

Search Quality Rating Guidelines

Version 1.0

November 2, 2012

Preface	4
Part 1: Rating Guidelines	5
1.0 Welcome to the Search Quality Rating Program!	5
1.1 URL Rating Overview	5
1.2 Important Rating Definitions and Ideas	5
1.3 The Purpose of Search Quality Rating	6
1.4 Raters Must Represent the User	6
1.5 Internet Safety Information	6
2.0 Understanding the Query	7
2.1 Understanding User Intent	7
2.2 Task Language and Task Location	8
2.3 Queries with Multiple Meanings	8
2.4 Classification of User Intent: Action, Information, and Navigation – “Do-Know-Go”	9
3.0 The Language of the Landing Page	12
4.0 The Rating Scale	12
4.1 Vital	12
4.2 Useful	13
4.3 Relevant	13
4.4 Slightly Relevant	14
4.5 Off-Topic or Useless	14
4.6 Unratable	14
5.0 Rating: From User Intent to Assigning a Rating	17
5.1 User Intent and Page Utility	17
5.2 Location is Important	18
5.3 Language is Important (This section is for Non-English Task Languages)	18
5.4 Multiple Interpretations	18
5.5 Specificity of Queries and Landing Pages	21
5.6 Common Rating Problems	21
6.0 Flags	25
6.1 Spam Flag	25
6.2 Pornography Flag	25
6.3 Malicious Flag	26
6.4 Compatibility between Ratings and Flags	26

Part 2: URL Rating Tasks with Query Locations	27
1.0 Query Locations	27
2.0 Location-Specific Rating Task Screenshot.....	29
3.0 Assigning a Rating When There is a Query Location	30
3.1 When Does the Query Location Matter?	30
Part 3: Webspam Guidelines	32
1.0 What is Webspam ?	32
1.1 The Relationship between Ratings and Spam	32
1.2 Why do Spammers Create Spam Pages?	32
1.3 When to Check for Spam.....	33
2.0 Looking for Technical Signals	33
2.1 Hidden Text and Hidden Links.....	33
2.2 Keyword Stuffing	34
2.3 Sneaky Redirects	35
2.4 Cloaking	36
3.0 Helpful Webpages vs. Spam Webpages.....	36
3.1 Pages with Copied Content and PPC Ads	37
3.2 Fake Search Pages with PPC Ads	39
3.3 Fake Blogs with PPC Ads.....	39
3.4 Fake Message Boards with PPC Ads	39
3.5 Copied Content that is NOT Spam	39
4.0 Commercial Intent	39
4.1 Thin Affiliates	40
4.2 Pure PPC Pages	41
4.3 Parked (Expired) Domains	41
4.4 Pages with Unhelpful Content and PPC Ads.....	42
5.0 Phishing Websites	42
6.0 Spam and the Resolving Stage	42
7.0 Conclusion	43

Preface

Google relies on raters, working in countries and languages around the world, to help us measure the quality of our search results, ranking, and search experience. These raters perform a variety of different kinds of “rating tasks” designed to give us information about the quality of different kinds of results in response to different kinds of queries. The data they generate is rolled up statistically to give us within the Google search team a view of the quality of our search results and search experience over time, as well as an ability to measure the effect of proposed changes to Google’s search algorithms. Raters’ judgments do not directly impact Google’s search result rankings. While a rater may give a particular URL a score, that score does not directly increase or decrease a given website’s ranking. Instead these scores are used in aggregate to evaluate search quality and make decisions about changes.

This document is a “Cliff’s Notes” version of our search quality rating guidelines. By this, we mean that it is not the entire version that raters use on a daily basis; however, it is a summary of the important topics. The raters’ version includes instruction on using the rating interface, additional rating examples, etc. These guidelines are used as rating specifications for search raters, and this document in particular focuses on a core type of rating task called “URL rating.” In a URL rating task, a rater is shown a search query from their locale (country + language) and a URL that could be returned by a search engine for that query. The raters “rate” the quality of that result for that query, on a scale described within the document. Sounds simple, right? As you’ll see, there are many cases to think through, and this document is used to guide raters on some of those cases and how to look at them.

Our search quality rating guidelines are in constant flux as we learn and search evolves over time. We’ve created this version especially for those individuals who want to understand better how Google thinks about relevance and quality of search results.

Part 1: Rating Guidelines

1.0 Welcome to the Search Quality Rating Program!

As a Search Quality Rater, you will work on many different types of rating projects. These guidelines cover just one type of search quality rating – URL rating.

Please take the time to carefully read through these guidelines. The ideas presented here are important for other types of rating. When you can do URL rating, you will be well on your way to becoming a successful Search Quality Rater!

1.1 URL Rating Overview

For each URL rating task you acquire, you will see a query and a URL. You will:

- Research the query
- Click on the URL to visit the landing page
- Assign a rating based on these guidelines

1.2 Important Rating Definitions and Ideas

Search Engine: A search engine is a website that allows users to search the Web by entering words or symbols into a search box.

Query: A query is the set of word(s), number(s), and/or symbol(s) that a user types in the search box of a search engine. We will sometimes refer to this set of words, numbers, or symbols as the “query terms.” Some people also call these “key words.” In these guidelines, queries will have square brackets around them. If a user types the words *digital cameras* in the search box, we will display: [digital cameras].

User Intent: When a user types a query, he is trying to accomplish something, such as finding information or purchasing an item online. We refer to this goal as the user intent.

Task Language and Task Location: Queries have a task language and task location associated with them and will look like this in these guidelines: **[digital cameras], Spanish (ES)**. This format indicates that the query *digital cameras* was typed into a search box by a Spanish reading user in Spain. Task locations are represented by a two-letter country code. The country code for Spain is ES. If the query had been typed by a Spanish reading user in Mexico, it would look like this: **[digital cameras], Spanish (MX)**.

For a current list of country codes, visit

http://www.iso.org/iso/country_codes/iso_3166_code_lists/country_names_and_code_elements.htm

Homepage (of a website): When we use the term “homepage,” we are referring to the main page of a website. It is the first page that users see when the website loads. The URL for the homepage of a website usually ends with .com, .edu, .org, .gov, etc., or the two-letter code for a country outside the US, such as .jp, .mx, .ru, etc. For example, <http://www.apple.com/> is the homepage of the Apple computer company website, and <http://www.mcdonalds.com/> is the homepage of the McDonald’s hamburger corporation website. We are aware that some countries use the term homepage to refer to the entire website of a company, organization, individual, etc. However, we use “homepage” to refer to the main page only.

Subpage: A page on a website that is not the homepage. For example, <http://www.apple.com/iphone/> is a subpage on the Apple website. An example of a subpage on the McDonald’s website is http://www.mcdonalds.com/usa/rest_locator.html.

Webpage or Web Page: Any page on a website. It may be the homepage or a subpage of the website.

URL: The URL is the Web address of the webpage you will evaluate, such as <http://www.microsoft.com>. It is important to look at the URL, but remember that you will evaluate the landing page.

Landing Page or Page: This refers to the webpage that you will evaluate. It is the page you see after you click on the URL. These guidelines will explain how to evaluate the content of the landing page. You may see ads and sponsored links on many landing pages. You will evaluate only the content posted by the webmaster. Your rating will not be based on ads or sponsored links on the page (even if they are related to the query).

Topic: The topic of the query is the focus or subject of the query; it is what the query is about. Users typing the query want to find pages on the Web that are related to the topic of the query.

Utility: The utility of the landing page is a measure of how helpful the page is for the user intent. Pages with good utility are helpful for users. Pages with no utility are useless. **Utility is the most important aspect of search engine quality, and is therefore the most important thing for you to think about when evaluating webpages.**

The Rating Scale will be described in detail in Section 4 of the Rating Guidelines (Part 1 of the General Guidelines), but here is a brief overview. For each task, you will assign exactly one of the following ratings:

Rating Scale	Description
<i>Vital</i>	A special rating category. See Section 4.1 of the Rating Guidelines.
<i>Useful</i>	A page that is very helpful for most users.
<i>Relevant</i>	A page that is helpful for many or some users.
<i>Slightly Relevant</i>	A page that is not very helpful for most users, but is somewhat related to the query. Some or few users would find this page helpful.
<i>Off-Topic or Useless</i>	A page that is helpful for very few or no users.
<i>Unratable</i>	A page that cannot be evaluated. A complete description can be found in Section 4.6 of the Rating Guidelines.

You will also assign any of the following **flags** that apply: **Not Spam**, **Maybe Spam**, **Spam**, **Porn**, and **Malicious**. They will be discussed in Section 6 of the Rating Guidelines.

1.3 The Purpose of Search Quality Rating

Your ratings will be used to evaluate search engine quality around the world. Good search engines give results that are helpful for users in their specific language and location. Please note that your ratings do not directly impact Google's search result rankings or ranking algorithms.

1.4 Raters Must Represent the User

It is very important for you to represent the user. The user is someone who lives in your task location and reads the task language, and who has typed the query in the search box.

You must be very familiar with the task language and task location in order to represent the experience of users in your task location. If you do not have the knowledge to do this, please inform your employer.

1.5 Internet Safety Information

In the course of your work, you will visit many different webpages. Some of them may harm your computer unless you are careful. Please do not download any executables, applications, or other potentially dangerous files, or click on any links that you are uncomfortable with. **We strongly recommend that you have antivirus and anti-spyware protection on your computer. This software must be updated frequently or your computer will not be protected. There are many free and for-purchase antivirus and anti-spyware products available on the Web.**

Here are links to Wikipedia articles with information about antivirus software and spyware:

http://en.wikipedia.org/wiki/Antivirus_software
<http://en.wikipedia.org/wiki/Spyware>

We suggest that you only open files with which you are comfortable. Please feel free to release rating tasks if they contain unknown or suspicious file formats.

The file formats listed below are generally considered safe if antivirus software is in place.

- .txt (text file)
- .ppt or .pptx (Microsoft PowerPoint)
- .doc or .docx (Microsoft Word)
- .xls or .xlsx (Microsoft Excel)
- .pdf (PDF) files

If you encounter a page with a warning message, such as “Warning-visiting this web site may harm your computer,” or if your antivirus software warns you about a page, you should not try to visit the page to assign a rating. You should instead assign a rating of **Unratable: Didn’t Load**. A description of this rating can be found in Section 4.6.1 of the Rating Guidelines.

You may also come across pages that require RealPlayer or the Adobe Flash Player plug-in. These are safe to download at:

<http://www.real.com/>
http://www.adobe.com/shockwave/download/download.cgi?P1_Prod_Version=ShockwaveFlash

Examples of pages that require Flash Player are: <http://www.ferrariworld.com> and <http://www.atraircraft.com>.

2.0 Understanding the Query

Before you can evaluate the task, you must understand the query. Please use an online dictionary or encyclopedia that is available for your task location, or do web research to help you understand all of the words in the query.

Important: If you use a search engine to research the query, please do not rely only on the ranking of results that you see displayed on the search results page. A query may have other meanings besides those represented in the top results. Do not assign a high rating to a webpage just because it appears at the top of a list of search results.

Here are some examples of the kinds of resources available on the Web that may be helpful:

Online encyclopedias:

http://en.wikipedia.org/wiki/Main_Page: the English language version of Wikipedia
<http://www.wikipedia.org/>: portal to other language/locale versions of Wikipedia

Translation tools:

<http://babelfish.yahoo.com/>
<http://www.wordreference.com/>
<http://translate.google.com/>

2.1 Understanding User Intent

In addition to understanding the meaning of the query, you must also consider user intent. What was the user trying to accomplish when he typed the query? You will need to understand user intent to evaluate the landing page.

Consider the query [tetris], English (US). Most English speaking users in the United States who type this query know that Tetris is a popular computer game. The most likely user intent is to play the game online.

Here are some other examples of queries and user intents:

Query	Likely User Intent
[Fedex], English (US)	Track a package or find a Federal Express location.
[calendar], English (US)	Find, customize, and print a calendar for the current month or year. Find a calendar that displays holidays. Find an online calendar to use to organize one's time.
[ebay], English (US)	Buy or sell merchandise on eBay, or navigate to the eBay homepage.

2.2 Task Language and Task Location

All queries have a task language and task location. Keeping these in mind will help you to understand the query and user intent. Users in different parts of the world may have different expectations for the same query.

Query	Query Meaning in the Task Location	Likely User Intent in the Task Location
[football], English (US)	American football played with a brown oval ball.	Find recent game scores, game schedules, pictures, team information, etc. for American football in the US.
[football], English (UK)	The game Americans call soccer, played with a round ball.	Find recent game scores, game schedules, pictures, team information, etc. for soccer in the UK or perhaps around the world.

2.3 Queries with Multiple Meanings

Many queries have more than one meaning. For example, the query [apple], English (US) might refer to the computer brand or the fruit. We will call these possible meanings **query interpretations**.

Dominant Interpretation: The *dominant interpretation* of a query is the interpretation that most users have in mind when they issue the query. For example, most users typing [windows], English (US) want results on the Microsoft operating system, rather than the glass windows on a wall. The dominant interpretation should be clear to you, especially after doing a little web research.

Common Interpretations: In some cases, there is no dominant interpretation. The query [mercury], English (US) might refer to the car brand, the planet, or the chemical element (Hg). While none of these is clearly dominant, all are *common interpretations*. Many or some people might want results related to these interpretations.

Minor Interpretations: Sometimes you will find less common interpretations. These are interpretations that few users have in mind. We will call these *minor interpretations*. Consider again the query [mercury], English (US). Possible meanings exist that even most English (US) users probably do not know about, such as Mercury Marine Insurance and the San Jose Mercury News. These are minor interpretations.

When you evaluate pages associated with a minor interpretation of the query, you will use lower ratings on the Rating Scale. In Section 5.4 of the Rating Guidelines, we will discuss in detail how to rate pages when the query has multiple interpretations.

2.4 Classification of User Intent: Action, Information, and Navigation – “Do-Know-Go”

Sometimes it is helpful to classify user intent for a query in one or more of these three categories:

- **Action intent** – Users want to accomplish a goal or engage in an activity, such as download software, play a game online, send flowers, find entertaining videos, etc. These are “**do**” queries: users want to do something.
- **Information intent** – Users want to find information. These are “**know**” queries: users want to know something.
- **Navigation intent** – Users want to navigate to a website or webpage. These are “**go**” queries: users want to go to a specific page.

An easy way to remember this is **Do-Know-Go**. Classifying queries this way can help you figure out how to rate a webpage. Please note that many queries fit into more than one type of user intent.

2.4.1 Action Queries – “Do”

The intent of an **action query** is to accomplish a goal or engage in an activity on the Web. The goal or activity may be to download, to buy, to obtain, to be entertained by, or to interact with a resource that is available on the Web.

Users want to **do** something. Here are some examples of goals and activities:

- Purchase a product.
- Download software for free or for money.
- Pay a bill online.
- Play a game online.
- Print a calendar.
- Send flowers.
- Organize photos or order prints online.
- Watch a video clip.
- Copy an image or piece of clipart.
- Take an online survey.
- View entertaining webpages, such as pictures, gossip, videos, etc.

Helpful pages for an action query are pages that allow users to do the activity or accomplish the goal.

Query	Likely User Intent	Description of The Landing Page
[geography quiz], English US)	Take an online geography quiz.	Page with a working online geography quiz.
[Beatles poster], English (US)	Find an image of a Beatles poster or perhaps purchase a Beatles poster.	Page on which to view or purchase a Beatles poster.
[download adobe reader], English (US)	Download software.	Official free download page on the Adobe website.
[fairy tale coloring pages], English (US)	Print coloring pages.	Trustworthy page with printable coloring pages.
[online personality test], English (US)	Take an online personality test.	Page with a working online personality test.
[what is my bmi?], English (US)	Calculate the BMI (body mass index).	Page with a working BMI calculator.
[good cop baby cop], English (US)	View the “Good Cop, Baby Cop” video.	Page on which to view this video.
[cute kitten pics], English US)	View photos of cute kittens.	Page of cute kitten photos to look at.
[Citizen Kane DVD], English (US)	Purchase this DVD.	Pages on which to purchase this DVD.
[flowers], English (US)	Order flowers online.	Pages on which to order flowers online.

Query	Likely User Intent	Description of the Landing Page
[play sudoku], English (US)	Play Sudoku online.	Pages on which to play Sudoku.
[calculate running pace], English (US)	Calculate running pace online.	Page with a working running pace calculator.
[bubble spinner 2], English (US)	Play Bubble Spinner 2 online or download the game.	Page on which to play and/or download this game.
[Spanish English dictionary], English (US)	Translate Spanish words into English or English words into Spanish.	Page that translates words from English to Spanish and Spanish to English.

2.4.2 Information Queries – “Know”

An **information query** seeks information on a topic. Users want to **know** something; the goal is to find information.

Helpful pages have high quality, authoritative, and comprehensive information about the query.

Query	Likely User Intent	Description of The Landing Page
[Switzerland], English (US)	Find travel and tourism information for planning a vacation or holiday, or find information about the Swiss geography, languages, economy, etc.	Page about Switzerland on a well-known travel guide. Informative CIA World Factbook webpage on Switzerland.
[cryptology use in WWII], English (US)	Find information about how cryptology was used in World War II.	United States Air Force Museum article about cryptology use during WWII.
[how to remove candle wax from carpet], English (US)	Find information on how to remove candle wax from carpet.	Page on a well-known magazine website with this information.

2.4.3 Navigation Queries – “Go”

The intent of a **navigation query** is to locate a specific webpage. Users have a single webpage or website in mind. This single webpage is called the **target** of the query. Users want to **go** to the target page.

The most helpful page for a navigation query is the navigational target page.

Query	Likely User Intent	URL of the Target Page	Description of the Target Page
[ibm], English (US)	Go to the IBM homepage.	http://www.ibm.com/	Official homepage of the IBM Corporation.
[youtube], English (US)	Go to the YouTube homepage.	http://www.youtube.com/	Official homepage of YouTube.
[ebay], Italian (IT)	Go to the Italian eBay homepage.	http://www.ebay.it/	Official homepage of eBay Italy.
[harvard college admissions], French (FR)	Go to the Harvard College admissions page on the Harvard University website.	http://admissions.college.harvard.edu/index.html	Harvard College Office of Admissions page on the official Harvard University website.
[best buy store locator], English (US)	Go to the store locator page on the Best Buy website.	http://www.bestbuy.com/site/olspage.jsp?id=cat12090&type=page	Store Locator page on the official Best Buy website.

Query	Likely User Intent	URL of the Target Page	Description of the Target Page
[sony customer support], English (US)	Go to the customer support page on the Sony website.	http://esupport.sony.com/	eSupport page on the official Sony website.
[outback steakhouse menu], English (US)	Go to the menu page on the Outback website.	http://www.outback.com/menu/	Menu page on the official Outback Steakhouse website.
[canon.com digital cameras], English (US)	Go to the digital cameras page on the Canon website. Although Canon is primarily known for its digital cameras, the target of the query is the digital cameras page, not the Canon homepage.	http://www.usa.canon.com/consumer/controller?act=ProductCatIndexAct&categoryId=113	Digital Cameras page on the official Canon website.
[facebook login], English (US)	Go to the login page on the Facebook website. Although users can log in from the Facebook homepage, the target of the query is the login page, not the homepage.	http://www.facebook.com/login.php	Login page on the official Facebook website.

2.4.4 Queries with Multiple User Intents (Do-Know-Go)

Many queries have more than one likely user intent. Please use your judgment when trying to decide if one intent is more likely than another intent. Here are some examples.

Query	Likely User Intent	URL of a Helpful Page	Description of The Landing Page
[download firefox], English (US)	Do and Go . This could be a “do” and a “go” query. Users want to download the web browser Firefox (“do” user intent). Many users may want to download the browser from the official Firefox website (“go” user intent).	http://download.cnet.com/mozilla-firefox/	The landing page is the Firefox browser download page on the cnet.com website, which is a well-known, respected website. Many users would feel comfortable downloading from this site. This page is helpful for the “do” user intent.
		http://www.mozilla.com/en-US/firefox/firefox.html	The landing page is the official Firefox browser download webpage. This page may be the target of the query and is helpful for the “do” and “go” user intents.
[Nikon digital cameras], English (US)	Do, Know, and Go . This could be a “do” and a “know” and a “go” query. Users are probably interested in a Nikon digital camera. Some users may have decided to buy a Nikon (“do”), but some may be researching the Nikon brand (“know”), and some may want to go to digital camera pages on the Nikon website (“go”).	http://www.target.com/s/nikon+digital+cameras	The landing page is the “Nikon” page on the target.com website. There are over 60 models of Nikon digital cameras for sale and the page has prices, specifications, and reviews. This page is helpful for both the “do” and “know” user intents.
		http://reviews.cnet.com/digital-camera-reviews/?filter=1000036_108496_&tag=centerColumnArea1.0	The landing page is the “Nikon Digital cameras” review page on the cnet.com website, with helpful information about many different Nikon digital cameras organized by price, resolution, digital camera type, and features. The page allows users to compare prices, features, etc. This page is helpful for the “know” user intent.
[ipad], English (US)	Do, Know, and Go . This could be a “do” and a “know” and a “go” query. Users are probably interested in buying an iPad (“do”), but some may be doing research (“know”), and some may want to go to iPad pages on the Apple website (“go”).	http://www.engadget.com/2011/03/09/ipad-2-review/	The landing page on the engadget.com website has a comprehensive review of the iPad. This page is helpful for the “know” intent.
		http://www.apple.com/ipad/	The landing page is the iPad product page on the official Apple website. This page may be the target of the query and is helpful for the “know” and “go” user intents.
		http://store.apple.com/us/browse/home/shop_ipad/family/ipad?mco=OTY2ODA0NQ	The landing page is the iPad page on the Store part of the official Apple website. Users can make a purchase and find information. This page may be the target of the query and is helpful for the “do,” “know,” and “go” user intents.

3.0 The Language of the Landing Page

You are expected to read and understand your task language and English. You are also expected to have some understanding of commonly used languages for your task location.

All landing pages will be flagged as one of the following:

- **The task language**
- **An acceptable language**
- **English**
- **Foreign Language**
- **None of the above**

Task Language: Use the flag that corresponds to your task language when the page content is entirely or mostly in the task language.

Acceptable Language: Use the flag that corresponds to the appropriate acceptable language when the page content is entirely or mostly in an acceptable language. Acceptable languages are other languages that are commonly used by a significant percentage of the population in the task location. The rating task will display the acceptable languages for the task location.

English: Use this flag when the page content is entirely or mostly English.

Foreign Language: Use this flag when you believe users in the task location would NOT be able to read/understand the content of the page.

None of the above: Use this flag when there is no language on the page to identify. Examples are pages that are completely blank, pages with images only, or pages with so much garbled text or so many encoding errors that you cannot identify the language.

For mixed language pages: Use your best judgment. Do not struggle with your selection of a language flag.

4.0 The Rating Scale

The rating scale offers five rating options that are based on user intent and the utility of the landing page: **Vital**, **Useful**, **Relevant**, **Slightly Relevant**, and **Off-Topic or Useless**. In addition, there is a rating category that will be used in special circumstances: **Unratable**.

4.1 Vital

The **Vital** rating is used for these very special situations:

- 1) The dominant interpretation of the query is navigation, and the landing page is the target of the navigation query.
- 2) The dominant interpretation of the query is an entity (such as a person, place, business, restaurant, product, company, organization, etc.), and the landing page is the official webpage associated with that entity.

In both cases, the query must have a dominant interpretation. If there is no dominant interpretation, it is not possible to assign a **Vital** rating.

Most **Vital** pages are very helpful. However, please note that this is not a requirement for a rating of **Vital**. Some **Vital** pages are “official,” but not very helpful.

4.1.1 Important Vital Concepts

Most queries do not have **Vital** webpages. Here are situations for which there is no **Vital** page.

- The query does not have a dominant interpretation.
- The query is not an entity or is not a navigation query.
- No official website or webpage exists for the entity.
- No person or entity can “own” the topic of the query.

Here are some examples of queries that do not have **Vital** pages:

Query	Vital Page	Description
[ADA], English (US)	No Vital page is possible	There is no dominant interpretation. The following entities are all common interpretations. Each interpretation has an official homepage, but none is Vital since there is no dominant interpretation. Americans with Disabilities Act American Dental Association American Diabetes Association
[knitting], English (US)	No Vital page is possible	This is an information query. Knitting is an activity anyone can do and that anyone can create a website for. There is no one official source for knitting information. No one can own this topic.
[diabetes], English (US)	No Vital page is possible	This is an information query. No person or entity can claim ownership of the query [diabetes].
[ipod reviews], English (US)	No Vital page is possible	[ipod] is an entity query, but [ipod reviews] is not. [ipod reviews] is an information query. Users are looking for information that many sites can provide.
[how old is britney spears?], English (US)	No Vital page is possible	[Britney Spears] is an entity query, but [how old is britney spears] is not. This is an information query. Users are looking for information that many sites can provide.

4.2 Useful

A rating of **Useful** is assigned to pages that are very helpful for most users. **Useful** pages should be high quality and a good “fit” for the query. In addition, they often have some or all of the following characteristics: highly satisfying, authoritative, entertaining, and/or recent (such as breaking news on a topic).

Useful pages are usually well organized and pages you trust. They are from information sources that seem reliable. **Useful** information pages are not “spammy.”

Please note that more than one page can be rated **Useful** for a query.

4.3 Relevant

A rating of **Relevant** is assigned to pages that are helpful for many or some users. **Relevant** pages have fewer valuable attributes than were listed for **Useful** pages. **Relevant** pages should still “fit” the query, but they might be less comprehensive, less up-to-date, come from a less authoritative source, or cover only one important aspect of the query.

Relevant pages must be helpful for users, in addition to being on-topic. **Relevant** pages should not be low quality. **Relevant** pages are average to good.

4.4 Slightly Relevant

A rating of **Slightly Relevant** is assigned to pages that are not very helpful for most users, but are somewhat related to the query. **Slightly Relevant** pages may be low quality and/or contain less helpful information. **Slightly Relevant** pages may serve a minor interpretation, have outdated information, be too specific, too broad, etc. to receive a higher rating.

A rating of **Slightly Relevant** should also be assigned to mobile landing pages (which are related to the query) that appear in regular URL rating tasks. Pages that are designed for mobile users are different from pages designed for regular desktop/laptop users. The content displayed is different (usually, much less content is provided) and the functionality of the page is different, too. Of course, if the mobile landing page is *unrelated* to the query, a rating of **Off-Topic or Useless** is appropriate.

4.5 Off-Topic or Useless

A rating of **Off-Topic or Useless** should be assigned to pages that are helpful for very few or no users. **Off-Topic or Useless** pages are unrelated to the query and/or have no utility.

A rating of **Off-Topic or Useless** also applies when there is lack of attention to an aspect of the query that is important for satisfying user intent.

You will also come across pages that are so unhelpful (and possibly deceptive) that they should be rated **Off-Topic or Useless**. For example, you may be given a page to rate that has links and ads and no actual content. The links redirect to other pages that lead to yet other links and ads. When nothing on the page is helpful to the user, it should be rated **Off-Topic or Useless**. These pages usually warrant the **Spam** flag, which you will learn about in Section 6.1 of the Rating Guidelines.

4.6 Unratable

You will assign **Unratable** to pages that you are unable to evaluate. Because you will encounter different types of unratable pages, please use the following categories of **Unratable** to describe the results:

- **Didn't Load**
- **Foreign Language**

Please note that you may assign more than one **Unratable** rating to a page. For example, if the landing page displays an error message in a foreign language and has no content (i.e. the page belongs in the **Didn't Load** category as described in Section 4.6.1 of the Rating Guidelines), it should be assigned both **Unratable: Didn't Load** and **Unratable: Foreign Language**.

4.6.1 Unratable: Didn't Load

Unratable: Didn't Load (usually referred to as just **Didn't Load**) is a special rating category for pages that truly do not load or have any content at all. These pages typically display some kind of web server or web application error message *and no other content*.

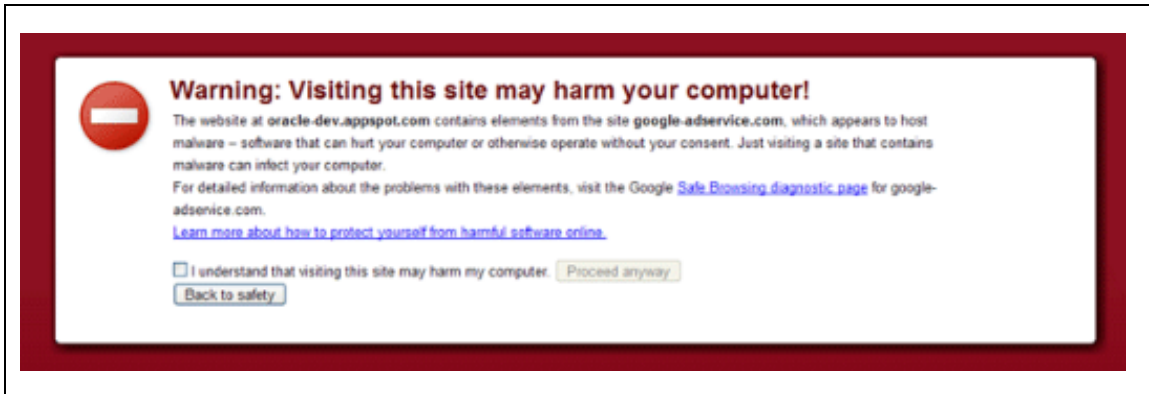
Pages that belong in the **Didn't Load** category include:

- Pages with error messages *and no other content on the page*.
- Pages with non-working redirects *and no other content on the page*.
- Completely blank pages.
- Pages with malware warnings, such as "Warning – visiting this web site may harm your computer!"
- Pages with certificate acceptance requests.

Please note that you should **not** assign a **Spam** or **Malicious** flag just because a security warning message or certificate acceptance request is displayed. There are some innocent pages that trigger these messages. If you encounter a warning message or certificate acceptance request, please assign a rating of **Didn't Load**. Do not assign a **Spam** or **Malicious** flag unless there is another reason to do so.

Descriptions of **Spam** and **Malicious** flags can be found in Sections 6.1 and 6.3 of the Rating Guidelines, respectively.

This is what a warning message might look like:



This is what a certificate acceptance request might look like:



See http://en.wikipedia.org/wiki/List_of_HTTP_status_codes for descriptions of different types of error messages. As you can see from this Wikipedia article, there are many types of web server errors and error messages. The most common types that you will see are:

- 401 - Unauthorized
- 403 - Forbidden
- 404 - Not Found
- 500 - Internal Error
- 503 - Service Unavailable

Pages that partially load or have some broken links should be rated on the rating scale according to their utility.

Here are examples of pages with these types of error messages (*and no other content*), which should be rated **Didn't Load**.

Query	Landing Page Error Message	Rating	Explanation
[Douglas Instruments], English (US)	"404 Not Found. Sorry the page you requested was not found on this server."	Didn't Load	A page that displays this generic 404 message and no content should be rated Didn't Load .
[SIAD], English (US)	"You are not authorized to view this page. You might not have permission to view this directory or page using the credentials you supplied."	Didn't Load	A page that displays this 403 error message and no content should be rated Didn't Load .
[electionwatch2009.com], English (US)	"Warning – visiting this web site may harm your computer!"	Didn't Load	Pages with warning messages should be rated Didn't Load .
[hat shipping], English (US)	"Website under construction."	Didn't Load	A page that is blank except for the words "Website under construction" should be rated Didn't Load .

In contrast, landing pages with error messages, **but which have content and/or working links**, should be rated according to their utility. Error messages on such pages are usually customized by the webmaster, but sometimes it is hard to tell. The important thing is to look for content and/or working links on the page.

Please note that sometimes **Didn't Load** error messages have links or text that could be mistaken for content, but these links and "content" are from the issuer of the generic message. They are **not** from the webmaster who created the landing page to be rated.

When you assign **Unratable: Didn't Load**, please copy and paste the error message that is displayed on the landing page in the comments section of the rating task.

Choosing a Landing Page Language for pages that do not load

You will choose a landing page language flag for every task you evaluate, even pages that do not load:

- Use the flag that corresponds to your **task language** for pages in your task language.
- Use the flag that corresponds to the appropriate **acceptable language** for pages in an acceptable language.
- Use the **English** flag for pages in English.
- Use the **Foreign Language** flag for pages in a language other than the task language, an acceptable language, or English.
- Use the **None of the above** flag when the page is blank, there is no language on the page, or the page does not load at all.

For a more complete description of the flags used to identify the language of the landing page, please see Section 3.0 of the Rating Guidelines.

4.6.2 Unratable: Foreign Language

Assign **Unratable: Foreign Language** when the page language is not in any of the following: the task language, an acceptable language, or English.

Most of the time, you will use the **Unratable: Foreign Language** rating whenever you choose the **Foreign Language** option for the language of the landing page.

The only time you will **not** use the **Unratable: Foreign Language** rating is when you are rating specific kinds of **Vital** pages. See Section 4.1.5 of the Rating Guidelines for information about rating **Vital** pages.

The **Unratable: Foreign Language** rating is appropriate for all other kinds of queries and all other foreign language pages, even if you personally understand the language on the page and believe you could assign a rating from the rating scale, or even if you can tell that the page is off-topic. When in doubt, please use **Unratable: Foreign Language**.

5.0 Rating: From User Intent to Assigning a Rating

In previous sections of the Rating Guidelines, you read about queries and the rating scale. In this section, we will put it all together. Here are the most important factors to consider when rating: user intent and page utility. This is true of all URL rating tasks, always.

Here are some of the other important ideas in this section:

- You must represent users in your task location. You must rate from a user perspective.
- Some queries have multiple interpretations or user intents. Unlikely interpretations or intents should be given lower ratings.
- Raters are different than users. Results that are helpful for raters are not necessarily helpful for users.
- Location is important. Good pages must be appropriate for the task location.

5.1 User Intent and Page Utility

It is very important to understand user intent. You will rate the landing page based on how well it fits the user intent behind the query. To do this, you may need to use:

- Your experience in the task location with the task language.
- Your common sense.
- Web research.

Hopefully, user intent will be easy to understand for most queries.

If you feel that a page is not helpful for a user, please give the page a low rating. A **Relevant** page must have some utility. A **Slightly Relevant** page has little utility, but is still on the right topic. An **Off-Topic** page has no utility and/or is not on the right topic.

Do not struggle with each rating. Give your best rating and move on. If you are having trouble deciding between two ratings, please use the lower rating. Sometimes, you may even have difficulty choosing among three ratings. When this happens, please use your best judgment.

Finally, although we do not base ratings only on the URL, it is sometimes helpful to look at the URL when rating. Here are the situations where the URL will be helpful:

- For spam identification.
- To notice redirects.
- For identification of some **Vital** pages.

Please remember that you must **ALWAYS** visit the landing page.

5.2 Location is Important

Good search engines return results that are “local,” which means that the results are good for users in their specific location. For example, if an English (US) user searches for [pizza], he is not interested in pizza restaurants in London, England. He wants pizza restaurants in the US. Important: Unless the query indicates otherwise, we will assume that most users want pages from their own location.

In most cases, you will need to lower the rating if the page content is from another country. Do not hesitate to lower the rating to Off-Topic if there is a mismatch between the task location and page that makes the result useless for a user in the task location.

5.3 Language is Important (This section is for Non-English Task Languages)

If your task language is English; for example (English (US), English (UK), English (CA), etc., you may skip this section.

Most of the time, you will use the **Unratable: Foreign Language** rating when the landing page is not in the task language, English, or an acceptable language (please see Section 4.1.5 of the Rating Guidelines for rating foreign **Vital** pages).

Landing pages in the task language are clearly a good choice for users in the task location.

Even though they are not considered foreign, landing pages in English or acceptable languages may not be a good “fit” for users in the task location. For example, in some countries there is a very high rate of English literacy. English pages may be a reasonable fit for locations with a high rate of English literacy, but in other locations where knowledge of English is somewhat rare, English landing pages may *not* be a good fit.

Additionally, some queries seem to “ask for” or “invite” English or acceptable language results, and some do not. When rating pages in English or in an acceptable language, please rate the page based on how helpful you think it is for users. Remember, you should use the **Slightly Relevant** rating for pages which are not very helpful for most users, but are somewhat related to the query.

In some locales, English is one of the official languages or a commonly spoken language. Users living in such locales would not be disappointed to see landing pages in English. For example, the Singapore government recognizes four official languages: English, Malay, Chinese, and Tamil, but English is the first and most dominant language in Singapore.

5.4 Multiple Interpretations

You will rate pages for some queries that have multiple interpretations and multiple user intents.

- In general, pages associated with minor interpretations and unlikely user intents should be rated lower.
- Pages for common interpretations of the query and reasonable user intents should *not* be lowered in rating.
- Only queries with a dominant interpretation can have **Vital** pages.

Here are some examples.

Query Interpretation	Example	Range of Ratings
<p>Dominant Interpretation: Of all the users who type the query, most users would want this interpretation.</p>	<p>[apple], English (US): Apple computers. Most users who type this query want results on Apple computers.</p> <p>[windows], English, (US): the Microsoft operating system. Most users who type this query want results on the Microsoft Windows operating system.</p> <p>[amazon], English (US): the popular website www.amazon.com. Most users who type this query want to go to the Amazon website.</p> <p>[median], English (US): the mathematical formula. Most users who type this query want results about the mathematical formula. Even though this query has a dominant interpretation, no Vital rating is possible since no one can own this query. The highest possible rating for this query is Useful.</p> <p>[guinea pig], English (US): the small furry animal often kept as a pet. Most users who type this query want results about the animal. Even though this query has a dominant interpretation, no Vital rating is possible since no one can own this query. Many webpages have information about guinea pigs. The highest possible rating for this query is Useful.</p>	<p>Vital to Off-Topic or Useless</p>
<p>Common Interpretation: Of all the users who type the query, many or some users would want this interpretation.</p>	<p>[apple], English (US): The fruit. Some users who type this query could want results about the fruit.</p> <p>[windows], English (US): The glass paned windows for a home. Many or some users who type this query could want results about glass windows for a house.</p> <p>[amazon], English (US): The rainforest or river in South America. Some users who type this query could want results about the river or rainforest.</p> <p>[ada], English (US): The American Dental Association, the American Diabetes Association, or the American with Disabilities Act. Many or some users could want information about any of these organizations.</p> <p>[mercury], English, (US): The car brand, the planet, or the chemical element. Many or some users could want information about the car, the planet, or the chemical element.</p> <p>[sandals], English (US): The open type of shoe or the chain of resorts located in the Caribbean Sea. Many or some users could want information about the open type of shoe or the chain of resorts.</p>	<p>Useful to Off-Topic or Useless</p> <p>There can be no Vital page if the interpretation is not dominant.</p>

Query Interpretation	Example	Range of Ratings
Minor Interpretation: Of all the users who type the query, few users would want this interpretation.	<p>[ada], English (US): The Atlanta Development Authority or the American Darters Association. Few users would want information about these interpretations.</p> <p>[mercury], English (US): The Mercury Magazine (published by the Astronomical Society of the Pacific) or Mercury Records (a record label in the U.K). Few users would want information about these interpretations.</p> <p>[hot dog], English (US): "Hot Dog," a movie that was in movie theaters in 1984. Few users would want information about this interpretation.</p>	<p>Relevant to Off-Topic or Useless</p> <p>The less likely you believe the interpretation is, the lower on the scale you should rate the associated result.</p>
"No chance" Interpretation: An interpretation so minor that almost no one would ever want this interpretation.	[guinea pig], English (US): A pig from New Guinea, which is an island country located near Australia. (There probably are pigs in New Guinea, but it is extremely unlikely that the user typing the query would have that interpretation in mind.)	Off-Topic or Useless

Please note that queries with a dominant interpretation *can* have common interpretations as well.

Query	Dominant Interpretation	Common Interpretation
[windows], English (US)	Microsoft operating system	glass windows that you see through
[kayak], English (US)	travel website	small, human-powered boat

In addition to multiple query interpretations, there may be many different possible user intents. Please decide whether a user intent is reasonable or likely. User intents that are less reasonable or less likely should also be lowered on the rating scale.

User Intent	Example	Range of Ratings
Likely user intent: Many or most users have these intents.	<p>[tetris], English (US): Play Tetris (a video game) online, or download the game.</p> <p>[flowers], English (US): Order flowers online, or learn about types of flowers or find pictures of flowers.</p> <p>[credit cards], English (US): Find a credit card company, apply for a card, or compare different brands of credit cards.</p> <p>[amazon], English (US): Go to Amazon.com.</p>	<p>Vital to Off-Topic or Useless</p>
Less likely user intent: Some or few users have these intents.	<p>[tetris], English (US): Research the history of Tetris.</p> <p>[flowers], English (US): Find a definition of the word "flower."</p> <p>[credit cards], English (US): Read an encyclopedia article on the history of credit cards.</p> <p>[amazon], English (US): Read an encyclopedia article about Amazon.com.</p>	<p>Relevant to Off-Topic or Useless</p> <p>Ratings should reflect how many users these pages would help.</p>

5.5 Specificity of Queries and Landing Pages

Some queries are very general and some queries are specific. And other queries are somewhere in between. Here are some examples that compare levels of specificity of English (US) queries:

Query	More Specific Query	Even More Specific Query
[chair]	[dining room chair]	[ikea "henriksdal" highback upholstered chair]
[cameras]	[Nikon cameras]	[Nikon d5000 slr]
[Toyota]	[Toyota hybrid]	[Toyota Prius 2010]
[library]	[Harvard library]	[Harvard Anthropology library]
[interview questions]	[interview questions for teachers]	[practice interview questions used for Teach For America]
[discount stores in houston]	[walmart stores in houston]	[walmart 9555 South Post Oak Road houston]

Good landing pages need to “fit” the specificity of query to be helpful for users who issued the query. When there is a mismatch between the query and the landing page, you will need to think carefully about how helpful the page is for users and rate accordingly.

When there is a mismatch between the query and landing page, assigning a rating can be difficult. You have to think about how helpful a page is for users and base your rating on that.

5.6 Common Rating Problems

Listed below are some common rating mistakes. Most of these mistakes have to do with user intent and the “fit” of the landing page to the query.

5.6.1 Dictionary or Encyclopedia Results

Dictionary or encyclopedia pages are often helpful to raters who are trying to understand the query. They can also sometimes be helpful for the user, but not when the user already understands the words in the query and is looking for something different.

Raters often give high ratings to pages for information user intents even when the query is an action query. For queries that clearly have action intent, information pages should not be rated above **Relevant**. Think about whether users want to know something or do something. Look at the content of the page and decide if the page is helpful for a “know” or “do” intent.

Please respect the “know” intent of product queries. Many people research items online before making a decision about whether to buy the item. Most product queries are “know” and “do” queries.

5.6.2 Queries that Ask for a List

Some queries seem to “ask for a list.” Here are a few principles to help you out when rating these types of queries:

- When the query seems to ask for a list that includes many, many possibilities, individual examples usually are not as helpful as a list.
- When the list of possibilities is short, then individual examples are helpful.
- Sometimes, there are very famous or popular examples on the list. In these cases, the individual famous or popular examples are helpful, even if the list of possibilities is long.

To summarize, if there are few items in the list, then high quality landing pages for individual items are helpful. If there are so many possibilities that any one item seems too specific, lists of results are usually more helpful, unless an individual item is very popular or highly expected.

5.6.3 Misspelled and Mistyped Queries

You will notice that some queries are misspelled or mistyped.

For **obviously** misspelled or mistyped queries, you should base your rating on user intent, not necessarily on exactly how the query has been spelled or typed by the user.

For queries that are **not** obviously misspelled or mistyped, you should assume users are looking for results for the query as it is spelled.

For the query, [federal expres], English (US), it is reasonable to assume that the user is looking for Federal Express at <http://www.fedex.com/us/>. For the query, [my sapce], English (US), it is reasonable to assume the user is looking for MySpace at <http://www.myspace.com/>. There are no other reasonable interpretations for these queries.

Then consider the query [John Stuart], English (US). Even though raters may believe that the user wants to go to pages associated with Jon Stewart, the well-known comedian and host of “The Daily Show” (a popular news satire TV show), we cannot assume that the query has been misspelled. There is a Las Vegas show producer named John Stuart, whose name exactly matches the spelling of the query, and it is very likely that there are “regular” people whose names match the spelling of the query, as well.

Important: Do not assume a query has been misspelled if there is a person or entity that matches the spelling in the query, or even if it is just reasonable that there might be such a person. Sometimes, people exist for whom there are no web results.

5.6.4 URL Queries

Some queries look like URLs. We will call these queries “URL Queries.”

Some URL queries are exact, perfectly-formed, working URLs, such as [www.ibm.com], English (US). Some queries that contain partial URLs, such as [ibm.com], English (US), become working URLs when you add “www.” or “http://” to the front of the URL. We will consider [www.ibm.com], English (US) and [http://www.ibm.com], English (US) to be the same query as [ibm.com], English (US). All of these are considered “URL queries.”

Some queries are website or webpage names, such as [yahoo], English (US) or [yahoo mail], English (US). These queries do not contain “.com” or “www” or other standard components of a URL. These are navigation or “go” queries, but we will not consider them URL queries.

Most queries are neither URL queries nor website/webpage name queries. Most of the time, queries contain terms that do not refer to a particular website or webpage.

Here are some examples of English (US) queries:

URL queries	Website Name/Webpage Name Queries (these are “go” queries, with no “URL parts”)	“Generic” Queries
[ebay.ca] [amazon.com] [people.com] [bbc.co.uk] [www.dealbook.com] [mail.yahoo.com] [news google.com] [tax form 1040 irs.gov] [rei.com]	[ebay] [amazon] [people] [bbc] [dealbook] [yahoo mail] [google news] [irs 1040 tax form official page] [rei kayak page]	[couches] [diabetes] [weight loss] [tax forms] [quilting]

Let's first discuss URL queries. Some URL queries are *not* "working URL" queries. The URLs do not load if you type or paste them into your browser address bar. However, we believe users have a specific page in mind. We will call these "imperfect URL queries." There are many types of imperfect URL queries. Here are descriptions of some of them:

- The query has the same format as a perfect URL query, but the page does not load. Here is an example: [www.UnitedStatesPassportProvider.com], English (US).
- The query has the same format as a perfect "working" URL query, but is obviously misspelled and does not "work." Here are some examples: [www.pizzzzzahut.com] and [www.mcriosoft.com].
- The query has a URL-like format, but contains extra words and/or spaces. Here is an example: [Australian open tennis tournament.com], English (US). We will call this an "imperfect URL query" because it contains "tournament.com" which is part of a URL, but there are spaces in the query.
- The query has a mix of words and URLs, such as [[barbie.com dress up games](#)], English (US).

Some URL queries can be extremely hard to rate. Although you will need to visit the landing page to see and evaluate the content, you will also need to look carefully at the URL of the landing page and the URL in the query. Do not just rate URL queries and results based on the appearance of the URL.

Trying to interpret user intent for imperfect URL queries is hard. It is very easy for users to mistype URLs.

If the query is a perfectly-formed, working URL, please consider *that* URL to be the dominant interpretation. The **Vital** rating should be given when the URL of the page exactly matches the URL in the query. Please note that sometimes the URL of the landing page may contain a longer string than the URL in the query, or look different in other ways. For example, for [imagehosting.us], English (US), both <http://imagehosting.us/> and <http://www.imagehosting.us/index.php?action=show&ident=1478428> should be rated **Appropriate Vital** since the landing page is the same.

If the query is not a perfectly-formed, working URL and/or does not load, please use your judgment to interpret user intent. Do not assign a rating of **Vital** unless there is little or no doubt that the page matches user intent.

Now let's talk about "website name" or "webpage name" queries, which are *not* URL queries. They are queries which contain the names of websites or webpages, and the *dominant* interpretation of the query is the website or webpage. Some website name queries have other meanings, besides the website.

Website or Webpage Query	Explanation
[kayak], English (US)	Users could be looking for a kayak (a type of boat), but Kayak is a very popular travel website. The website kayak.com is the dominant interpretation.
[youtube], English (US)	YouTube is one of the most popular websites on the Web.
[ebay], English (US)	eBay is one of the most popular websites on the Web.
[webmd], English (US)	WebMD is a very popular medical information website.
[twitter], English (US)	Twitter is a very popular website.
[cafepress], English (US)	Cafepress is a website where users can buy t-shirts and other gifts and even have them custom-made.
[addicting games], English (US)	AddictingGames is a very popular game website.
[rei kayak page], English (US)	Users want to go to the "kayak" page on the REI website.

Here are some examples of queries which are **not** website queries and are **not** URL queries. Website names exist that match these queries, but those websites are probably not what users have in mind. These queries do not have **Vital** pages.

Generic Query	Explanation
[birdcages], English (US)	Users are probably interested in researching or buying a birdcage. This is a generic query. There is no Vital page. There is a store with the URL birdcages.com, but many stores sell birdcages.
[kamasutra], English (US)	Users are probably interested in learning about the Kama Sutra or reading the Kama Sutra text. There is no Vital page. There is a store with the URL kamasutra.com, but that probably is not the dominant interpretation of this query.

Generic Query	Explanation
[weightloss], English (US)	Users are looking for weight loss information, and there are many good authoritative pages with weight loss information. There is a website weightloss.com , which has helpful, common sense information about losing weight, but users probably are not trying to go to that page.
[couches], English (US)	Users are interested in researching or buying a couch. There are many good websites that sell couches. There is a website couches.com , but there is nothing in the query that indicates users want to go to couches.com .

Keep in mind that just about any query can be turned into a URL by adding “.com” to the query terms. However, without the “.com” included in the query, you should not assume the query is a website name.

In other words, just because the query is [couches] does not mean that the result <http://www.couches.com> is what the user wants. Please be careful with “generic” queries. A commonly used spam technique is to create websites with generic names.

When users issue URL queries, the intent is to go to a specific page. That page should be rated **Vital**. It can be very hard to rate “non-Vital” pages for URL queries. Sometimes, the **Vital** page is the only helpful result for a URL query. But sometimes, other pages are helpful as well. Ratings for pages with information about the queried website can range from **Off-Topic or Useless** to **Useful**.

5.6.5 New and Old Pages

Information or “know” queries may be about recent or past events. The landing page should be rated based on fit to the informational need of the query. Some queries demand very recent results. Most of the time, you need to consider the content of the page rather than the date on the page.

For some queries, timeliness is very important. Queries for recent events and recurring events need pages with recent content. We assume that users who type queries looking for results from an election, sporting event, or other type of annual competition are looking for the most recent results, not results from previous years.

Please note, however, that, depending on when annual events occur, the most helpful pages may be for the past event or the current/upcoming event. If the event took place several months ago, the most helpful pages would probably be about the past event. If the event will take place in a few months, the most helpful pages would probably be about the upcoming event. You will have to use your judgment.

If the landing page appears to be the official page of the event, it should get a **Vital** rating, whether the content is about the past or upcoming event.

Information queries may need recent results as well. For example, if the query is [population of Paris], English (US), users are looking for the most current population numbers.

On the other hand, if the query is [population of France in 1813], the issue is not how “new” or “recent” the page is, but whether it has the information requested. Sometimes “old” pages are the only good source of information about past events. “Old” pages are not necessarily “outdated” or bad. It depends on the query and the page content.

If the landing page you are given to rate is a search engine page with an empty search box and no results displayed, then the page has no connection to the query and should get a rating of **Off-Topic or Useless**.

If the landing page is a set of results from a search engine, the page could be very helpful to users. Depending on how helpful the page would be, ratings can range from **Useful** to **Off-Topic or Useless**.

6.0 Flags

In addition to assigning a rating from the rating scale, you will also assign flags to mark special types of pages.

6.1 Spam Flag

You must decide if the page is should be assigned a **Spam** flag by looking for spam signals that you will learn about in the Webspam Guidelines (Part 4 of the General Guidelines).

Not Spam: If you do not believe that a page has been designed using deceptive web design techniques, you should assign a **Not Spam** flag.

Maybe Spam: If you find a page to be “spammy,” but you do not feel comfortable saying that the webmaster definitely designed the page using deceptive web design techniques, you should assign a **Maybe Spam** flag.

Spam: If you believe that a page has been designed using the deceptive web design techniques described in the Webspam Guidelines,” you should assign a **Spam** flag.

If you choose either **Maybe Spam** or **Spam**, you must include a comment explaining why.

6.2 Pornography Flag

Please apply the **Porn** flag to all porn pages. A page will be considered porn if it has pornographic content, including porn images, links, text, pop-ups, and/or ads. An image may be considered porn in one culture or country, but not another. Please use your judgment and knowledge of the task location.

6.2.1 Clear Porn Intent

If the user intent behind a query is clearly **not** pornographic, a landing page that serves a pornographic interpretation of the query and/or has porn for its main content should be rated **Off-Topic or Useless** and assigned a **Porn** flag. For example, consider the query [car pictures]. In any task language, a page showing a nude female reclining on the hood of a car should be rated **Off-Topic or Useless** and assigned a **Porn** flag, even though there is a car in the picture.

The reasons we are asking you to do this are the following:

- When the user intent is clearly not porn, a porn result that serves a “no chance” interpretation or has porn as its main content should be considered to have no utility.
- Uninvited porn is a very bad experience for many users and is an indication of poor search engine quality.

6.2.2 Possible Porn Intent

Some queries have both non-porn and porn interpretations. For example, all of the following English (US) queries are possible porn intent queries, but they also have a non-porn intent: [girls], [gay], [thong], [breast], [sex], [spanking]. We will call these queries “possible porn intent” queries.

For these queries, please assume that the non-porn interpretation is dominant, even if you think users are looking for porn. For example, please assume that the dominant interpretation of [spanking], English (US) is the discipline technique used by parents on a child (the non-porn interpretation). Rate the porn interpretation as a minor interpretation, even if you think most users are looking for porn.

6.2.3 Clear Porn Intent

For very clear porn queries where no other intent is possible, assign a rating to the porn landing page using the rating scale without lowering the score. Even though there is porn intent, the page should still be assigned a **Porn** flag.

Please note that you should not simply rate all porn pages for porn queries as **Relevant** or **Useful**. Even though the query is porn and the result is porn, the page must fit the query to have utility and get a high rating.

Pages that provide a poor user experience - such as pages that try to download malicious software - should also receive low ratings, even if they have some images appropriate to the query.

Porn stars, porn movies, names of specific porn websites, etc., can have **Vital** pages. Be consistent in assigning a **Porn** flag to all porn pages, even when the rating is **Vital**.

Please do not assign a **Porn** flag to a non-porn page, just because the query has porn intent. If the landing page is not porn, it should not be flagged.

6.3 Malicious Flag

A page should be assigned a **Malicious** flag if:

- You are forced to quit your browser due to prompts that keep coming back and will not go away.
- There are attempts to download spyware, Trojans, viruses, etc.

Please note that pop-ups that you are able to close are not malicious, even if it takes a couple of tries to get rid of them.

Please do not assign a **Malicious** flag just because the browser gives you a warning message or certificate acceptance request. Assign a **Malicious** flag only under the conditions listed above. If you encounter a page with a warning message, such as “Warning-visiting this web site may harm your computer,” or if your antivirus software warns you about a page, you should not try to visit the page to assign a rating. You should instead assign a rating of **Unratable: Didn’t Load**.

6.4 Compatibility between Ratings and Flags

Please be aware that **Unratable** pages can be assigned **Spam**, **Porn**, and/or **Malicious** flags. Here are some examples:

- The page is in a foreign language, but has porn images.
- The page is in a foreign language, but there is hidden text.
- The page does not load, but you can tell from the URL that it is a sneaky redirect. You will learn about sneaky redirects in the Webspam Guidelines (Part 4 of the General Guidelines).
- The page does not load, but has porn ads.
- The page is in a foreign language, but you cannot close a pop-up on the page and you are forced to quit your browser.

Part 2: URL Rating Tasks with Query Locations

1.0 Query Locations

All URL rating tasks have a task location, which is usually the country location.

Some URL rating tasks also have a “query location,” which is associated with the geographic location of the user when he or she issued the query. The query location may be a zip code, town, city, city and state, etc. Usually, the query location is automatically detected by the search engine, but may come from the user’s stated preferences.

For narrowly defined query locations, such as specific zip codes or towns, the relevant location may extend beyond the specified zip code or town boundaries. Remember that real users are sometimes looking for the nearest stores or restaurants. If those happen to be outside the specified location, that may be acceptable to the user. You will have to use your judgment about what is reasonable.

Here are some important things to know about tasks with query locations:

- You will rate from the perspective of someone living in the query location.
- Local pages (pages associated with the query location) that are helpful should receive high ratings.
- Pages that would be helpful to users in any query location should also receive high ratings.
- When the query is an entity, such as a business, organization, school, etc., and the entity has both an official homepage and official location-specific webpages, a rating of **Appropriate Vital** will apply to both the entity’s homepage and the appropriate query location-specific webpage.

Important: Sometimes, users specify a location when they type a query. For example, in the query [pizza hut, Marietta, Georgia], the user has specified “Marietta, Georgia” as the location of interest. Some tasks have both a Query Location and a location specified in the query. When this happens, you should rate with respect to the location specified in the query, rather than the Query Location.

Here are examples of three types of tasks:

- The task has a location specified in the query.
- The task has a Query Location.
- The task has both a Query Location and a location specified in the query.

Task Type	Screenshot	Description												
<p>This is <i>not</i> a location-specific task because it does <i>not</i> have a Query Location.</p> <p>Notice, however, that a location is specified in the query.</p>	<table><tr><td>Query</td><td>pizza hut san francisco</td></tr><tr><td>URL</td><td>http://www.yelp.com/biz/pizza-hut-san-francisco</td></tr><tr><td>Task Location</td><td>United States (US)</td></tr><tr><td>Task Language</td><td>English</td></tr><tr><td>Other Acceptable Languages</td><td><i>None</i></td></tr></table>	Query	pizza hut san francisco	URL	http://www.yelp.com/biz/pizza-hut-san-francisco	Task Location	United States (US)	Task Language	English	Other Acceptable Languages	<i>None</i>	<p>The user wants Pizza Hut information for the San Francisco area.</p>		
Query	pizza hut san francisco													
URL	http://www.yelp.com/biz/pizza-hut-san-francisco													
Task Location	United States (US)													
Task Language	English													
Other Acceptable Languages	<i>None</i>													
<p>This is a location-specific task because it has a Query Location.</p>	<table><tr><td>Query</td><td>pizza hut</td></tr><tr><td>Query Location</td><td>***** San Francisco *****</td></tr><tr><td>URL</td><td>http://www.yelp.com/biz/pizza-hut-san-francisco</td></tr><tr><td>Task Location</td><td>United States (US)</td></tr><tr><td>Task Language</td><td>English</td></tr><tr><td>Other Acceptable Languages</td><td><i>None</i></td></tr></table>	Query	pizza hut	Query Location	***** San Francisco *****	URL	http://www.yelp.com/biz/pizza-hut-san-francisco	Task Location	United States (US)	Task Language	English	Other Acceptable Languages	<i>None</i>	<p>The query was issued by a user living in San Francisco.</p> <p>We can assume that the user is looking for a Pizza Hut restaurant in San Francisco.</p>
Query	pizza hut													
Query Location	***** San Francisco *****													
URL	http://www.yelp.com/biz/pizza-hut-san-francisco													
Task Location	United States (US)													
Task Language	English													
Other Acceptable Languages	<i>None</i>													
<p>This is also a location-specific task because it has a Query Location.</p> <p>Notice, however, that a location is specified in the query.</p>	<table><tr><td>Query</td><td>pizza hut san francisco</td></tr><tr><td>Query Location</td><td>***** New York *****</td></tr><tr><td>URL</td><td>http://www.yelp.com/biz/pizza-hut-san-francisco</td></tr><tr><td>Task Location</td><td>United States (US)</td></tr><tr><td>Task Language</td><td>English</td></tr><tr><td>Other Acceptable Languages</td><td><i>None</i></td></tr></table>	Query	pizza hut san francisco	Query Location	***** New York *****	URL	http://www.yelp.com/biz/pizza-hut-san-francisco	Task Location	United States (US)	Task Language	English	Other Acceptable Languages	<i>None</i>	<p>The query was issued by a user living in New York.</p> <p>However, because the query contains “san francisco,” we know that the user is looking for Pizza Hut restaurants in the San Francisco area, even though the Query Location is New York.</p>
Query	pizza hut san francisco													
Query Location	***** New York *****													
URL	http://www.yelp.com/biz/pizza-hut-san-francisco													
Task Location	United States (US)													
Task Language	English													
Other Acceptable Languages	<i>None</i>													

2.0 Location-Specific Rating Task Screenshot

The Location-Specific URL rating task page is similar to the standard URL Rating task page, except that it displays additional information associated with the Query Location.

Information	Standard URL Rating Task Page	Location-Specific URL Rating Task Page
Query Location	Standard URL Rating task home does not have this information.	***** New York ***** ***** 90210 ***** ***** Dallas, TX ***** ***** TX *****

Location-Specific URL Rating Task Page

[rater homepage](#) → rating task [johndoe@gmail.com](#) [[rater homepage](#) · [recently completed tasks](#) · [logout](#)]
Language: **English (US)**

Rating Task - icq

[search results: [google](#)] · [release task](#)

Query	icq
Query Location	***** San Francisco, CA *****
Query Description	This is a location-specific rating task for the Query Location described above.
URL	http://www.mobicro.info/
Task Location	United States (US)
Task Language	English
Other Acceptable Languages	None

3.0 Assigning a Rating When There is a Query Location

In some tasks, the query location will be an important consideration in the rating you assign. For example:

Query: [IHOP restaurants], English (US)

Query Location: Boston, MA

The query location is an important consideration. Users in Boston who type this query are interested in IHOP restaurants in the Boston area, not other locations.

However, in many tasks the query is *not* associated with a specific location and the query location will not be a consideration at all. The rating you assign will be the same rating you would have assigned if the task did not have a query location. For example:

Query: [amazon.com], English (US)

Query Location: Boston, MA

The query location is not a consideration at all. Amazon.com is a website that is not associated with a specific location.

The query location makes a difference when the landing page would be more helpful to users in some locations than users in other locations.

3.1 When Does the Query Location Matter?

Here are some examples that demonstrate when the query location matters and when it does not.

Query	Query Location	URL	Likely User Intent	Does the Query Location Matter in this Example?	Explanation
[facebook], English (US)	Birmingham, AL	http://www.facebook.com/	The user in Birmingham, Alabama wants to go to the Facebook website at www.facebook.com	No , because Facebook is a website that is not associated with a specific location.	The landing page is equally helpful to users in Birmingham, Alabama and other locations. It should be rated Appropriate Vital for any query location, or if there is no query location specified in the task.
[Benihana], English (US)	New York, NY	http://www.benihana.com/	The user in New York City wants information about the Benihana restaurant in New York City or to go to the Benihana homepage.	No , because the homepage of the entity should get an Appropriate Vital rating, even if a location-specific webpage exists.	The official Benihana homepage should be rated Appropriate Vital for New York City or any other query location, or if there is no query location specified in the task.
[Benihana], English (US)	New York, NY	http://www.benihana.com/location/newyorkwest-nyc/	The user in New York City wants information about the Benihana restaurant in New York City or to go to the Benihana homepage.	Yes , because users in New York City are interested in Benihana restaurants in New York City, not other locations.	The landing page is the official webpage for the Benihana restaurant located in New York City. It should be rated Appropriate Vital for the query location. However, it would be rated Other Vital for other query locations or Slightly Relevant if there is no query location specified in the task.

Query	Query Location	URL	Likely User Intent	Does the Query Location Matter in this Example?	Explanation
[ice rink], English (US)	College Station, TX	http://www.arcticwolfice.com/	The user in College Station, Texas wants information about local ice rinks.	Yes , because users in College Station are interested in ice rinks in College Station, not other locations.	The landing page is the official homepage of Arctic Wolf Ice Center, the only ice rink in College Station and therefore the dominant interpretation for this query location. It should be rated Appropriate Vital for the query location. However, it should be rated Off-Topic or Useless for other query locations or Slightly Relevant if there is no query location specified in the task.
[weather conditions], English (US)	Las Vegas, NV	http://www.wunderground.com/US/NV/Las_Vegas.html	The user in Las Vegas, Nevada wants information about local weather conditions.	Yes , because users in Las Vegas are probably interested in the weather in Las Vegas, not other locations.	The landing page has information about the current weather conditions in Las Vegas. It should be rated Useful for the query location. However, it should be rated Off-Topic or Useless for other query locations or Slightly Relevant if there is no query location specified in the task.
[patriots], English (US)	Concord, MA	http://www.patriots.com/	The dominant interpretation of this query for the user in Concord, Massachusetts is the New England Patriots football team. The user wants information about the team or to go to the team's official homepage.	Yes , because the New England Patriots football team is very popular with users in New England (where Concord, Massachusetts is located). It is highly likely that users in Concord issuing this query have this football team in mind.	The landing page is the football team's official homepage. It should be rated Appropriate Vital for the query location. However, it should be rated Useful for query locations outside New England, or if there is no query location specified in the task, because it is a common interpretation of the query.

Part 3: Webspam Guidelines

1.0 What is Webspam ?

Webspam is the term for webpages that are designed by webmasters to trick search engines and draw users to their websites. In these guidelines, we sometimes refer to webspam as “spam,” and webmasters who use deceptive techniques as “spammers.”

In the coming pages, you will learn how to identify some of these deceptive techniques. When you see them being used, you will assign a **Spam** flag. Please note that pages that are merely annoying, junky, or low quality, such as pages with lots of pop-ups or ads, are not necessarily spam.

1.1 The Relationship between Ratings and Spam

In the Rating Guidelines, Part 1 of the General Guidelines, you learned that landing pages are rated according to their utility to users for a particular query. You would not be able to assign a rating to a page without knowing the query.

Spam flags do **not** depend on a relationship between the query and the landing page. A page should get a **Spam** flag if it is created using deceptive techniques - no matter what the query is or how helpful the page might be.

Some spam pages are very low quality and have little or no content which would be helpful for users. These pages will usually be assigned a low rating, either **Slightly Relevant** or **Off-Topic or Useless**, in addition to the **Spam** flag.

Other spam pages, which are not as low quality and have some helpful content, may be assigned a rating of **Slightly Relevant** or **Relevant**.

In some specific cases, it is also possible for a page to receive a **Vital** rating, and also be assigned a **Spam** flag. For example, if there is a sneaky redirect and the landing page is the target of the query, the page will get a **Vital** rating and a **Spam** flag. You will learn about “sneaky redirect” spam in Section 2.3 of the Webspam Guidelines.

1.2 Why do Spammers Create Spam Pages?

Spammers typically create spam pages to make money. Sometimes, they make money directly, by placing moneymaking links on the spam page.

There are two primary types of moneymaking links:

- **Pay-Per-Click (PPC)** ads
- **Affiliates** (introduced in Section 4.1 of the Webspam Guidelines)

When hosting PPC ads on a webpage, each time ads are clicked, the site owner gets paid. Another term for PPC ads is “sponsored links.” PPC ads appear on many, many webpages. Some pages with PPC ads are spam, but many pages with PPC ads are not. Pages should not be assigned a **Spam** flag if they are created to provide substantive information or help to users. Pages are spam if they exist primarily to make money and not help users.

Sometimes, spam pages do not have moneymaking links. These spam pages are created to change search engine rankings or even to do harm to users’ computers with sneaky downloads. They are spam because they use deceptive techniques, even though you are unable to see how they are making money.

1.3 When to Check for Spam

There are some pages, such as the main page of a well-known website (e.g. <http://www.apple.com>), that you may feel do not need to be evaluated for spam. However, even webmasters for highly reputable websites occasionally use deceptive techniques. Therefore, we ask that you use the following two quick and easy spam detection techniques on **all** webpages that you evaluate.

- Apply “Ctrl-A” (or apply “⌘” and “A” for Apple computer users) to the landing page to look for hidden text. You will learn about using “Ctrl-A” in Section 2.1.1 of the Webspam Guidelines.
- Scroll all the way down and to the right on the page to look for hidden text on areas of the page outside the normal viewing area. You will learn more about hidden text outside the normal viewing area in Section 2.1.5 of the Webspam Guidelines.

2.0 Looking for Technical Signals

When evaluating a page for spam, you should start by looking for the following “technical signals”:

- Hidden text and hidden links.
- Keyword stuffing.
- Sneaky redirects.
- Cloaking with JavaScript redirects and 100% frames.

This section describes these technical signals and provides tips and tools on how to identify them.

2.1 Hidden Text and Hidden Links

Webmasters add hidden text and/or hidden links to lure search engines and users to their pages. Hidden text is visible to the search engine, but not to the user, who might find it distracting or annoying. Here are some things you should know about hidden text:

- It may be completely invisible to the human eye.
- It may be in the same color as the background color on the page, or in a color that is so close to the background color that it almost invisible and will not be noticed.
- It may be formatted in a very, very small font size (e.g., 1-point) so that it will not be noticed.
- It may be placed outside the normal viewing area. For example, there may be a large blank space between the normal viewing area and a “hidden” area of text all the way at the bottom of the page or far to the right.
- Sometimes there is just a line or two of hidden text, but you may even see a whole page of it.
- Most hidden text is there to trick the search engine, but occasionally you will find hidden text that is not spam. For example, if the webmaster merely hides the date of an update, it is not spam.

Hidden text may be revealed by:

- Applying Ctrl-A (or “⌘” and “A” for Apple computer users).
- Disabling CSS.
- Disabling JavaScript.
- Viewing the source code.
- Looking outside the normal viewing area.

2.1.1 Apply Ctrl-A to the Landing Page

After you have clicked on the URL, simultaneously press the “**Ctrl**” and “**A**” keys (the keyboard shortcut for “Select All” for PC users), or “**⌘**” and “**A**” or “**Command**” and “**A**” (the keyboard shortcuts for Apple computer users) and then scroll down the whole page. This technique sometimes reveals text that has been hidden.

Tiny text is not always exposed using Ctrl-A. You should be suspicious of horizontal lines or bars on the page because sometimes they contain hidden text. A simple technique for revealing this type of hidden text is to select and copy the suspicious line or bar, paste it in your word processor, and increase the font size. You may also try using the techniques described below.

2.1.2 Disable CSS

Disabling CSS sometimes reveals hidden text. You do not need to check every page for hidden text in CSS, but please do check if the page is suspicious.

2.1.3 Disable JavaScript

Spammers sometimes use JavaScript to hide text.

Important: When you are done looking for spam on a particular page, please remember to go back and enable JavaScript. If you do not do this, certain features on pages you open will not work.

2.1.4 View the Source Code

Viewing the source code sometimes reveals hidden text. Look for large areas of keyword stuffing in the source code. Keyword stuffing is discussed in Section 2.2 of the Webspam Guidelines.

Please note that a **Spam** flag should *not* be assigned when the keyword stuffing appears in the meta tags only. Meta tags are easy to identify because they start with the words “meta name.” Here is an example:

2.1.5 Look Outside the Normal Viewing Area

Be suspicious of large blank areas on the bottom and far right portions of the page. Use the vertical and horizontal scroll bars to see if appears there is text on the portion(s) of the page outside the main viewing area.

2.2 Keyword Stuffing

Keyword Stuffing: Webmasters sometimes load pages with an excessive amount of keywords. Here are descriptions of what you might see:

- Keywords repeated many times on the page.
- Words that are related to keywords repeated many times on the page.
- Multiple misspellings of keywords on the page.
- Pages with a large amount of what look like gibberish or random keywords.
- Pages that appear to be programmatically or automatically generated text that doesn’t really make sense.

Webmasters also sometimes load pages with *irrelevant* keywords on topics that are unrelated to the query, such as mortgages, cell phones, ringtones, gambling, weather, etc. Whether the keywords are related or unrelated to the query, the intent is to draw search engines and users to the page.

It is sometimes difficult to decide when the keywords on a page should be considered keyword stuffing. We ask you to assign a **Spam** flag if you think the number of keywords on the page is excessive and would be annoying and

distracting to the real user. If you do not feel the number of keywords would bother the user, please do not assign a **Spam** flag.

Please note: Hidden text and keyword stuffing often go together. Hidden text frequently contains keyword stuffing.

Recognizing keyword stuffing

Some keyword stuffing is visible to the human eye and you will not have to use any special techniques to see it. In other cases, it is hidden. You will discover *hidden* keyword stuffing by using the techniques in Section 2.1 of the Webspam Guidelines. **Important:** *hidden* keyword stuffing will **always** be considered spam (unless it is only in the source code meta tags).

2.2.1 Keyword Stuffing in the URL

URLs may also contain keyword stuffing. These URLs are computer-generated based on the words in the query and are often formatted with many hyphens (dashes) in them. They are a strong spam signal.

Here are some examples of keyword stuffing in the URL. We have removed the hyperlinks from these examples because some of them have stopped working and others have become malicious. You do not need to click through to the landing page in order to see that there is keyword stuffing in the URL and that they are spam.

- <http://frat-boy-blog-gay.grandbrooklynlodge.cn/boy-brief-frat-in-their-wet.html>
- <http://brazilian-model-alexandra.wantloweryour.cn/brazilian-model-adriana-lima.html>

2.3 Sneaky Redirects

Sneaky Redirects: We call it a sneaky redirect when a page redirects the user from a URL on one domain to a different URL on a different domain, *with spam intent*. Search engines “see” the first page, while the user is sent to a different page and sees different content. Here are some other things you should know about sneaky redirects:

- While being redirected, you may notice that the page redirects through several URLs before ending up on the landing page.
- Sneaky redirects may take the user to one of several rotating domains; so clicking on the same URL several times may send you to different landing pages each time.
- Some sneaky redirects take users to well-known merchant websites, such as Amazon, eBay, Zappos, etc.

Recognizing sneaky redirects

- **Compare the two URLs:** Compare the URL in the rating task to the URL of the landing page to see if it makes sense that one would redirect to the other. A redirect from a company’s old homepage to its new homepage on a different domain is not sneaky. Redirects from one page on a domain to another page on the same domain are also not sneaky.
- **Look at the domain registrants:** If you suspect that a sneaky redirect has taken place, you should check to see “whois” the registrant (or owner) of the two domains. If the registrant is the same, the redirect is less likely to be sneaky. Please see Section 2.3.1 of the Webspam Guidelines for instructions on checking “whois.”

2.3.1 Using “Whois”

Here are instructions for checking “whois” the domain registrant:

1. Go to the site of a “whois” provider. Here are two you can use: <http://www.domaintools.com/> and <http://whois.mtgsy.net/default.php>. Some computers also allow you to run a command like `whois example.com` from a terminal window.
2. Enter the URL of one domain in the search box on the “whois” page. Sometimes, you will need to delete some leading or following characters. For example, if the URL is <http://supportaj.dell.com/support/>, you will enter just “dell.com” in the search box of the whois provider.
3. Open another “whois” page.
4. Enter the URL of the other domain in the search box on the second “whois” page.
5. Compare the domain registrants for the two URLs. If you find that they have the same domain domain registrant, you will **typically** conclude that the page is not spam. If they are different and do not seem related, it is probably spam.

Please be aware that domains with the same domain registrant can look very different. For example, Barnes and Noble, the bookseller, owns the following domains: www.barnesandnoble.com, www.bn.com, and www.books.com.

2.4 Cloaking

It is called “cloaking” when the webmaster shows different pages to the search engine and the user. True cloaking is somewhat rare, but spammers do use other methods to show different pages to search engines than to users. Two such techniques used by spammers are:

- JavaScript redirects
- 100% frame

2.4.1 JavaScript Redirects

Spammers use JavaScript redirects to show one page to search engines while sending users to a different page. Looking at the page first with JavaScript enabled and then with JavaScript disabled reveals the differences.

2.4.2 100% Frame

Webmasters sometimes cloak what users see by using frames. Two frames (pages) exist, but one frame takes up 100% of the screen. The user sees one frame (page), but the search engine sees both frames.

3.0 Helpful Webpages vs. Spam Webpages

Search engines want to display webpages that are helpful to users. In this section, you will learn how to determine if pages with ads on them are spam, or if they have utility to the user. We will talk about:

- Pages with PPC ads and other content, which *are designed to help users in a way that adds substantial value*.
- Pages with PPC ads and other content, which *primarily exist to make money without adding much value*.

Some pages contain PPC ads only, or have very, very little on them besides the PPC ads. We refer to these pages as “pure PPC” pages. You will learn more about pure PPC pages in Section 4.2 of the Webspam Guidelines. When the page containing PPC ads adds substantial value for users, it is not spam. Here are examples of content that can be helpful to users:

- **Price comparison functionality:** Some webpages offer price comparisons for shoppers looking to make a purchase. The shopper then has ability to take price into consideration. Even if the user has to click an affiliate link to go to another site to place the order, it can be helpful to have price comparisons on the page.
- **Product reviews:** Some pages provide *original* product reviews that are helpful to the user in deciding whether to make a purchase. Items that are commonly reviewed are books, electronics, and hotels.
- **Recipes:** Some pages provide recipes. If the recipes on the page are helpful, for example, if the recipes are original or the page includes reviews of original or non-original recipes, the page is not spam.
- **Lyrics, quotes, proverbs, poems, etc.:** Some pages display this type of content. If the page is designed to help users find song lyrics or poems, etc., it is not spam.
- **Contact information:** Some pages provide contact information for companies. If the contact information includes physical addresses, phone numbers, maps, etc., the page is helpful and not spam.
- **Coupon, discount, and promotion codes:** Some affiliate pages provide coupon, promotion, or discount codes for the consumer, in addition to a link to the merchant. Since these types of codes are helpful to the user, they provide added value.

Please note that recipes, lyrics, quotes, poems, etc. do not usually have authoritative pages. Anyone can obtain and put this content on webpages.

3.1 Pages with Copied Content and PPC Ads

Copied content refers to content that has been copied from other sources. Webmasters sometimes use special “scraper” software to search the Web for content to put on their websites that is related to specific keywords. Content can also be taken from another website using the simple “copy and paste” method.

3.1.1 Copied Text and PPC Ads

Content that has been copied from sources such as Wikipedia (<http://www.wikipedia.org/>) and the Open Directory Project (<http://www.dmoz.org/>), sites that allow the distribution of their content and may even encourage it, is still considered to be copied content.

Copying content from such sources is not necessarily illegal, nor is it plagiarism. Webmasters who copy content usually do not claim to be original content creators and may, in fact, assign credit to the originator of the content. However, even if they do give credit to others, it is considered to be copied content.

These copies are often old, not updated, and may not be trustworthy. Users want information they can trust. A copy of a Wikipedia article on an unknown website accompanied by ads offers little utility to users. We will call a page spam if it has copied content and is created to make money from ads on the page.

3.1.2 Feeds and PPC Ads

Web publishers (such as the BBC, CNN, Usenet, CNet, NYTimes, Atom, and others) publish information online that is readily available to users through RSS (Really Simple Syndication) and XML (Extensible Markup Language) feeds. Companies, such as Searchfeed.com, provide feeds of PPC ads and links to most qualifying webmasters. A page that just contains freely available feeds and PPC ads, and was created just to make money, is spam.

3.1.3 Doorway Pages

Doorway pages are sets of pages that have been created for search engines to deliver the user to a common destination page. The pages all look very much the same and do not provide meaningful content for users. Here is an example: <http://www.hair-removal-hair-laser.com/>. This page contains links for all of the states in the US. Clicking on a link makes you think that you are getting a customized page for that state, but if you click on another link, you will find that every page is really the same. These pages are spam.

3.1.4 Templates and Other Computer-Generated Pages

Some websites use templates to mass-reproduce webpages automatically. The content is usually copied from sources that provide such content. You will learn to recognize templates, which usually follow a generic format or pattern. Look for slight keyword variations that suggest automated use of a keyword suggestion tool. If the keyword is “mortgage,” you may see words such as “mortgages,” “mortgage loan,” “mortgages loans,” etc. in the title, snippets, and/or URL.

These spam pages contain links to other pages that usually contain some combination of copied content, PPC ads, and other spam links. Clicking on links on these pages will land you on other pages on the same domain with similar content and links.

3.1.5 Copied Message Boards

Sometimes you will see copied message boards (user forums, or copies of Usenet posts) and ads. When the page contains only the copied message board and PPC ads, the page is spam.

3.1.6 Recognizing Copied Content

Here are some things you can do to help you recognize copied content:

- **Search for an exact sentence from the text on the page:** Copy and paste a distinctive sentence in the search box of a search engine. When you paste the sentence in the search box, put quotation marks around it so that the search engine will search for the exact string of words. You may want to pick a series of regular words without punctuation or special characters, just to make sure that your search isn't complicated by the punctuation characters. From the search results displayed, you may find where the content originated. If the content is original and has not been copied from another source, it probably was written to be helpful to users.
- **Look for PPC ads surrounding the content.** Wikipedia and DMOZ do not display ads. If you see Wikipedia or DMOZ content and PPC ads with no original content on the page, it is spam.
- **Become familiar with the format of Wikipedia and DMOZ pages:** The section headings and links on Wikipedia pages usually follow the same format. DMOZ pages use a directory pathway that is easy to recognize. In addition, DMOZ pages have these links: “submit a site” and “become an editor,” which also often appear on copied pages.

3.1.7 Other Tips for Recognizing Spam Pages

Here are some additional ways to recognize spam pages:

- **Look at URL formatting:** Look for URL formatting that suggests that a template or other automation was used to create it. Often, you will see keywords contained in the URL, separated by hyphens. Here is an example: <http://nzealand.co.nz/blog/thelawmail/2007/12/29/com-search-extreme-belladonna-users-search-expired-domain-names-search-expired-domains/>.
- **Look to see if the page appears to have been created to help users:** Look for features, such as lyrics, recipes, quotes, contact information, phone numbers, physical addresses, original reviews, a working comment box, etc.
- **Look for suspicious, computer-generated grammar:** Look at the text on the page. When it is computer-generated, it often looks like “gibberish,” which means that it does not make sense. You may also see hyperlinked keywords inside the text. Think about whether it seems as if the page was created by a human or by a machine: pages created by machines are usually not designed to be helpful for users and are usually spam.

3.2 Fake Search Pages with PPC Ads

A fake search page is a page with a list of links that looks like a page of search results. You will see a “search box” on the page, but if you submit a new query in the search box, you just get a different page of links. If you click on a few of the links, you will see that the page is just a collection of PPC links disguised as search engine results.

3.3 Fake Blogs with PPC Ads

A fake blog contains fake blog entries that are either nonsensical or copied from another source. Fake blogs often contain keyword stuffing, which is described in Section 2.2 of the Webspam Guidelines. The page exists so that the PPC links on the page will be clicked. PPC links may appear within the text of the fake blog entry, or on other parts of the page. Fake blogs may appear to allow the user to post a comment, but the feature does not work. Fake blogs are spam.

Spammed Blogs: Spammed blogs are different from fake blogs. A spammed blog is a real working blog with real blog entries, but has been spammed with entries that contain PPC ads and/or porn links. We do not want to penalize a blog because someone else has put spam on it. If you believe that the blog is a good, legitimate blog that has been spammed by someone else, please do **not** assign a **Spam** flag.

3.4 Fake Message Boards with PPC Ads

A fake message board is similar to a fake blog. It contains what appear to be “messages,” but are not. The text in the message may be nonsensical or it may contain PPC links. Fake message boards may appear to have comment, registration, and login sections, but either these features do not work at all, or you are redirected back to the same page. On real message boards, you will see responses to posts. On fake message boards, either there are no responses, or the responses themselves are spam.

Copied Message Boards with PPC Ads: You may also find entire message boards that have been copied. If you suspect this has happened, copy and search for a snippet of text. Copied message boards are spam.

Spammed Message Boards: Spammed message boards are different from fake message boards. A spammed message board is a real message board with real posts and real responses, but which posts with PPC ads and/or porn links have spammed. We do not want to penalize a message board because someone has put spam posts up on it. If you believe the message board is a good, legitimate message board that has been spammed, please do **not** assign a **Spam** flag.

3.5 Copied Content that is NOT Spam

Some copied content is not spam. Here are some examples: lyrics, poems, proverbs, quotes, etc. This type of content has no unique or central authority.

If the page you are evaluating appears to be from a legitimate lyrics, poetry, etc. website, do not assign a **Spam** flag. If you think the page exists primarily to make money, you should assign a **Spam** flag.

4.0 Commercial Intent

In this section, we will talk about how spammers make money and how to look for commercial intent.

Most spam pages have commercial intent. Spammers create spam pages to make money and earn commissions when users make a purchase on an affiliate merchant site or when they click on a PPC ad.

Pages are spam if they exist primarily to make money and not to help users.

Please remember: Some spam pages do not have obvious moneymaking intent. If a page is created to change search engine rankings or even to do harm to users' computers with sneaky downloads, it is spam even though you are unable to see how the page is making money.

4.1 Thin Affiliates

An **affiliate** earns money from affiliate commissions. An affiliate website exists primarily to make money. Affiliates often show content from other "real" merchant sites, such as Amazon or eBay, or a good hotel or travel website. When users click on links to buy products or make reservations, they are oftentimes redirected to the real merchant page.

A **thin affiliate** is a site that offers little additional information and does not offer substantial value to users compared to many other sources on the Web. For example, an affiliate that has only copied content from the merchant site is considered a thin affiliate. This is a moneymaking spam technique.

4.1.1 Recognizing Thin Affiliates

To help determine if a page is a thin affiliate, you can do the following:

- **Look for original content on the page.** The quality of an affiliate page or site depends on how much added value, usefulness, or original/additional information is available on the page that is not easily available elsewhere on the Web. If the page has the same "cookie-cutter" text or functionality found on dozens or hundreds of other sites, it is more likely to be spam.
- **Look at the domain registrants.** If clicking a button takes you to another page, check to see "whois" the registrant (or owner) of the two domains. If the registrant is the same, the page is typically not a thin affiliate. Please follow the instructions for checking "whois" in Section 2.3.1 of the Webspam Guidelines.

4.1.2 Recognizing True Merchants

Features that will help you determine if a website is a true merchant include:

- a "view your shopping cart" link that stays on the same site.
- a shopping cart that updates when you add items to it.
- a return policy with a physical address.
- a shipping charge calculator that works.
- a "wish list" link, or a link to postpone the purchase of an item until later.
- a way to track FedEx orders.
- a user forum that works.
- the ability to register or login.
- a gift registry that works.

Please note the following:

- A page does not need to have all of these features to be considered a true merchant.
- Yahoo! Stores are true merchants – they are typically not thin affiliates.
- Some true smaller merchants take users to another site to complete the transaction because they use a third party to process the transaction. These merchants are not thin affiliates.

Many large web retailers offer affiliate programs. Some of the most common examples are Amazon.com, eBay.com, Zappos.com, Allposters.com, Hotels.com, Orbitz.com, and Overstock.com.

4.2 Pure PPC Pages

We refer to pages with PPC ads only (or with PPC ads and very little other content on them) as pure PPC pages. The spammer makes money when a link is clicked. No purchase is necessary. Pure PPC pages may have links to other spam pages that also contain PPC ads. Pure PPC pages are spam. Fake directory pages also can be considered pure PPC pages.

4.3 Parked (Expired) Domains

Definitions of “Domain”: The word “domain” can have two different meanings for raters:

- It can refer to one of the elements in the DNS (Domain Name System), such as .com, .org, .edu, .net, .gov, .it, .uk, .cn, .es, etc., that organize Internet addresses.
- It can refer to the set of words (URL) that identifies the web address of a specific entity, such as microsoft.com, harvard.edu, baidu.cn, etc.

In this section, when we use the word “domain,” we are referring to the second meaning.

When companies go out of business, are acquired by another company, change their name, or fail to pay their domain registration fee, the domain name “expires” and may be purchased by someone else.

Parked Domains: Spammers sometimes buy expired or expiring domains and put their own content on the page. Such sites are referred to as “parked domains” or “expired domains.” Their value to spammers is in their pre-existing links. Pages that previously linked to the expired domain will now link to the spammer’s page.

Spammers also purchase the following kinds of domains, which we will also refer to as parked domains, since they are similar in appearance:

- Domains which are close in spelling to real domains, hoping that users will mistype the domain name or URL and land on their websites, which contain PPC ads.
- Domains that users might type when looking for a website to use.

A typical parked/expired domain contains some or all of the following:

- A list of sponsored links.
- A list of popular categories.
- A list of categories that contains the keywords.

Recognizing Parked/Expired Domains

- **Look at the links.** Typically, all of the links on a parked domain are paid links. There is little to no original content on the page.
- **Look at the domain name (URL).** On a parked domain, the domain name (URL) often has little or nothing to do with the content on the webpage. You may see the keywords, but the links are usually generic and the linked pages are not really associated with the query.
- **Look at the page on the Internet Archive.** Go to <http://www.archive.org/index.php> to enter the URL and view the page as it appeared previously, when its original owner maintained it. If the original site was different, it is probably a parked domain.

You will soon become familiar with the format of parked / expired domains.

4.4 Pages with Unhelpful Content and PPC Ads

Some webpages with content are created just for the purpose of putting ads on them; writers are paid by spammers to create articles on a wide range of topics. Often the articles are very generic and do not provide a lot of good information, but they are original. You will not find the articles on another website. Although you may be convinced that the intent is to deceive, if the content makes sense and appears to be original, you will not be able to assign a **Spam** flag to such pages. You will have to use your judgment.

- Look at the content and ask yourself some of these questions:
 - Is the content likely to be helpful to users or is it too general, too poorly written, or gibberish?
 - Does the page provide substantial value when compared to other pages in search results?
 - Does the page have an excessive amount of ads that distract from or interfere with the main content?
 - Would you trust the content?
 - Would you be comfortable giving your credit card information to the site?
- Try to determine if the page was made by a human or by a computer.
- Try to determine why the page was created.

5.0 Phishing Websites

Phishing is an attempt by unscrupulous people to obtain sensitive information from Internet users. Some of you may have received emails in your own email accounts that look as if they're from legitimate companies, but upon closer inspection are not. Often these emails ask for sensitive information.

The landing page in the following task also asks for sensitive information and is another type of phishing.

Query [runescape gold], English (US)

URL <http://www.gprunescape.com/>

This landing page should make users (and raters) very suspicious and cautious. The spelling and grammar are bad and unprofessional, and the page feels “spammy.” What is most worrisome is that the page asks for the user’s bank password and pin number!

Even though we would not want to interact with the page, this type of phishing does not go against the Webspam Guidelines and the page should not be flagged as spam or malicious.

Please remember to only flag pages that fall in one of the spam categories described in the guidelines. Some phishing pages may be spam, but this one is not.

6.0 Spam and the Resolving Stage

It is not uncommon for tasks to go into the “resolving” stage because raters disagree on whether a page should be assigned **Unratable: Didn’t Load** or a rating from the rating scale and a **Spam** flag. The disagreement occurs because raters see different pages when they click on the link in the task. These differences may be due to timing, or they may be due to browser version and/ or setting differences.

When a task goes into the resolving stage for this reason and the page you see matches the criteria for **Unratable: Didn’t Load**, please take another look. Since other raters see a spam page, it is obvious that they are looking at something different from what you see. Here are some things you can try:

- Update to the most current browser version.
- Look at the source code or disable JavaScript.

If you still do not detect spam, do not assign a **Spam** flag.

Please be aware that spam pages frequently stop loading after a period of time. If you detect spam one day, but the page does not load for you the next day, please do not change your rating, (i.e., do not remove the **Spam** flag).

7.0 Conclusion

Spam recognition is a skill that is developed through practice and exposure.

Remember to look at the page as a whole. Spam pages usually have some of these characteristics:

- PPC ads are usually very prominent on the page, and it is obvious that the page was created for them.
- If you do a text search, you will find that the content has been copied.
- If you visually remove all of the spam elements from the page (PPC ads and copied content), there is little or no value remaining.

Good pages usually have these characteristics:

- The page is well-organized. There may be ads on the page, but they are well identified and not distracting.
- If you do a text search, the original page is usually the first result displayed.
- The page will have value to the user. A good search engine would want the page in a set of search results.

Here are the spam flags that you will use:

- **Not Spam:** If you do not believe that a page is spam, you should assign a **Not Spam** flag.
- **Maybe Spam:** If you find a page to be “spammy,” but you do not feel comfortable saying that the page is definitely spam, you should assign a **Maybe Spam** flag.
- **Spam:** If you believe that a page has been designed using the deceptive web design techniques described in these guidelines, you should assign a **Spam** flag.

When unsure which flag to use, remember to ask yourself these questions:

- Does the page provide the user with a good search experience?
- Does the page contain original content that would be helpful to users?
- Do you think the page should be included in a set of search results?
- Is the page designed for users? Is there a human element to the page?
- If you removed the PPC ads and copied text from the page, is there useful content remaining?

If you answer “yes” to these questions, the page is probably **not** spam.