APPROACHING DESIGN AND DEVELOPMENT OF
A DENTAL INFORMATICS WEB PORTAL

by

Ella Epshteyn, RDH, BS

A CAPSTONE PROJECT

Presented to the Department of Medical Informatics and Outcomes Research and Oregon Health & Science University

School of Medicine

in partial fulfillment of the requirements for the degree of

Master of Medical Informatics

May 2003

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I thank my advisor, Susan Price, MD, for her generous guidance and support throughout this project. Her insight and enthusiasm were invaluable. I am also grateful to Dr. Judith Logan, Dr. William Hersh and Dr. Denice Stewart. This project could not have happened without their knowledge and expertise. In addition, I thank all members of the American Medical Informatics Association Dental Informatics Working Group and staff members of Oregon Health & Science University's Dental Informatics Department who offered their help and feedback at any point throughout the project. And finally, special thanks to my husband Eugene for supporting my dreams and ambitions.

ABSTRACT

In recent years, an enormous amount of information has appeared on the Internet including many pages related to Dental Informatics. However, information on any one particular topic is usually sparse and difficult to find. One solution to this problem is to create a Web portal unique to a particular domain such as Dental Informatics. This project started development of a methodology for creating small domain Web portals. The domain of Dental Informatics was used as an example.

Several phases of the project have been completed including selection of the Content Management System, design of the preliminary hierarchy of categories and partial development of an automatic classifier prototype. While the initial version of the classifier is promising, some additional refinement is needed. This approach to building small domain portals seems feasible. However more research is needed before final conclusions can be drawn about its effectiveness.

Introduction

Computerization is changing dentistry. New electronic tools are emerging to assist with the diagnosis and treatment of oral diseases and injuries. The majority of dental offices now use computers for practice management including but not limited to tasks such as billing and scheduling. Electronic dental records, while not widespread, have a potential to revolutionize dental practice by allowing faster, easier access (as well as remote access and information exchange), and offering more effective record keeping including electronic restorative and periodontal charting, decision support and other features. Intraoral cameras are becoming an integral part of patient education and remote consultations. Digital radiography eliminates the necessity of darkroom and chemicals, decreases processing time, solves storage problems, provides easier retrieval, and, most importantly of all, significantly reduces the amount of radiation exposure to a patient. New application software packages have been developed to allow digital images to be stored, transmitted, reproduced, or manipulated to enhance diagnostic quality. The Internet is proving to be an extremely useful source of diagnostic and therapeutic information and continuing education.[1, 2, 3, 4, 5, 6]

Despite the tremendous growth of Dental Informatics as a field in the recent years (defined by Schleyer et. al. as “the application of computer and information science to improve dental practice, research, education and management,”) many dentists are still unfamiliar with the term[7]. Yet many dental practices are going digital, with new modalities such as electronic billing and digital radiography no longer a luxury but soon to be a necessity.[1, 2] When considering a change, such as switching to electronic dental records, the Internet may be the best place to look for information on the existing systems and research.

In recent years, an enormous amount of information has appeared on the Internet including many pages related to Dental Informatics. This is good news for dental professionals looking for information on anything from teledentistry and electronic dental records to on-line continuing dental education. However, information on a particular topic (such as continuing education courses concerned with early caries detection, for example) is usually sparse and difficult to find.[3] Web catalogues (also called directories) such as Yahoo (http://www.yahoo.com) and Looksmart (http://www.looksmart.com/) attempt to make searching for information easier by organizing it into categories.[8] Considering the amount of information available, however, it is impossible for such a wide spectrum system to cover every single domain in the detail that is often necessary for effective searching. Thus, small domains such as Dental Informatics are either left out or lack depth. One solution to this problem is to create a Web portal unique to a particular domain. Such a portal would cover most major topics in the field organized for easy navigation, perhaps as a hierarchy of categories. The idea is not new (there are quite a few portals on the Internet devoted to small domains). Yet, there is no clear cut methodology to approach the development of a portal in a relatively new field everyone has a slightly different understanding of, such as Dental Informatics. How are categories developed? How can we assure that all major topics are covered as completely as possible? How can we make maintenance of the portal as easy and inexpensive as possible? These are just a few of many questions that one could encounter when designing such a portal. This project will attempt to develop a methodology for development of small domain Web portals, as well as to answer some of these and other questions as a cornerstone for future work.

Background

About Portals

The Concise Oxford Dictionary (9th ed.) defines the word "portal" as "a doorway or gate, etc., esp. a large one and elaborate one"[9]. While there is no standard definition in regards to a Web portal, it appears that the main objective of most Web portals is to provide a single access point to information[10]. In addition, many portals offer a variety of services to users such as e-mail, chat rooms, auctions, and calendar services among others.

Portals vary in purpose. They can generally be broken down into three groups: commercial (personal), enterprise (corporate) and knowledge. Commercial portals contain "personalized, but generic content" and are available for free on the Web. Any Web search site such as My Yahoo (http://my.yahoo.com) or My Excite (http://my.excite.com) may be considered a commercial portal. Enterprise portals contain "personalized, generic and business-related content". The major goal of such a portal is to share business-related information such as corporate data. WebGov (http://w3.gsa.gov/webgov/webgov.nsf), a government portal with links to various federal websites, is an example of an enterprise portal. Knowledge portals offer personalization based on user roles, habits or preferences, as well as generic, business-related, and domain-specific content. The Department of Defense Network (http://www.nic.mil/), which provides information to the US military worldwide, is an example of a knowledge portal.[11]

Portals can also be subdivided into horizontal portals (also called Horizontal Enterprise Portals (HEP) and megaportals), and vertical portals (Vertical Enterprise Portals (VEP)). HEPs are public Web sites that attempt to provide their users "with all the services they might need." These types of portals usually include shopping, weather, stock prices, news, and other services. An example of such a portal is MyExcite. HEPs can typically be personalized to individual needs of users (an example of such personalization is creating a stock portfolio). [12]

Some users, however, may need a different type of personalization, specifically the type that is related to organization-specific information. VEPs deliver "organization-specific" user-oriented information. Authentication is usually required to access VEPs and the entry pages are unique for each user. An example of VEP may be a college portal offering students access to their grades and schedules, while allowing teachers to post course materials and access and maintain student information (such as grades and assignment feedback). [12]

Web portals can deal with multiple domains or be domain-specific. Domain is generally defined as "an area under one rule, a realm". Dental Informatics is a domain. The field can be considered a specialty of Medical Informatics, which, according to Van Bemmel, "comprises the theoretical and practical aspects of information processing and communication, based on knowledge and experience derived from processes in medicine and healthcare". Schleyer et al. identify several goals of Dental Informatics including improving patient outcomes, improving the efficiency of dental care delivery, and supporting research and education. [7, 13]

When approaching a project as big and complex as designing a portal, certain questions have to be asked ahead of time. How is information going to be organized and where will it be stored? What software packages will be used? Who are the potential users? How much is the project going to cost? The following sections explore some of these issues and explain the choices made in the development of the Dental Informatics portal.

Managing the Portal's Content

Building a basic website is not difficult even for a novice. Programs such as Front Page Express and Dream Weaver allow users to create Web pages even without an extensive knowledge of Hypertext Markup Language (HTML). However, managing a portal with even a few pages can become challenging. Different types of files such as text, graphics, video, and audio files, email archives, etc. all need to be "organized and made easily accessible." [14] A Content Management System (CMS) helps to do just that. It allows users to maintain site content, while providing necessary security, "without learning web programming, design or other web technologies." By using a user-friendly interface (such as a Web browser), authorized users can write text, format pages, upload files and manipulate data among other tasks.[15]

A well-implemented CMS can make controlling a site's content easier and provide cost savings in the long run. However, there are many systems available of different types and sizes, and choosing the wrong one may prove to be disastrous. Therefore, a careful evaluation and selection of the right CMS prior to developing a portal may be crucial to the success of the portal. [14, 15]

Search Engine vs. Directory

This section provides a brief distinction between search engines and directories. The goal of both is to make searching the Internet easier for the user, but the search mechanisms are different.

Search engines utilize programs called "spiders" or "agents" to "crawl" the Web in search of new websites. These websites are then automatically indexed using various algorithms and inserted into the search engine's database.[16]

The directories, also known as catalogs, are organized into categories and allow browsing. Traditionally, human editors have been used to manually index and categorize websites into these categories.[16] Manual (or human) indexing is defined as assigning indexing terms to documents (usually from a controlled vocabulary) by human indexers.[17] If adapted for this project, this approach may eventually require the development of a Dental Informatics controlled vocabulary, since the existing dental vocabularies (such as Index to Dental Literature) may lack granularity when describing documents related to Dental Informatics. While many directories still utilize manual indexing and categorization, automation may become a necessity considering the fast growing number of Internet documents.

Identifying the Audience

Identifying the intended audience may be one of the most important steps in the development of a portal. There are several groups of people that play a role in dentistry and, subsequently, in Dental Informatics:

- *Dental Professionals* – These include dentists, dental hygienists and dental assistants.
- *Dental Researchers* – This group is comprised of people involved in the academia, such as researchers or teachers.
- *Dental Patients* – The growth of the Internet created the so-called "informed consumer" and the need for organized dental patient information.
- *Commercial Parties* – These include vendors of dental clinical and educational software, third party payers, equipment and materials suppliers, and others.

Identifying potential users prior to development of the portal may help visualize their information needs and therefore help create a more complete user-oriented portal.

Methods

Due to the lack of resources and other constraints no comprehensive evaluation of Content Management Systems was performed in this project. Zope v. 2.5 was chosen primarily because of its free cost and widespread popularity and availability. Since Dental Informatics is a very small domain, no problems were anticipated due to the selection of Zope as the CMS. However, the lack of a comprehensive evaluation could prove problematic in the development of larger domain portals.

About Zope v. 2.5

Z Object Publishing Environment (Zope) is an open source framework for building Web applications.[18] Zope comes with a built-in Web server and is interoperable with Web servers such as Apache and IIS as well as any other Web server that supports the Common Gateway Interface (CGI). It is written in Python and is available for both Windows and Unix.[19, 20]

Zope architecture has many components, which are brought together into one system[16]. The features include but are not limited to the following:

- A built-in search engine and cataloging system
- Relational database connectivity
- Collaboration services
- Security with user log-ins
- Open Standards support
- Extensibility via Python [19, 20]

Several well-known companies and organizations have already adopted Zope including WebMD, Verizon Wireless, NASA / Space Telescope Science Institute, NATO and many others.[21]

Adapting Zope

Several issues arose in the course of this stage of the project. The first issue was to find a hosting service that supported Zope. The Zope's home page provided a list of hosting services.[21] Two major factors influenced the selection of a local provider. Since limited resources were allocated to this project, cost was a major consideration. The other factor was that, besides Zope access, the provider allocated a shell account to users. That allowed more flexibility in the project, such as the use of Perl scripts and database manipulation from the command line (for certain queries not supported by Zope).

Another issue was learning enough Zope features within the time frame allowed for this project. Zope has a "long and steep" learning curve.[19] Although documentation for Zope is abundantly available on the Internet, many features take time to learn. Unfortunately, support was not readily available from the hosting service. It often took weeks to get a reply and much e-mail remained unanswered.

Organizing DIP

The ability to browse among categories was the main reason for designing Dental Informatics Portal (DIP) as a directory. Since manual categorization is becoming challenging for larger domain portals, it was decided to implement an automatic classifier instead of manual categorization to prevent potential scalability problems as the domain expands. The development process of the automatic classifier is described later in this document.

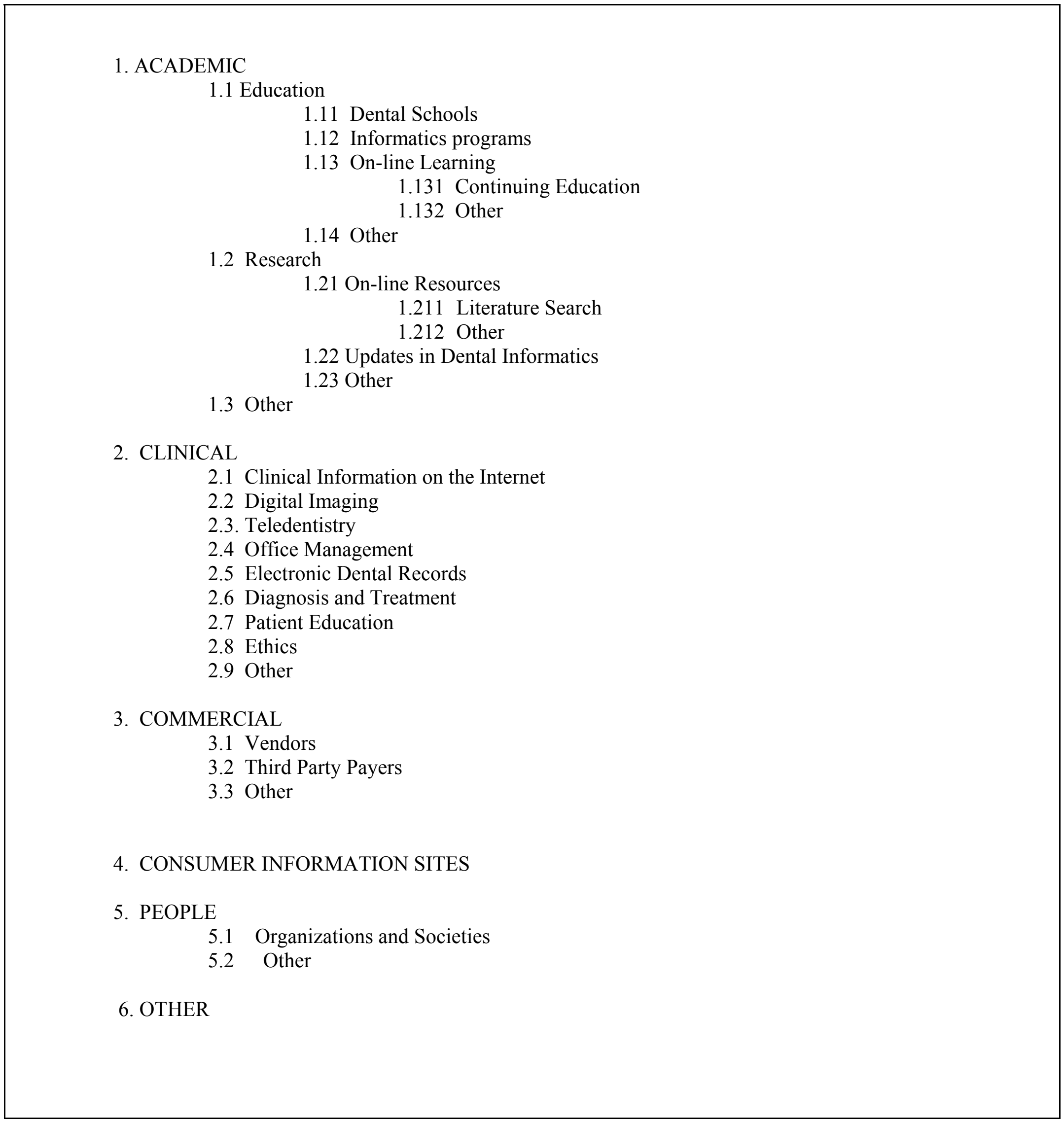
Although not the only way to organize a portal, a hierarchy of categories is the most commonly used and described structure on the Internet and in the related literature. Therefore it became a structure of choice for DIP.

The four groups described in the background section (dental professionals, researchers, patients and commercial parties) became the first level categories in the DIP's hierarchy. A separate category was added to include organizations and societies related to all groups. The subcategories were derived from a review of the related literature and Web pages on the Internet (Fig.1). In addition, an "other" category was added to each level to provide for websites that did not fit into any particular pre-defined category. As validation, a number of experts in the field (such as members of the American Medical Informatics Association (AMIA) Dental Informatics (DI) working group and faculty members of Dental Informatics programs) were contacted through the AMIA e-mail list as well as directly by e-mail and asked to evaluate the preliminary categories and make suggestions on the improvement of the hierarchy. Unfortunately, despite multiple postings to the list, there were very few responses. Suggestions were taken via an anonymous Web form and recorded in the Zope database. No major changes were proposed at that time.

Developing an Automatic Classifier

As mentioned before, it is becoming challenging to categorize fast growing numbers of Internet documents. Therefore, it was decided to develop an automatic classifier instead of manual categorization.

Figure 1. DIP hierarchy.



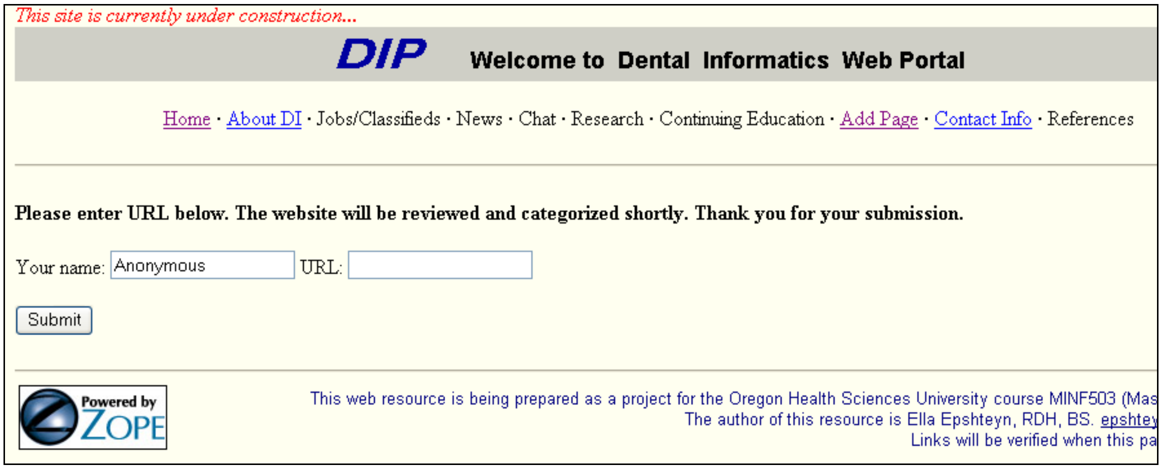
1. ACADEMIC
    - 1.1 Education
        - 1.11 Dental Schools
        - 1.12 Informatics programs
        - 1.13 On-line Learning
            - 1.131 Continuing Education
            - 1.132 Other
        - 1.14 Other
    - 1.2 Research
        - 1.21 On-line Resources
            - 1.211 Literature Search
            - 1.212 Other
        - 1.22 Updates in Dental Informatics
        - 1.23 Other
    - 1.3 Other
2. CLINICAL
    - 2.1 Clinical Information on the Internet
    - 2.2 Digital Imaging
    - 2.3. Teledentistry
    - 2.4 Office Management
    - 2.5 Electronic Dental Records
    - 2.6 Diagnosis and Treatment
    - 2.7 Patient Education
    - 2.8 Ethics
    - 2.9 Other
3. COMMERCIAL
    - 3.1 Vendors
    - 3.2 Third Party Payers
    - 3.3 Other
4. CONSUMER INFORMATION SITES
5. PEOPLE
    - 5.1 Organizations and Societies
    - 5.2 Other
6. OTHER

Automatic categorization can be defined as a problem of automatically assigning documents (in this case websites) to pre-defined categories [22, 23]. The term "clustering" is sometimes used interchangeably with "categorization". There is, however, an important difference between the two terms. Clustering does not sort documents into predefined categories, but rather builds previously unknown categories depending on the contents of the text.[24] Clustering could be very useful in the development of additional categories from the contents of the "other" categories. This work, however, is outside of the scope of this project.

At this stage the portal is designed so that users themselves can submit relevant pages for categorization (Figure 2). Submissions will be reviewed to rule out spam and automatically categorized by a Perl classifier. Eventually, a spider may be used to locate pages on the Internet, however it is beyond the scope of this project.

Only a small portion of the classifier has been completed. It is hypothesized that Perl's regular expressions can be effectively used to categorize Web pages. Regular expressions are templates that define collections of possible strings (with a string being a sequence of non-white space characters) [25]. An example of a regular expression in Perl would be

Figure 2. DIP URL Submission Page.



This site is currently under construction...

DIP Welcome to Dental Informatics Web Portal

Home · About DI · Jobs/Classifieds · News · Chat · Research · Continuing Education · Add Page · Contact Info · References

Please enter URL below. The website will be reviewed and categorized shortly. Thank you for your submission.

Your name: Anonymous URL:

This web resource is being prepared as a project for the Oregon Health Sciences University course MINF503 (Mas
The author of this resource is Ella Epshteyn, RDH, BS. epshtey
Links will be verified when this pa

\bcomput.*. The symbol "\b" in Perl is a word-boundary anchor, in this case marking a beginning of the word, while ".*" means "any number of any characters". In this case ***\bcomput.**** would match the words "computing", "compute", "computational", but not "recomputed". In a more complex example "**digit.*\s+(imag.*|rad.*)**" would match both "digital radiography" and "digitized image". The goal of this part of the project is to demonstrate the concept of use of regular expressions in this context, not to achieve a perfect method for website categorization.

Several tasks had to be performed prior to development of the classifier. A convenience sample of approximately 130 Dental Informatics Web pages was selected by searching the Internet. Search terms included general terms such as "dental informatics" and "dental computing" as well as narrower terms such as "digital radiography", "computerized dental records" and "continuing dental education". During the initial planning stages of the project, it was hoped that the same Dental Informatics experts that provided feedback regarding the hierarchy would also manually categorize the sample. However, it turned out to be more difficult to recruit volunteers than originally anticipated. Therefore, the sample was manually categorized by the author of this project. A single category of the hierarchy with the largest number of websites, "education," was selected for development of the classifier prototype. The websites in that subset were then assigned into two sets – training set and evaluation set. Each set contained 13 pages. The training set of websites was used to derive a set of regular expressions while the evaluation set was used to test the classifier. It was expected that a single Web page could belong in more than one category, however identifying multiple categories was beyond the scope of this project.

The goal was to develop a Perl program that would correctly identify at least a portion of websites in the evaluation set. Since this set had already been manually assigned to this category, the only two choices for the program would be either putting a website into this category or not doing so.

The next step was to analyze the websites in the training set. Several supporting scripts were written for this purpose. The first one utilized Perl's *LWP::Simple* module to parse a website's HTML and to separate markups from the body of the document. The parsed text in both the markups and body was then filtered to remove the stop words and to stem every word in the document. Although stemming was not crucial in this case, it was anticipated that using a stemmed form of the words would produce more accurate regular expressions. The Porter stemming algorithm was used for this purpose because of its availability [26]. This particular algorithm only removes suffixes and not prefixes. More research might result in selection of a more complete algorithm. The complete code for these scripts can be found in Appendix A.

Automatic categorization methods can utilize the analysis of markups (or meta-tags), which represent attributes of the words and structure of the documents.[27] For example, text inside of a <TITLE> tag as well as "description" or "keyword" meta tags can be more important for categorization than the rest of the text. It was anticipated that the final program would first look at the text inside of markups before moving on to the rest of the document. Unfortunately, few documents in the training set contained useful information in meta-tags. Therefore markup analysis was not a useful technique on this set .

Table 1. Examples of regular expressions with mapped synonyms.

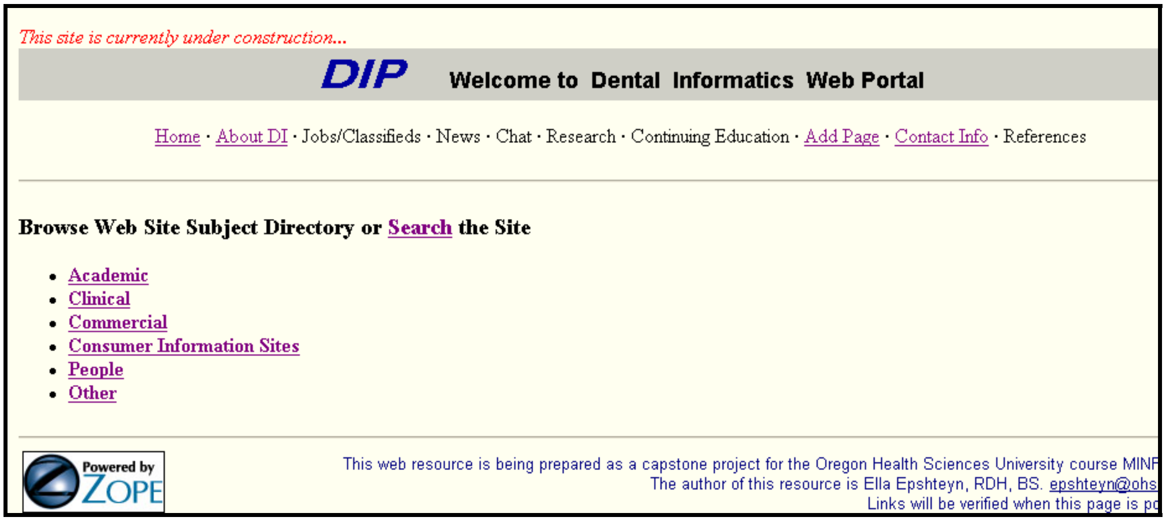
<table>
<tr><th>Words and Phrases</th><th>Synonyms</th><th>Perl Regular Expressions</th></tr>
<tr><td>clinical education</td><td>education, training</td><td>clinic\w+\s+(educ|train)</td></tr>
<tr><td>dental school</td><td>school, college, institution</td><td>dental\.+(school|college|institut)</td></tr>
<tr><td>computer assisted learning</td><td>assisted, aided</td><td>comput\w+\s(assist\w+|aid\w+)\s+(learn|</td></tr>
<tr><td></td><td>learning, education, training</td><td>educ|train)</td></tr>
<tr><td>educational software</td><td>educational software, computer tutorial</td><td>(educ\w+\s+softwar|comput\w+\s+tutor)</td></tr>
</table>

A set of regular expressions was derived following parsing and filtering websites in the training set. Basic mapping of synonyms was performed at this point. Several examples of these expressions with mapped synonyms are presented in Table 1. A Perl program was then written to match these regular expressions with the text in the evaluation set. The code for this script can be found in Appendix B.

Results

An empty portal shell is in place at this time. The home page of the portal is illustrated in Figure 3. The website submission page is connected to Zope's relational database ready for website submissions. Manual categorization by human editors can be performed to make the portal functional.

Figure 3. Dental Informatics Portal's home page.



This site is currently under construction...

DIP Welcome to Dental Informatics Web Portal

Home · About DI · Jobs/Classifieds · News · Chat · Research · Continuing Education · Add Page · Contact Info · References

Browse Web Site Subject Directory or Search the Site

- Academic
- Clinical
- Commercial
- Consumer Information Sites
- People
- Other

Powered by ZOPE

This web resource is being prepared as a capstone project for the Oregon Health Sciences University course MINF
The author of this resource is Ella Epshteyn, RDH, BS. epshteyn@ohs
Links will be verified when this page is po

Testing the portion of the classifier developed for this project produced promising results. All of the 13 pages in the evaluation set, previously manually categorized to the selected category, were identified as belonging to that category by the automatic classifier. When tested with 10 random websites not related to dentistry or Dental Informatics (selected by browsing different categories on Yahoo), the false positive rate was 0. However, when the classifier was tested with 10 Dental Informatics websites from other categories, 7 out of 10 websites contained regular expressions that would identify them as belonging to the selected category when they did not belong. Since no work on differentiating between the categories had been performed, these findings did not present a problem but rather demonstrated that more work was needed to create an effective classifier.

Discussion

Several problems were encountered in the course of the project. Zope documentation indicated that Zope supported Perl scripts as well as Python scripts. This would have been a very useful feature since several supporting Perl scripts were going to be utilized for various tasks throughout the project. However, the hosting provider failed to install the Perl module for Zope and was not responsive to requests for help. As mentioned before, cost was a major consideration in the selection of this particular provider. In the future, it may be beneficial to perform a comparison study of different hosting providers and select one based on the quality of their technical support rather than cost.

Recruiting Dental Informatics experts was another issue. Despite multiple postings to the AMIA e-mail list and other attempts to recruit experts, the author received very few responses. Since validation of categories may be important, perhaps other methods of recruiting domain experts or other methods of validation should be explored.

Developing a methodology is a process. Most phases of this project will continue to be improved in the future (for a example, more categories may be added to the hierarchy). The automatic classifier will go through many more stages of development. A comprehensive markup analysis as well as, perhaps, other methods of categorization may be added. The ability to tell the difference between the categories is an important feature still to be developed. That could be based on combinations and frequencies of occurrences of regular expressions. It is expected that the classifier will occasionally fail to categorize websites outside of the evaluation set. However, even negative results provide a valuable insight into how to improve the approach to this development. Website classification into any of the "other" categories should not be seen as failure by the classifier to recognize the website, but rather an opportunity to come up with additional categories or regular expressions to further categorize websites. Some websites may still have to be manually indexed. A limited controlled vocabulary for Dental Informatics may have to be developed for this purpose. Ideally, a comprehensive vocabulary of dental and dental informatics terms would need to be developed or built upon the existing vocabularies. Future work could also include adding a spider to locate Web pages, automatic maintenance of the portal, and the demonstration of usability. While it is difficult to draw any conclusions about the effectiveness, scalability and generality of the described methodology, this project may be a good starting point for development of small domain portals.

Summary and conclusion

To summarize, the following phases of the project have been completed:

- Selection of the Content Management System
- Design of the preliminary hierarchy
- Partial development of the automatic classifier prototype

Considering limited functionality of the classifier and the small samples of websites in both the training and evaluation sets, conclusions about the effectiveness of the classifier prototype could not be drawn at the time of completion of this project. Any conclusions about the effectiveness of methodology used for DIPs design would be premature as well. A comprehensive usability study will need to be performed in the future to assess the effectiveness of this approach.

As the field of Dental Informatics continues to evolve, more information will become available on the Internet. Large commercial portals such as Yahoo are not capable of offering adequate depth and detail to domains this small. Although more research is needed to develop an effective methodology for their design, small domain portals such as DIP may be the solution. A Web portal specific to Dental Informatics may provide users with a way to efficiently locate and retrieve information on specific topics. Having this information may in turn help dental offices to keep up with the updates in the field, help patients to make informed decisions about their care, provide people in academia with important research questions and encourage vendors to develop new technologies.

References:

1. Bauer JC P. The digital transformation of oral health care. Teledentistry and electronic commerce. JADA 2001;132.

2. Schleyer TKL D, PhD. Digital dentistry in the computer age. JADA 1999;131.

3. Schleyer TKL D, PhD, Pham T. Online continuing dental education. JADA 1999;130.

4. Versteeg CH SG, van der Stelt PF. Efficacy of Digital Intra-oral Radiography in Clinical Dentistry. J Dentistry. 1997;25:215-24.

5. Miles DA LR, Parks ET. Digital X-rays Are Here; Why Aren't You Using Them? J Calif Dent Assoc. 1999;27:926-33.

6. Paurazas SB GJ, Pink FE, Hoen MM, Steiman HR. Comparison of diagnostic accuracy of digital imaging by using CCD and CMOS-APS sensors with E-speed film in the detection of periapical bony lesions. Oral Surg Oral Med Oral Pathol Oral Radiol Endod. 2000;89:356-62.

7. Schleyer T. D, PhD, Spallek H., DMD, PhD. Dental Informatics. A Cornerstone of Dental Practice. JADA 2001;132.

8. Nielsen J. Jakob Nielsen's Alertbox:Intranet Portals: The Corporate Information Infrastructure, 1999.

9. The Concise Oxford Dictionary of Current English: Clarendon Press, 1995.

10. Pickering C. A Look Through the Portal. Software Magazine 2001;21.

11. Types of Portals.
http://www.dtic.mil/dtic/dtic-e/portal/types.html (05-03-2003)

12. Katz R. Web Portals & Higher Education: John Wiley & Sons, Inc, 2002.

13. Van Bemmel JH MM. Handbook of Medical Informatics: Houten/Diegem, 1997.

14. Paul F. Enterprise Technology: Choosing the Right Content Management System. PC World 2001: July.
http://www.pcworld.com/resource/printable/article/0,aid,50428,00.asp (04-28-2003).

15. Lurie I. A Web Content Management Blueprint.
http://www.portentinteractive.com/library/cmsexplained.pdf (04-28-2003).

16. Search Engines vs. Directories
http://www.aizee.dk/english/about_search_engines.asp (05-03-2003).

17. Hersh W. Information Retrieval: A Health Care Perspective: Springer-Verlag New York, Inc., 1996.

18. Browning P. Zope - a Swiss Army Knife for the Web?
http://www.ariadne.ac.uk/issue25/zope/ (04-28-2003).

19. Litt S. Zope: Quick and Simple. Linux Productivity Magazine 2002:1:4.
http://www.troubleshooters.com/lpm/200211/200211.htm (04-28-2003).

20. Latteier A., Pelletier M. The Zope Book: New Riders Publishing 2002.

21. Zope Home Page.
http://www.zope.org/ (04-28-2003).

22. Ko Y., Park J., Seo J. Automatic Text Categorization using the Importance of Sentences. http://acl.ldc.upenn.edu/coling2002/proceedings/data/area-28/co-269.pdf (04-28-2003).

23. Luo H. Experiments on Automatic Categorization of Broadcasting News. http://www.ee.columbia.edu/~luoht/research/paper/text_cat.pdf (04-28-2003).

24. What is automatic text filtering and categorization? http://www.nada.kth.se/theory/humanlang/textfilter.html (04-28-2003)

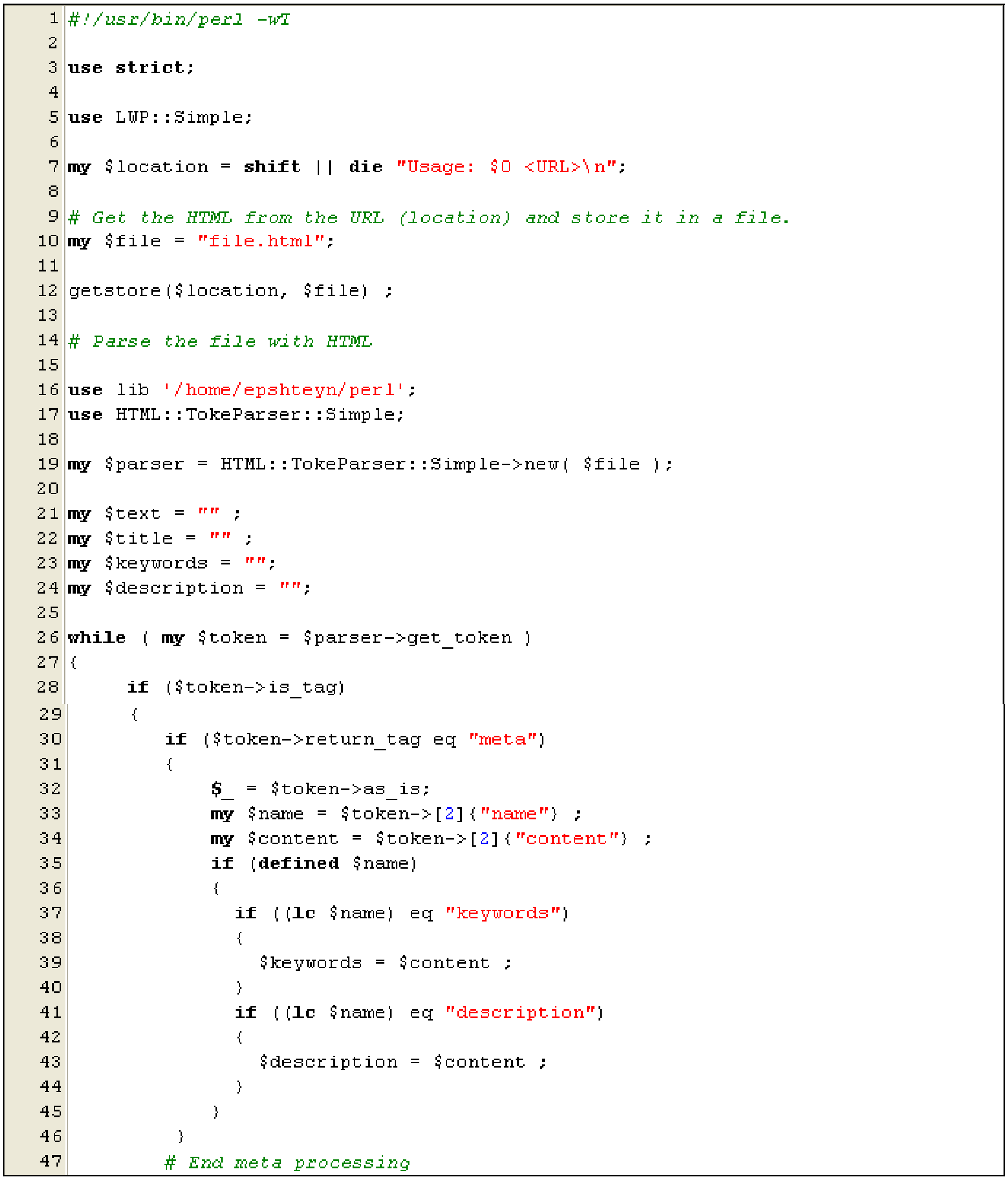
25. Schwartz R. Learning Perl: O'Reily & Associates, Inc., 1993.

26. Porter Stemming Algorithm. http://www.tartarus.org/%7Emartin/PorterStemmer/perl.txt (04-29-2003)

27. Multilevel Automatic Categorization for Web Pages. http://www.isoc.org/inet98/proceedings/1x/1x_5.htm (05-03-2003)

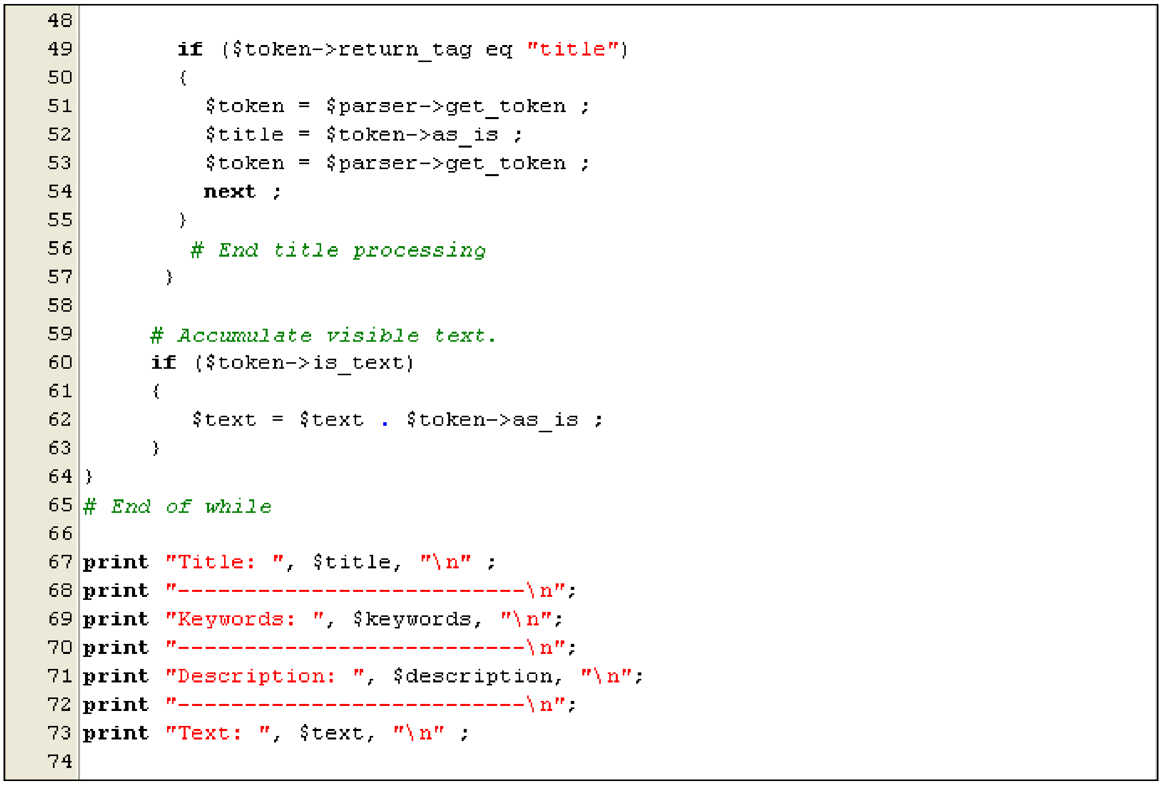
Appendix A. Supporting Perl Scripts.

a. ParseHTML.pl: Accesses URL, parses, separates markups.



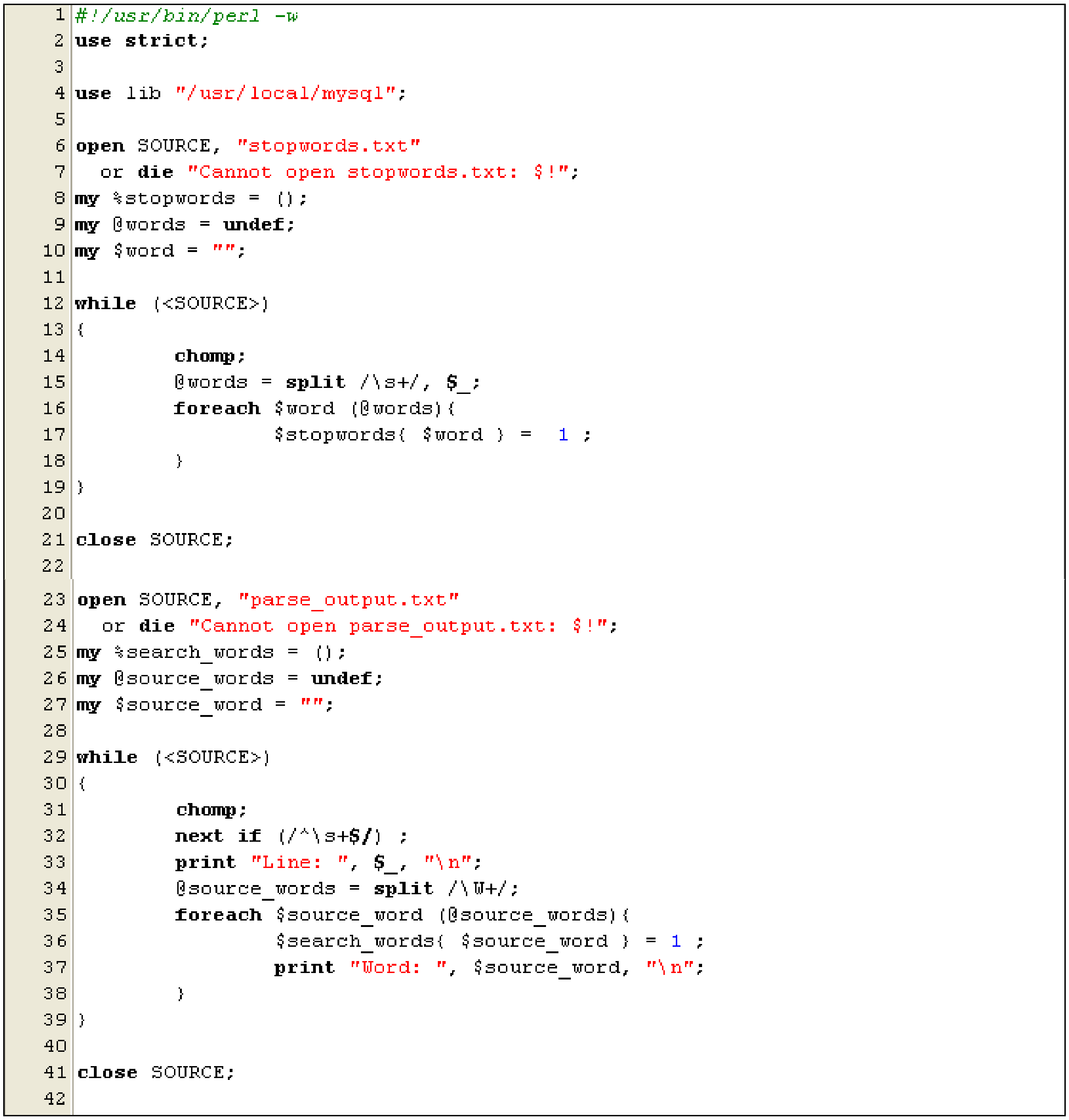
```
#!/usr/bin/perl -wT

use strict;

use LWP::Simple;

my $location = shift || die "Usage: $0 <URL>\n";

# Get the HTML from the URL (location) and store it in a file.
my $file = "file.html";

getstore($location, $file) ;

# Parse the file with HTML

use lib '/home/epshteyn/perl';
use HTML::TokeParser::Simple;

my $parser = HTML::TokeParser::Simple->new( $file );

my $text = "" ;
my $title = "" ;
my $keywords = "";
my $description = "";

while ( my $token = $parser->get_token )
{
     if ($token->is_tag)
     {
        if ($token->return_tag eq "meta")
        {
            $_ = $token->as_is;
            my $name = $token->[2]{"name"} ;
            my $content = $token->[2]{"content"} ;
            if (defined $name)
            {
              if ((lc $name) eq "keywords")
              {
                $keywords = $content ;
              }
              if ((lc $name) eq "description")
              {
                $description = $content ;
              }
            }
         }
        # End meta processing
```

continued from previous page.



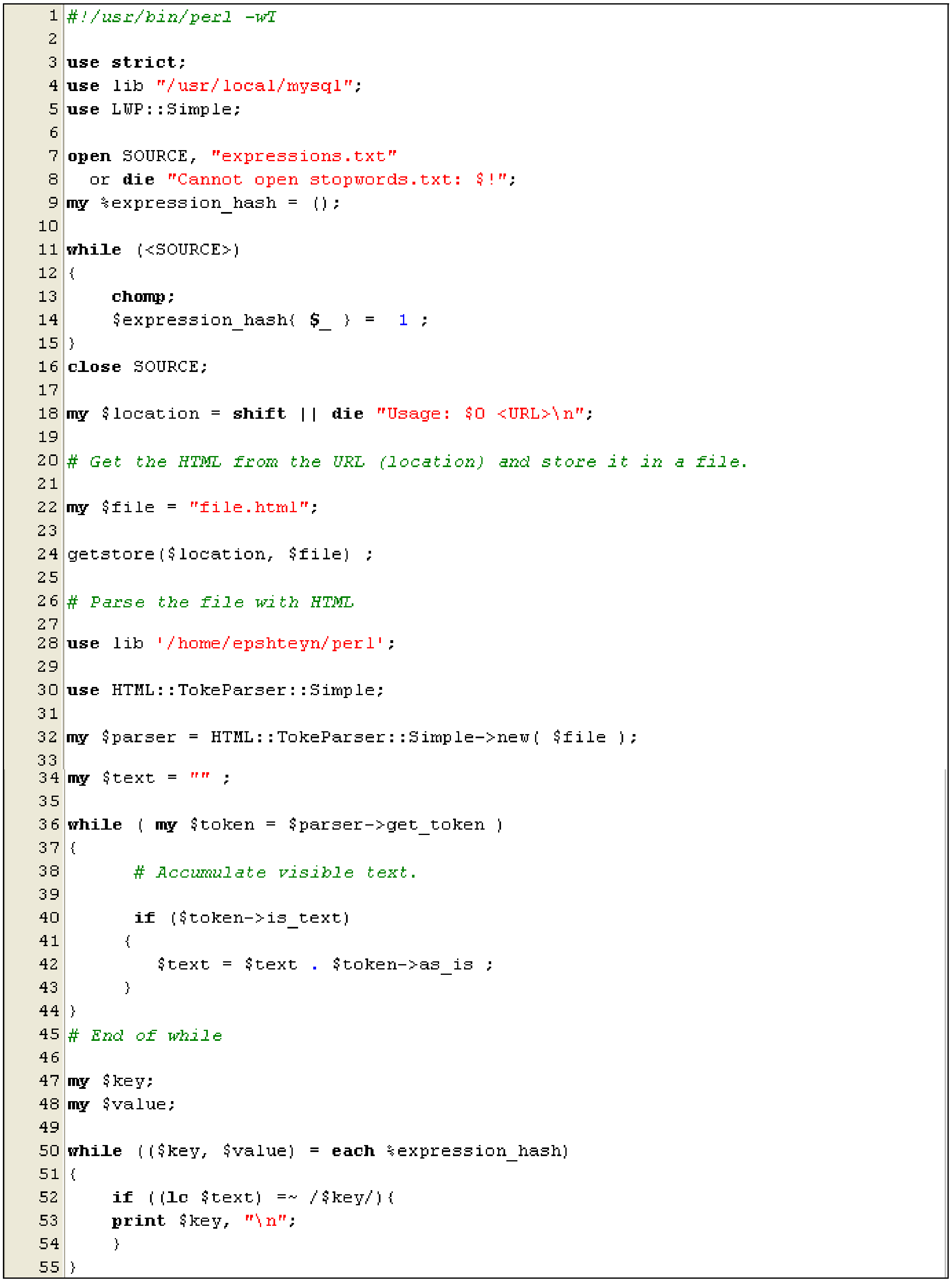
```perl

       if ($token->return_tag eq "title")
       {
         $token = $parser->get_token ;
         $title = $token->as_is ;
         $token = $parser->get_token ;
         next ;
       }
        # End title processing
      }

     # Accumulate visible text.
     if ($token->is_text)
     {
        $text = $text . $token->as_is ;
     }
}
# End of while

print "Title: ", $title, "\n" ;
print "--------------------------\n";
print "Keywords: ", $keywords, "\n";
print "--------------------------\n";
print "Description: ", $description, "\n";
print "--------------------------\n";
print "Text: ", $text, "\n" ;

```

b. FilterStopwords.pl: Opens ParseOutput.txt, removes stop words.



```
#!/usr/bin/perl -w
use strict;

use lib "/usr/local/mysql";

open SOURCE, "stopwords.txt"
  or die "Cannot open stopwords.txt: $!";
my %stopwords = ();
my @words = undef;
my $word = "";

while (<SOURCE>)
{
        chomp;
        @words = split /\s+/, $_;
        foreach $word (@words){
                $stopwords{ $word } =  1 ;
        }
}

close SOURCE;

open SOURCE, "parse_output.txt"
  or die "Cannot open parse_output.txt: $!";
my %search_words = ();
my @source_words = undef;
my $source_word = "";

while (<SOURCE>)
{
        chomp;
        next if (/^\s+$/) ;
        print "Line: ", $_, "\n";
        @source_words = split /\W+/;
        foreach $source_word (@source_words){
                $search_words{ $source_word } = 1 ;
                print "Word: ", $source_word, "\n";
        }
}

close SOURCE;

```

Appendix B. Automatic Classifier: matches regular expressions to text from a website.



```perl
#!/usr/bin/perl -wT

use strict;
use lib "/usr/local/mysql";
use LWP::Simple;

open SOURCE, "expressions.txt"
  or die "Cannot open stopwords.txt: $!";
my %expression_hash = ();

while (<SOURCE>)
{
    chomp;
    $expression_hash{ $_ } =  1 ;
}
close SOURCE;

my $location = shift || die "Usage: $0 <URL>\n";

# Get the HTML from the URL (location) and store it in a file.

my $file = "file.html";

getstore($location, $file) ;

# Parse the file with HTML

use lib '/home/epshteyn/perl';

use HTML::TokeParser::Simple;

my $parser = HTML::TokeParser::Simple->new( $file );

my $text = "" ;

while ( my $token = $parser->get_token )
{
      # Accumulate visible text.

      if ($token->is_text)
     {
        $text = $text . $token->as_is ;
     }
}
# End of while

my $key;
my $value;

while (($key, $value) = each %expression_hash)
{
    if ((lc $text) =~ /$key/){
    print $key, "\n";
    }
}
```