OPIM 311, Spring 2010, Case Assignment #5: Text Mining & Competitive Intelligence

Steven O. Kimbrough

19 April 2010
Due: 2010-05-03: Case assignment #5 due by midnight on webCafé.
Contents

1 Overview 3

2 Background & Context 3

3 Activity Framework 7
   3.1 Formulate .............................................................. 7
   3.2 Identify ................................................................. 7
   3.3 Obtain ................................................................. 8
   3.4 Produce ............................................................... 8

4 Uses and Validities 8

5 Assignment Tasks and Deliverables 11
   5.1 Preparatory Steps: Directory structure ............................. 11
   5.2 Assignment Tasks and Deliverables: Specific ....................... 12

A The os.path module 14

B The re module 15

C The string module and string methods 15

D Reference works 15
   D.1 Text Mining ............................................................ 15
   D.2 Content Analysis ..................................................... 16
   D.3 Toolkit ............................................................... 16
   D.4 Folksonomies ........................................................ 16

E Sources of Free Content for Business/Technology Material 17
   E.1 Scientific Documents – Free Journals ............................ 17
   E.2 Patents & Trademarks Websites .................................... 17
   E.3 Legal Websites ........................................................ 18
   E.4 Corporate Financials Websites ..................................... 19
   E.5 Product Specifications (Directories) ............................... 19
   E.6 Selected Business Websites (Free Content) ...................... 20

References 21
1 Overview

This assignment is about using Python (and more generally, scripting languages) to support competitive intelligence activities by employing text mining (recovering patterns of information from bodies of text).

There is considerable evidence that word patterns in text documents can be used to make predictions, or at least to support plausible diagnoses. A recent paper by Tetlock et al. (2008) (see also the earlier Tetlock (2007)) makes a credible case that word patterns in news stories carry information that can be used to predict stock prices. That study used an ontology of “negative” tags developed independently as part of the General Inquirer project. See http://www.wjh.harvard.edu/~inquirer/ and http://www.webuse.umd.edu:9090/. You can see the entire collection of tag sets at http://www.webuse.umd.edu:9090/tags/. We’ll be working with the negative tags http://www.webuse.umd.edu:9090/tags/TAGNeg.html (used by Tetlock et al., 2008) as well as the positive tags http://www.webuse.umd.edu:9090/tags/TAGPos.html You should consider innovative uses to which these tag sets might be applied.

What we’ll be doing is to process a small number of text files. In practice, these could be SEC filings, 10-K reports (annual reports) for companies listed on major exchanges, excerpts from blogs, newspaper articles, and so on; just about anything. We’ll produce two reports, one that counts occurrences of positive tag words and one that counts negative tag words. The reports will be written to files in a form that can easily be imported into Excel or some other spreadsheet program. Examples of the files are available: outPut.txt for the raw output and word-reports.xls for the Excel workbook file.

The results are at least moderately interesting for the small number of cases (only 11) we examine. Clearly, scaling up and careful statistical and/or graphical examination is called for if any solid conclusions are to be reached. But that is beyond the scope of this exercise. Instead, the point is to show how certain information can be extracted and presented or organized. Scaling up is then a straightforward matter.

I have strived to keep this rather complex exercise as simple as possible. If you can do this exercise, you can do a lot more, given more time, time beyond the scope of an introductory course. Besides taking away the skills to be able to do this, I hope also that you will be able to imagine ways in which the lessons can be generalized and put to exciting use in practice.

2 Background & Context

I’ll let others handle the necessary definitions.

A broad definition of Competitive Intelligence is the action of gathering, analyzing, and applying information about products, domain constituents, customers, and competitors for the short term and long term planning needs of an organization. Competitive Intelligence (CI) is both a process and a product.[1] The process of collecting,
storing and analyzing information about the competitive arena results in the actionable output of intelligence ascertained by the needs prescribed by an organization.

Key points of this definitions:

1. Competitive Intelligence is an ethical and legal business practice. (This is important as CI professionals emphasize that the discipline is not the same as industrial espionage which is both unethical and usually illegal).
2. The focus is on the external business environment. [2]
3. There is a process involved in gathering information, converting it into intelligence and then utilizing this in business decision making. CI professionals emphasize that if the intelligence gathered is not usable (or actionable) then it is not intelligence.

A more focused definition of CI regards it as the organizational function responsible for the early identification of risks and opportunities in the market before they become obvious. This definition focuses attention on the difference between dissemination of widely available factual information (such as market statistics, financial reports, newspaper clippings) performed by functions such as libraries and information centers, and competitive intelligence which is a perspective on developments and events aimed at yielding a competitive edge.

The term CI is often viewed as synonymous with Competitor analysis but Competitive Intelligence is more than analyzing competitors it is about making the organization more competitive relative to its existing set of competitors and potential competitors. Customers and key external stakeholders define the set of competitors for the organization and, in so doing, describe what could be a substitute for the business, votes, donations or other activities of the organization. The term is often abbreviated as CI, and most large businesses now have some Competitive Intelligences functions with staff involved often being members of professional associations such as the Society of Competitive Intelligence Professionals.

The Society of Competitive Intelligence Professionals (SCIP) is an organization for those who are interested in learning more about Competitive Intelligence. Established in 1986, it provides education and networking opportunities for business professionals, and provide up to date market research and analysis. “Members of the SCIP have backgrounds in market research, strategic analysis, science and technology.”

(From http://en.wikipedia.org/wiki/Competitive_intelligence, 2009-04-10.)

Here are SCIP’s vision and mission statements:

SCIP Vision and Mission Statements
VISION:
Better decisions through competitive intelligence.

MISSION:
SCIP will be the global organization of choice for professionals engaged in competitive intelligence and related disciplines. SCIP will be the premier advocate for the skilled use of intelligence to enhance business decision-making and organizational performance.

(From http://www.scip.org/, 2009-04-10.)

Back to the Wikipedia for the term text mining:

Text mining, sometimes alternately referred to as text data mining, roughly equivalent to text analytics, refers generally to the process of deriving high-quality information from text. High-quality information is typically derived through the dividing of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities).

(From http://en.wikipedia.org/wiki/Text_mining, 2009-04-10.)

Text mining is a term, body of literature, and bundle of techniques that has recently come to the fore in, broadly, the information sciences communities (typically, in schools of information, computer science and computational linguistics groups, and information systems groups in business schools). Content analysis is a much older term, body of literature, and bundle of techniques, dating from the 1950s and associated, broadly, with humanities and communication studies. Both are relevant to competitive intelligence.

Content analysis is a research tool used to determine the presence of certain words or concepts within texts or sets of texts. Researchers quantify and analyze the presence, meanings and relationships of such words and concepts, then make inferences about the messages within the texts, the writer(s), the audience, and even the culture and time of which these are a part. Texts can be defined broadly as books, book chapters, essays, interviews, discussions, newspaper headlines and articles, historical documents, speeches, conversations, advertising, theater, informal conversation, or really any occurrence of communicative language. Texts in a single study may also represent a variety of different types of occurrences, such as Palmquist’s 1990 study of
two composition classes, in which he analyzed student and teacher interviews, writing journals, classroom discussions and lectures, and out-of-class interaction sheets. To conduct a content analysis on any such text, the text is coded, or broken down, into manageable categories on a variety of levels—word, word sense, phrase, sentence, or theme—and then examined using one of content analysis’ basic methods: conceptual analysis or relational analysis

(From http://writing.colostate.edu/guides/research/content/pop2a.cfm 2009-04-11.)

The Wikipedia as expected has an entry, although it is perhaps not up to the expected quality. Still, it is useful:

Content analysis is a methodology in the social sciences for studying the content of communication. Earl Babbie defines it as “the study of recorded human communications, such as books, websites, paintings and laws.” It is most commonly used by researchers in the social sciences to analyze recorded transcripts of interviews with participants.

Content analysis is also considered a scholarly methodology in the humanities by which texts are studied as to authorship, authenticity, of meaning. This latter subject include philology, hermeneutics, and semiotics.

Harold Lasswell formulated the core questions of content analysis: “Who says what, to whom, why, to what extent and with what effect?.” Ole Holsti (1969) offers a broad definition of content analysis as “any technique for making inferences by objectively and systematically identifying specified characteristics of messages.” Kimberly A. Neuendorf (2002) offers a six-part definition of content analysis:

“Content analysis is an indepth analysis using quantitative or qualitative techniques of messages using a scientific method (including attention to objectivity-intersubjectivity, a priori design, reliability, validity, generalizability, replicability, and hypothesis testing) and is not limited as to the types of variables that may be measured or the context in which the messages are created or presented.”

(From http://en.wikipedia.org/wiki/Content_analysis 2009-04-11.)

Not that I endorse every word or sentiment, but these are close enough to the mark to let us proceed.
3 Activity Framework

The purpose of this section is to present a simple, introductory framework for describing how to support competitive intelligence through text data mining. Our framework has four steps—

1. **Formulate** questions of interest.
2. **Identify** documents or texts likely to support answering the questions of interest.
3. **Obtain** collections of the documents of interest.
4. **Produce** answers to the questions of interest and produce reports.

—which we elaborate upon briefly in the sections that follow immediately.

3.1 Formulate

A main aspect of the art of text mining is to identify valuable questions that it can contribute to answering. Like any tool, there are some things it does well and some it does not do well. We cannot characterize the possibilities here, except to say that our general approach is to apply groups of related words (“ontologies” or “taxonomies” etc.) to various texts and that there is some accumulation of evidence that this can often be used to detect performance and make predictions. In the present exercise we are only concerned with the techniques of assembling such reports. That they may have value is presumed on good evidence. What in a particular case their value is, is something beyond the scope of the present exercise.

3.2 Identify

Being able to identify documents that credibly have the information needed is a necessary and easily neglected achievement. There is a tendency to think of the Web as holding everything and of Google as giving us effective access. This is unfortunate and far from accurate. Even when the documents are located on the Web, finding them and separating them from documents we don’t want remains an enduring challenge. Perhaps the best possibility is to find focused collections of documents and download from these. Examples:

1. ‘Screen scraping’ from Internet search engines, such as Google and Yahoo!
   
   Also, these engines typically have APIs that allow you do retrieval URLs from searches under program control. Much easier this way.

2. SEC filings. www.sec.gov This is where we got the documents for this exercise.


4. Organization Web sites. Use Wget to download any available files.
5. News reports and periodicals.
6. Other regulatory filings.
7. Court filings and legal proceedings.
8. Organization archives. See records management systems, ECM (enterprise content management), and so on. See, e.g.,
   - http://www-01.ibm.com/software/data/content-management/
   - The trade association is http://www.aiim.org/

See appendix E, page 17. The original document is ‘Mac Free Sites.doc’, which is posted on webCafé.

3.3 Obtain

The technical requirements for actually obtaining required documents depend entirely on the documents in question and range in difficulty from trivial to dauntingly hard and labor intensive. Being clever or insightful in identifying the required documents can help a lot.

3.4 Produce

Given formulated questions and having obtained adequate collections of documents that plausibly contain answers to our questions, it remains to produce those answers.

Present-day methods for producing information from collections of text, for text mining, amount largely to identifying and counting words and word phrases (called n-grams), then using this information to construct features which are in turn examined statistically (including graphically and in report tables). The present exercise is a case in point. Feature construction and statistical analysis can be surprisingly subtle and fraught with difficulties, although much of value can be done simply. This is a subject that lies on the frontier of current knowledge and know-how.

4 Uses and Validities

In text mining we process documents in order to produce data that we hope will be of value. These data are, in any event, noisy, biased, and in many ways problematic. Better techniques can be helpful, so there is unlimited scope for innovation in text mining. Even so, we must be reconciled to data from text that are in many ways of much lower quality than primary data we often encounter. [Witten and Frank] (2005, page 351) express the point well:
Data mining is about looking for patterns in data. Likewise, text mining is about looking for patterns in text: it is the process of analyzing text to extract information that is useful for particular purposes. Compared with the kind of data we have been talking about in this book, text is unstructured, amorphous, and difficult to deal with. Nevertheless, in modern Western culture, text is the most common vehicle for the formal exchange of information. The motivation for trying to extract information from it is compelling—even if success is only partial.

(But why the point should be limited to “modern Western culture” is a mystery to me.)

In the end it comes down to “Is the game worth the candle?”, “Is the view worth the hike?”, “Is the effort worth the reward?” Often—increasingly so in text mining—it is.

This raises questions of uses and validities. To what uses can text mining be put effectively? and How can we assess or understand the validity of a given application? A little framework, given in Table 1, will be helpful.

<table>
<thead>
<tr>
<th>Type of Use</th>
<th>Type of Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Statistical</td>
</tr>
<tr>
<td>Prediction</td>
<td></td>
</tr>
<tr>
<td>Detection</td>
<td></td>
</tr>
<tr>
<td>Filtering</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Uses-validities framework for text mining

Points arising:

1. How to judge whether a model is statistically valid is a more or less straightforward thing. Standards and methods are broadly agreed to. Of course, finding a statistically valid model is far from straightforward or non-trivial. The achievement of Tetlock et al. (2008) (among other investigators) is to find statistically valid models (and methodology for generating models) by which predictions may be made from data derived by text mining. The news here is quite positive. Demonstrations have been made that statistically valid predictive models can be developed based on text mining. The ongoing challenge in this area is to find testable opportunities. Stock prices, the subject of (Tetlock et al., 2008), are atypical in having so much high accuracy data that can be subject to prediction and response.

2. Statistical validity is, of course, the best kind for empirical models. But we can’t always have the best. Sometimes, even usually, we need to assess the face validity of a result and if the judgment is favorable, we use the result. It is simply the case that under the circumstances to hand we have to rely on our judgment (or judgments, if they can be collected). It is for this kind of validity, face validity, that text mining is very promising, especially in the scope of its potential applications, compared to statistical validity.

3. In the world of data not derived from text we often use graphics, careful tabular displays and exploratory data analysis methods to arrive at plausible conjectures, thereby achieving face
validity. This can be, and is commonly, done for purposes of making predictions (making forecasts; establishing trends and patterns) and for purposes of detection (noticing changes, events, departures from established pattern).

All of this can be, and is starting to be, done routinely for text mining. See for example Blog-Pulse [http://www.blogpulse.com/](http://www.blogpulse.com/) Many organizations regularly scan the web, chat rooms, blogs, etc. for chatter about their products, their competitors, their suppliers. Visualizations based on text mining have been invented and shown to be effective. (See for example [Dworman et al., 2000](http://opim-sky.wharton.upenn.edu/~sok/sokpapers/2000-1/jasis-2000-dworman-kimbrough-patch.pdf); see also [Dworman, 1999a](http:).

Another kind of example involves **folksonomies**, folk taxonomies generated by collaborative tagging. See [http://www.steve.museum/](http://www.steve.museum/) for a very substantial application in the world of museums and archives.

4. Support for detection of events, relationships, changes, etc. with face validity is well within range of small, scripting-based projects, such as our current exercise. Note that by using the Excel spreadsheets, you can not only see the summary information at the top, you can explore the data. Sorting is an especially useful tool here. How do the firms differ in their patterns of words?

   (Note that horizontal (“left to right”) sorting would be informative, too, but Excel is not up to the task with such a large data set. You could do it in Python, however.)

5. Internet search engines, such as Google and Yahoo!, do retrieval, as in the field of **Information Retrieval**.

   Information retrieval (IR) is the science of searching for documents, for information within documents and for metadata about documents, as well as that of searching relational databases and the World Wide Web. There is overlap in the usage of the terms data retrieval, document retrieval, information retrieval, and text retrieval, but each also has its own body of literature, theory, praxis and technologies. IR is interdisciplinary, based on computer science, mathematics, library science, information science, information architecture, cognitive psychology, linguistics, statistics and physics.

   Automated information retrieval systems are used to reduce what has been called “information overload”. Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications.

Text mining is something quite different. Search and retrieval find documents (records). Text (and data) mining find patterns and information in and among the records. Search and retrieval are record-oriented; mining is pattern-oriented (Dworman et al., 2000).

The two concepts merge to a degree, however, in the case of filtering applications. A conventional IR engine (think: Internet search engine) may retrieve a collection of putedly relevant documents, but these documents themselves need to be culled in various ways to find what is most interesting. That culling is the job of a filtering engine.

5 Assignment Tasks and Deliverables

5.1 Preparatory Steps: Directory structure

This exercise continues where Case #4 left off. Recall that in Case #4 we downloaded a number of files from the Web and stored them as strings (text files) in the following directory structure:

downloads/
   ATT/
      ATT 10K 2008.txt
   BerkshireHathaway/
      Berkshire Hathaway 10K 2008.txt
   ...
   Walmart/
      Walmart 10K 2009.txt

If we abstract (generalize) this directory structure, we have (replacing downloads with originals):

originals/
   <taxon/>*
   <source file>*

Here and below, we indicate directories by a trailing (ending) / in their names. rltr/ = root of the little text repository. A trailing * indicates 0 or more such entities. A name of a directory or a file surrounded by angle brackets, <...>, indicates a directory or file of the mentioned type. Here is the basic directory structure we shall use in Case #5. Notice that it contains the directory structure used in Case #4.
There are two files under the miscindex/ directory:
  miscindex/
      negative-words.txt
      positive-words.txt

The originals/ directory may contain 1 or more <taxon/> subdirectories, each of which contains (we can assume for now) exactly 1 text (string) file. For development and testing purposes, you may use the files you downloaded in Case #4. We shall post other test directories later.

5.2 Assignment Tasks and Deliverables: Specific

Given a directory of originals/ as described above, your task is to produce a report that counts the number of occurrences of the positive words (or negative words, as the case may be) in each of the text documents under the originals/ directory. As a means of doing this, you will first have to ‘tokenize’ the original text files, removing HTML tags et cetera and reducing them to lists of words (as much as possible).

Specifically, do the following:

1. Convert the documents under originals/ to simple text (string) documents consisting of words separated by white space. These documents should be stored under tokenized/, which should mirror originals/.

Note well:

(a) You can get a list of the files and directories in originals/ by using os.listdir(). However, when processing the list you should add a check to exclude any unwanted files. You only want directories and you will need to create directories with the same names under tokenized/. See appendix A for some helpful methods from the os.path module.

(b) Your original files may be (typically will be) HTML files, which include a lot of things in HTML tags, and these should be removed.

For example,
should be reduced (at least) to:

'\nForm 10-K

10-K
1
d10k.htm
FORM 10-K

'What you need to do is to use regular expressions to remove every pair of angle brackets and what is between them.

(c) When you remove something, say an HTML tag, you should probably leave a blank space where it was, rather than eliminate everything.

(d) There may (will) remain a number of HTML special characters, such as &nbsp; . They all begin with & and end with ; (and do not otherwise contain either & or ;). You should remove these, too.

(e) Words that occur at the end of sentences will have trailing punctuation, which has to be eliminated. Hint: Use regular expressions to extract the words from their trailing punctuation.

(f) Count as words only strings consisting of letters of the Roman alphabet and the dash. Think: [a-zA-Z\-].

The end result of this task is that you convert all the files in originals/ to text files in tokenized/, with the same subdirectory structure. The resulting text files should insofar as possible consist of words in lower case separated by white space, without punctuation or HTML markup.

2. Write to a file a report in CSV form, word-reports.csv, suitable for importing into Excel, including formulas. Specifically:

(a) Read in and load up as a dictionary either the negative words, negative-words.txt, or the positive words, positive-words.txt, as options. Here is a simple and not very elegant, but workable way to do this. Early on in your program file, tokenize.py, add a line like this

# Pick your tag set:
ontology = 'positive-words.txt'
and use
ontology = 'negative-words.txt'

for the other case.

(b) Process each file under tokenized/, and determine the number of times each positive (negative) word appears in the file. Note:

- In the General Inquirer tag sets there are a number of cases in which more than one sense of a word is identified. For example, we find this in the positive tags:

  ADVANCE#1 IAV Pos SUPV Travl Pstv Strng Actv | 47% verb:
  To move or bring forward, improve, promote
  ADVANCE#2 Pos Noun Pstv Strng Actv Means | 20% noun-adj:
  A moving forward, improvement, approach, in front, prior
  ADVANCE#3 Pos Modif Virtue Pstv Psv | 20% adj:
  "Advanced"--forward, in front, progressive

  You will need to reduce this to one term, in this case, “advance”. So you have to drop the # and what appears after it.

(c) Write out a text file, word-reports.csv, with the following format or structure. The first column should begin with two empty cells, followed by all of the words in the positive (or negative) word list, one word per row. The first row should consist of an empty cell, followed by the names of the taxa (or company names) of the source files. The second row should consist of an empty cell, followed by a formula for Excel to calculate the total of the entries in the cells below it, for each of the rows. The remaining cells should display the counts of the number of occurrences of the word (row heading) in the document associated with the taxon (column heading). This file should be in comma-separated form (CSV) and should display properly when imported into Excel.

3. You should turn in two files: tokenize.py (for step 1, above) and generatereport.py (for step 2 above).

4. You should also turn in a short report, just a page or two. Write anything that you think will be helpful to the grader in examining what you did.

A  The os.path module

See [http://docs.python.org/3.0/library/os.path.html?highlight=os.path#module-os.path](http://docs.python.org/3.0/library/os.path.html?highlight=os.path#module-os.path)

Especially useful:

- os.path.exists(path)

  “Return True if path refers to an existing path. Returns False for broken symbolic links. On some platforms, this function may return False if permission is not granted to execute os.stat() on the requested file, even if the path physically exists.”
os.path.exists

• os.path.split(path)

“Split the pathname path into a pair, (head, tail) where tail is the last pathname component and head is everything leading up to that. The tail part will never contain a slash; if path ends in a slash, tail will be empty. If there is no slash in path, head will be empty. If path is empty, both head and tail are empty. Trailing slashes are stripped from head unless it is the root (one or more slashes only). In nearly all cases, join(head, tail) equals path (the only exception being when there were multiple slashes separating head from tail).”

os.path.splitext(path)

• os.path.splitext(path)

“Split the pathname path into a pair (root, ext) such that root + ext == path, and ext is empty or begins with a period and contains at most one period. Leading periods on the basename are ignored; splitext(‘.cshrc’) returns (‘.cshrc’, ”).”

B The re module

See: http://docs.python.org/3.0/library/re.html?highlight=re#module-re

HOW-TO http://docs.python.org/3.0/howto/regex.html#regex-howto

C The string module and string methods

See: http://docs.python.org/3.0/library/string.html?highlight=str#module-string

But especially see:
http://docs.python.org/3.0/library/stdtypes.html#string-methods

D Reference works

For your edification. Not needed to do this assignment.

D.1 Text Mining

1. (Feldman and Sanger 2007).
2. (Han and Kamber, 2006). Mainly on data mining. Has some brief but useful material on text mining.

3. (Konchady, 2006).

4. (Witten and Frank, 2005). Mainly on data mining. Has some brief but useful material on text mining.

D.2 Content Analysis


2. (Neuendorfl, 2002)

3. (Popping, 2000)


D.3 Toolkit

1. (Friedl, 2006)

D.4 Folksonomies

Collaborative tagging systems. A folksonomy is what a “folk classification” system is called. Were a hot idea a few years ago. Intriguing possibilities for organizing texts, but few clear “killer aps” so far. Still, it’s interesting. There are important tag-based systems, including: http://del.icio.us/ (the first!), http://technorati.com/, http://flickr.com/


And the inevitable Wikipedia entry:

Folksonomy (also known as collaborative tagging, social classification, social indexing, and social tagging) is the practice and method of collaboratively creating and managing tags to annotate and categorize content. Folksonomy describes the bottom-up classification systems that emerge from social tagging.[1] In contrast to traditional subject indexing, metadata is generated not only by experts but also by creators and consumers of the content. Usually, freely chosen keywords are used instead of a controlled vocabulary.[2] Folksonomy (a portmanteau of folk + taxonomy) is a user-generated taxonomy.

(From http://en.wikipedia.org/wiki/Folksonomy, accessed 2009-4-11.)

E Sources of Free Content for Business/Technology Material

Assembled by M. Halperin, Penn librarian, 2/20/2009.

E.1 Scientific Documents – Free Journals

Directory of Open Access Journals
http://www.doaj.org/
Physet Free Physics Journals
http://physnet.uni-oldenburg.de/PhysNet/journals.html
Chemistry Free Chemistry Journals
http://www.abc.chemistry.bsu.by/current/fulltext.htm
Biology Free Biology Journals
Medicine Free Medical Journals
http://www.freemedicaljournals.com/

E.2 Patents & Trademarks Websites

http://gethelp.library.upenn.edu/guides/business/patentslistweb.html

• Delphion - Patent Information - Intellectual Property Research
  http://www.delphion.com/
• Derwent - Some free info.
• European Patent Office
• Google Patent Search
• Intellectual Property Mall - Franklin Pierce Law Center
• NBER U.S. Patent Citations
• PatentCafe - Includes a patent glossary.
• PatentLaw Links.com
• United States Copyright Office - Library of Congress
• United States Patent and Trademark Office - U.S. Dept.of Commerce
• USPTO Glossary
• World Intellectual Property Organization(WIPO)

E.3 Legal Websites

http://gethelp.library.upenn.edu/guides/business/legalweb.html
• ABI World – “The premier site for bankruptcy information.”
• Business.GOV Business Laws - Legal & regulatory info SBA
• Corporate Crime Reporter
• Economic Crime Survey - Price Waterhouse Coopers
• FreeAdvice.com
• Getting the Deal Through –International corporate law and policy
• InterNet Bankruptcy Library, http://bankrupt.com/
• LawCrawler
• Legal Definitions.com
• Legal Engine
• Legal Forms
• Patents and Trademarks
• Securities Class Action Clearinghouse
• Taxes
• Technology Deals
• Transparency International - Combats corruption worldwide.
• U.S. Supreme Court Opinions - 1889-present Full text searching
• vLex
E.4 Corporate Financials Websites

http://gethelp.library.upenn.edu/guides/business/corpfinweb.html

- AnnualReports.com
- Annual Reports Service - The Public Register [Registration Required]
- Companies requiring CEO, CFO Certification of Financial Statements - SEC
- EasyReports.Net
- EDGAR Company Search
- EDGAR Database of Corporate Information - U.S. SEC
- IRIN - Investor Relations Information Network - Annual reports in Adobe Acrobat
- PRARS Annual Report Service
- Public Reference Room - SEC
- Search-sec.com inc
- SEC Info - US (SEC EDGAR) and Canada (CSA SEDAR)
- SEDAR - Canadian public company filings.
- The Annual Reports Library

E.5 Product Specifications (Directories)

http://gethelp.library.upenn.edu/guides/business/compdirectref.html

- Chemical Week Buyer’s Guide
- Corporate Information
- The Definitive ISP Buyer’s Guide
- The Electronic Blue Book of Building and Construction
- Global Sources - Supplier and product info for buyers worldwide
- Local Google - Locate local businesses by U.S. city, town, or zip code
- Hoover’s Online
- Institute Of Management Consultants USA - Directory
• Manufacturing.net - Manufacturing Yellow Pages
• O’Dwyer’s Directory of Public Relations Firms
• The Scannery - Websites of companies worldwide
• Thomas Regional Electronic Network Directory (TREND)
• Thomas Register of American Manufacturers
• zapdata.com – D&B – 14 million companies [Registration required]
• Partminer (fee required)
  [http://www.partminer.com/entsol.jsp?sub=svcs]
• GlobalSpec.com [http://www.globalspec.co]

E.6 Selected Business Websites (Free Content)
[http://gethelp.library.upenn.edu/guides/business/businesswebsites.html]

• Accounting
• Banks and Banking
• Business News+
• Business Reference
• Compensation
• E-commerce
• Economic Conditions and Statistics
• Energy & Environmental Resources
• Entrepreneurship and Small Business
• Finance and Investment
• Forecasting
• Government Resources – Business
• Health Care
• Industries
• Information Technology
• Job Hunt
• Labor Legal Aspects of Business
• Management
• Market Research
• Marketing & Advertising
• Media
• Multinational Business
• Nonprofit Organizations
• Patents & Trademarks
• Public Relations
• Real Estate
• Risk & Insurance
• Search Engines
• Taxes
• Trade
• Transportation
• Working Papers

References


