Telltale User Documentation

Grace Crowder
crowder@cs.umbc.edu

October 2, 1995

1 Introduction to Telltale

Telltale is a dynamic hypertext environment project that uses n-grams [Sha48] to build hypertext links between documents. Telltale allows users to view documents from a corpus, creating links to documents that are similar [Dam95].

This paper describes a Telltale system based on [PN95].

2 Screen Layout

The telltale screen layout contains several different areas [see Figure 1]. The following sections describe the areas of the screen.

2.1 Menu Bar

The menu bar appears at the top of the Telltale window. It contains the file menu.

2.1.1 File Menu

There are several commands in the File Menu:

- load database

  Allows the user to specify an existing telltale database to be loaded.

- save database

  Allows the user to pecify where to store the current telltale database.

- document tags

  Displays and allows changes to the text tags expected in the incoming text.

- quit

  Quit telltale.

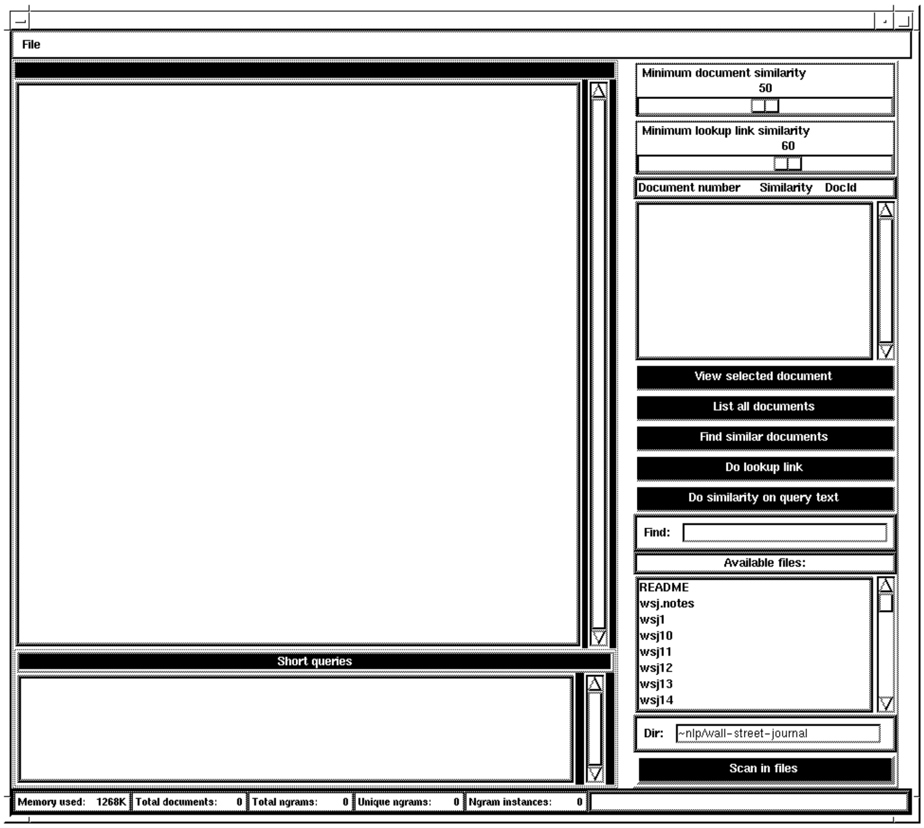

Figure 1: Telltale Screen Layout

2.2 Document Window

The document window displays the current document. When Telltale is first loaded, there is no current document. See the blank document window in figure 1.

To display a document in this window, select the document from the document list and chose the *View selected document* command.

2.3 Query Window

The query window displays the current query. Queries may be typed in or selected from a displayed document and pasted into this window. See figure 2.

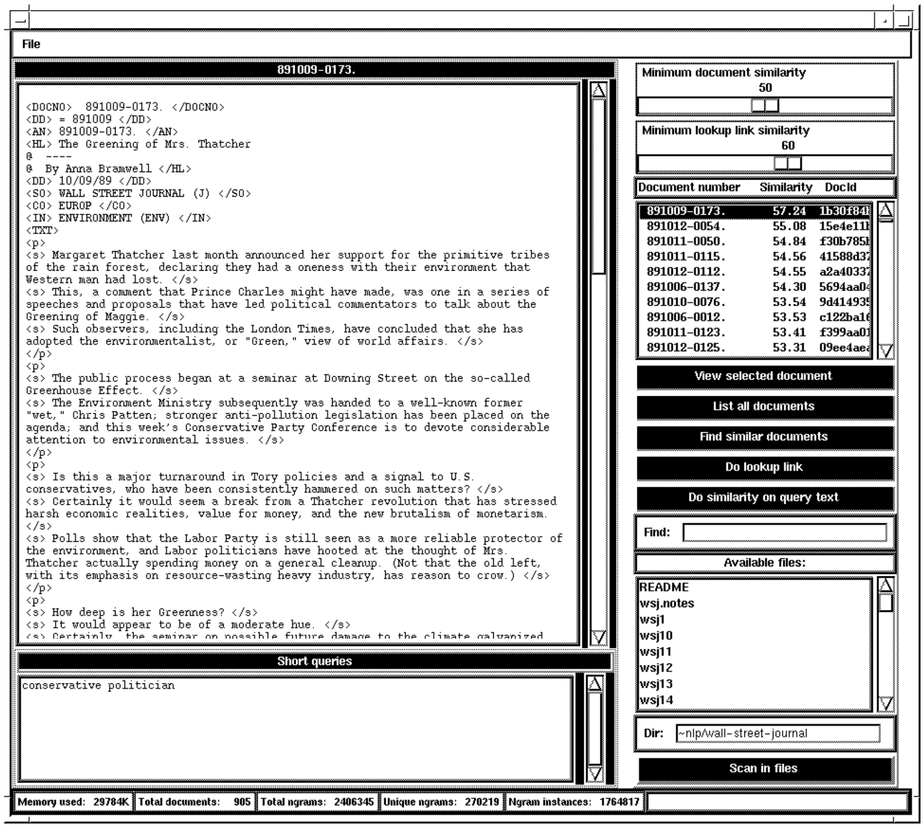

Figure 2: Query text used for Do similarity on query text

2.4 Document Similarity slider

The document similarity slider is used to specify the minimum document similarity that is required before a document is considered as an answer to a query or as being similar enough to the target document. This value is used by the *Find similar documents* and *Do similarity on query text* commands.

2.5 Lookup Link slider

This has not yet been implemented.

2.6 Document List

This window contains a list of currently available documents. When Telltale is initially loaded, this list is empty. As documents are scanned in, they are added to the document list which will be displayed in this window when the *List All Documents* command is executed. When there is a set of documents in response to a similarity request, the documents are listed in this window by similarity score.

2.7 Commands

There are six Telltale commands:

- View selected document

  This command causes the currently selected document from the document list to be displayed in the document window.

- List all documents

  This command causes all currently loaded documents to be listed in the document list.

- Find similar documents

  This command finds documents similar to the document currently displayed in the document window. The result will be a list of similar documents in the document list window. These documents will be in descending order by similarity score. Not suprisingly, the original document will be at the top of the list as it is 100% similar to itself.

- Do lookup link

  This has not yet been implemented.

- Do similarity on query text

  This command finds documents similar to the query currently displayed in the query window. The result will be a list of similar documents in the document list window. These documents will be in descending order by similarity score.

- Scan in files

  This command scans in the files highlighted in the file list box which contains a list of text files in the currently specified directory.

2.8 Status Bar

The Status bar contains useful information about the resource usage of Telltale. From left to right, the status bar contains:

- Memory used

  Tells how much memory is being used by telltale.

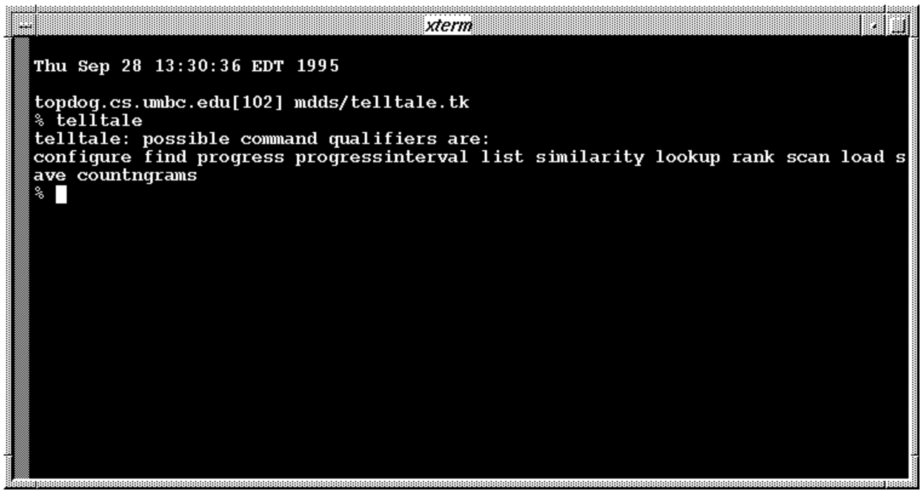

Figure 3: Telltale command line

- Total documents

  Tells how many documents have been scanned in to Telltale.

- Total n-grams

  Tells how many n-grams exist in the scanned in text.

- Unique n-grams

  Tells how many unique n-grams exist in the scanned in text.

- N-gram instances

  Contains the sum of the unique n-gram counts for each of the scanned in documents.

- % of task complete

  Tells how much of the current task is complete.

3 Telltale commands

All telltale commands are available when you run telltale.tk. This program opens up a command line in the window from which it's started (see figure 3). These commands can be typed on that command line (of course, the results may not be immediately reflected in the GUI). Telltale.tcl supports only the text-based commands listed in this section.

- configure

  This command is used to set many configuration parameters:

  Note: that some of the configure parameters are specified by appropriate entries in telltalerc.

  - –docstart “$< doc >$”

    Specifies a text flag.
  - –docend “$< /doc >$”

    Specifies a text flag.
  - –textstart “$< txt >$”

    Specifies a text flag.
  - –textend “$< /txt >$”

    Specifies a text flag.
  - –docnostart “$< docno >$”

    Specifies a text flag.
  - –docnoend “$< /docno >$”

    Specifies a text flag.
  - –n-gram 5

    Specifies the size of the n-grams for this run of telltale. Note that it is a bad idea to change n in the middle of a telltale session.

    Think carefully before changing this value. Currently n can range from 1 to 8.
  - –addinstance 1

    Specifies that n-grams in scanned in documents are to be added to the n-gram data structures. If -addinstance is 0, then n-gram instances will remain zero even though the documents read in contain n-grams
  - –affectweights 1

    Specifies that n-grams in scanned in documents are to affect similarity scoring.
  - –docreports 0
  - –usestopchars 0

    A 0 for this parameter specifies that stop characters (such as punctuation) are to be treated as blank characters in scanned documents. A 1 for this parameter specifies that stop characters (such as punctuation) are to be replaced by blank stop-character blank in scanned documents. In addition, when n-grams are computed, any n-grams containing stop characters will be discarded. This prevents n-grams from spanning sentences. It also removes most text markup tags from the n-gram lists.

  - –rankmethod 1
  - –rankamt1 0.000000
  - –rankamt2 0.000000
  - –scorefilescale 1e+09

- find

  Find gets information about a document (the containing file, offset within the file, etc.)

- progress and progressinterval

  Progress and progress interval are used to implement status messages. There's no need to use them in the tcl interface.

- list

  Displays a list of scanned documents.

- similarity

  Returns a list of documents that are similar to the given query.

  Usage: telltale similarity $< docId|string >$ [minsim]

- lookup

  Not yet implemented.

  Usage: telltale lookup $< docId|string >$ [minsim]

- rank

  Not yet implemented.

  Usage: telltale rank $< docId|string >$ [minsim]

- scan

  Specifies a file to be scanned.

  Usage: telltale scan $< file >$

- load

  Specifies an already existing telltale database to be loaded.

  Usage: telltale load database_name

- save

  Specifies a name for a telltale database to be stored.

  Usage: telltale store database_name

- countngrams

  Countngrams counts the number of n-gram postings and unique n-grams given a database and a particular filtering function. The default is to include all postings and n-grams, but this can be adjusted to experiment with different filtering mechanisms.

References

[Coh95] Jonathan Cohen. Highlights: Language- and Domain-Independent Automatic Indexing Terms for Abstracting. *JASIS*, April 1995.

[Dam95] Marc Damashek. Gauging Similarity with n-Grams: Language-Independent Categorization of Text. *Science*, Vol. 267, pp. 843-848, 10 February 1995.

[PN95] Claudia Pearce and Charles Nicholas. TELLTALE: Experiments in a Dynamic Hypertext Environment for Degraded and Multilingual Data. To appear in JASIS, 1995.

[Sha48] Claude Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, Vol 27, pp. 379-423, 623-656, July, October 1948.