will see in the lecture.

(6) What kind of labels would be relevant for the plot-summary problem?

Once we have annotated a text with the *labels* we have chosen we can start looking at some statistical properties of this corpus. The two main numbers we can calculate at this point would be the *token count* and the *type count*. The first will tell us how many tokens (words, punctuation

marks, etc.) there are in the whole corpus. The type count will tell us how many *different* tokens there are in the corpus. For instance, the word war may appear 10.000 times in the corpus and this number would be added to the token count, but only once in the type count.

Have a look now at the first plot summary in the list by Brian Rathjen.

(7) Can you see any words that appear in the text often? Calculate how many times the following words appear in that summary

1. couple
2. letter
3. puzzle
4. prize
5. round

(8) Do these words by themselves give you an idea of what the plot says? Can you think of any other strategy to extract parts of the text that are useful to get an idea of the main contents of the plot?

(9) How can we calculate the total number of types (type-count) for a given corpus?