User Manual for soblogitsgood

Yubin Kim, Daniel Burstyn, Gobaan Raveendran, Nathaniel Flath

November 25, 2009

Contents

1 Introduction 2

1.1 Outline . . . 2

1.2 Product Overview and Motivation . . . 2

2 Search Basics 2

2.1 Some Basic Facts . . . 2

2.1.1 How to read search results . . . 2

2.2 Break it down . . . 3

2.2.1 How to read the breakdown . . . 3

2.3 Help Mode . . . 4

3 Advanced Features 5

3.1 Sentiment Search API . . . 5

3.2 Data Access API . . . 5

4 Glossary 6

1 Introduction

1.1 Outline

This document will first cover what our product, So Blog It's Good, does, and the motivation for creating it. Common, simple ways to use it are covered next, along with descriptions of each of our screens. We then cover the advanced features and how to use our developer API. Finally, a glossary of terms is presented.

1.2 Product Overview and Motivation

soblogitsgood at its core is a large system that crawls the internet and collects blog information. Blogs in particular are an interesting data source because, unlike most web pages, they are typically written by a single person from their point of view. If we filter out this particular data we can investigate what the big topics the blogging population (the "blogosphere") is talking about.

soblogitsgood can be used in two very distinct ways. The first is its web interface, which allows users to query the database looking for **opinions** on a topic. A typical Google search will typically return a wide range of informative results. However, user will frequently want to hear about a product or service from another actual person. soblogitsgood can be used to find user stories and not just feature lists.

The second use is more advanced and aims to provide access to the collected data through an API. This API is available to developers who want to perform their own analysis on our corpus of blogs. This saves them the trouble of having to collect and clean all that data.

2 Search Basics

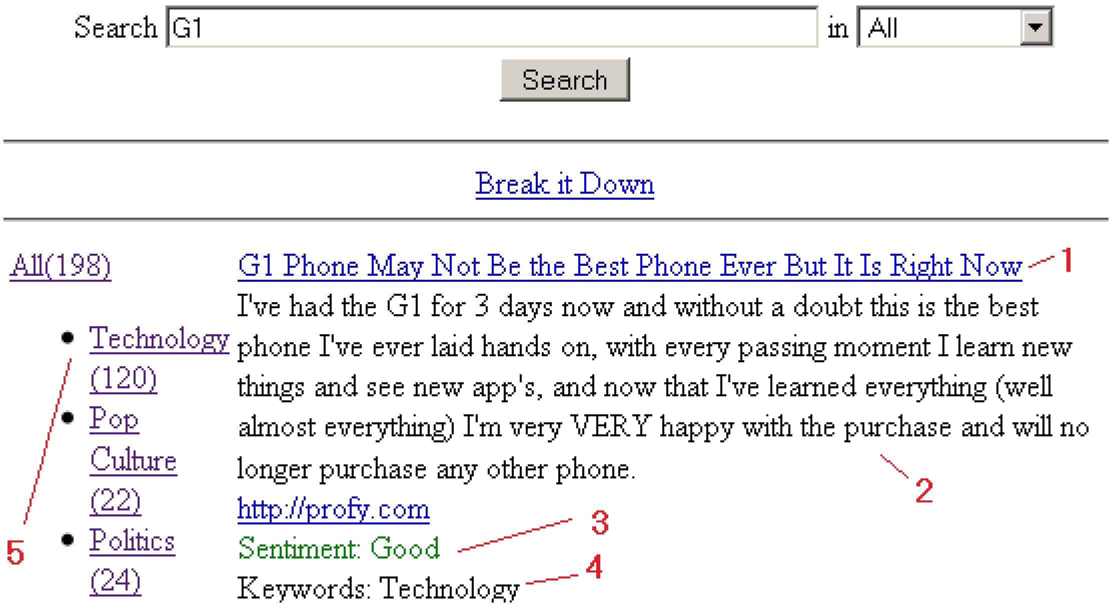
Search is simple: just type whatever comes to mind in the search box, hit Enter or click on the Search button, and soblogitsgood will search the blogosphere for blogs that are relevant to your query. Unfortunately, due to the nature of natural languages, it may be hard to get exactly the results you were looking for without refinement. Thus, the following tips can help you refine your technique to make the most of your searches.

2.1 Some Basic Tips

- **Select a category.** Certain terms in natural languages become almost impossible to disambiguate, for example logs about opera's can either be about the theatrical performances, or the browser, or many other words. Furthermore, results can be obtained faster and more filter functions can be used when a category is selected, due to the nature of the search algorithms used by soblogitsgood.
- **Search is always case insensitive.** Searching for new york times is the same as searching for New York Times.

2.1.1 How to read search results

soblogitsgood's goal is to provide you with results that are clear and easy to read. The diagram below points out four features that are important to understanding the search results page:



Search: G1 in: All

All(198)

- Technology (120)
- Pop Culture (22)
- Politics (24)

G1 Phone May Not Be the Best Phone Ever But It Is Right Now

I've had the G1 for 3 days now and without a doubt this is the best phone I've ever laid hands on, with every passing moment I learn new things and see new app's, and now that I've learned everything (well almost everything) I'm very VERY happy with the purchase and will no longer purchase any other phone.

http://profy.com

Sentiment: Good

Keywords: Technology

1. The Title: The first line of any search result is the title of the webpage.
2. The Snippet: A description of or an excerpt from the webpage.
3. The Sentiment: The basic stance the article has about the topic.
4. Keyword: The category this blog falls under.
5. Bins: Summary of categories for each result.

All these features are important in determining whether the page is what you need. The title is what the author of the page designated as the best short description of the page.

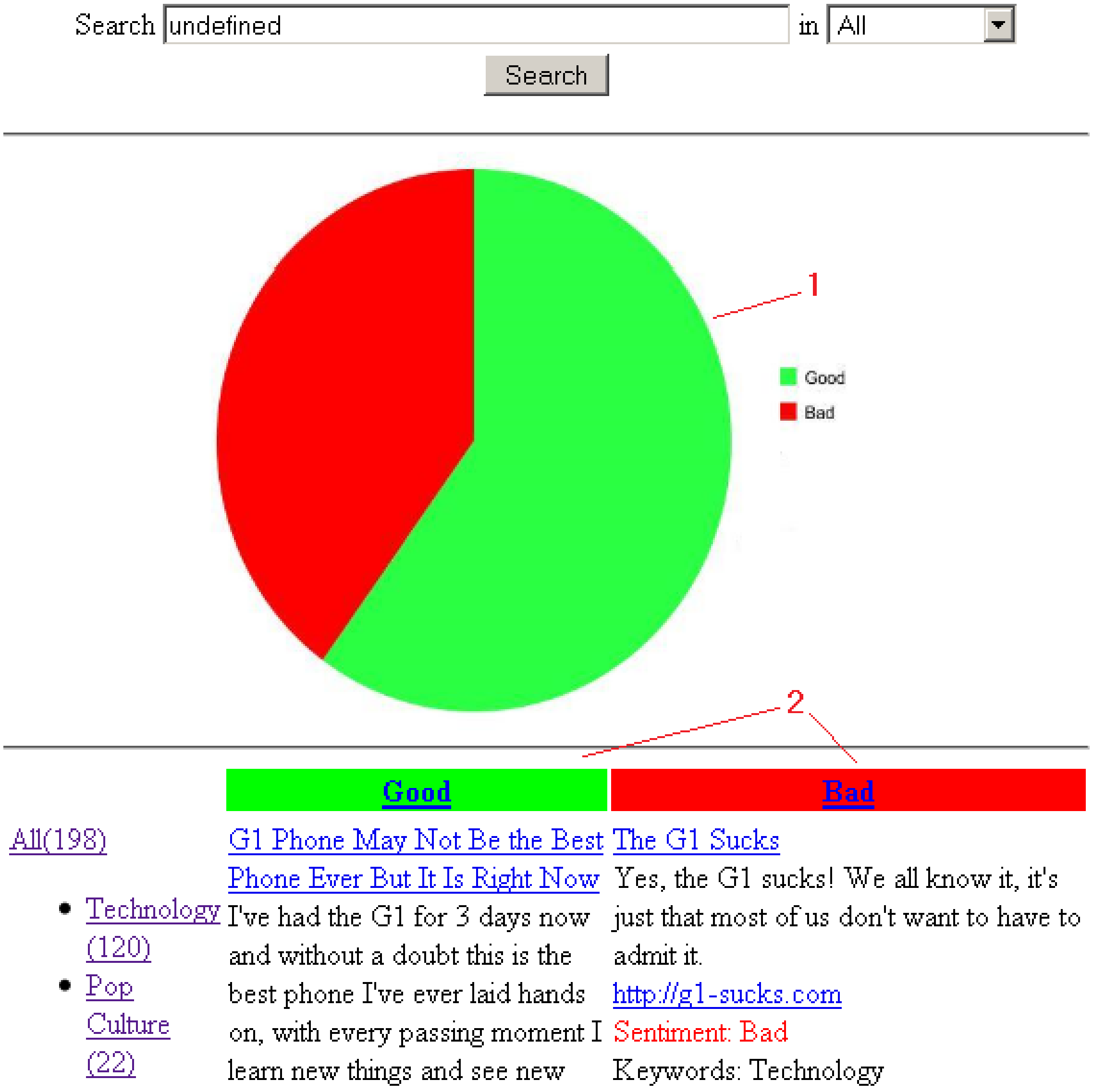
The snippet is soblogitsgood's algorithmic attempt to extract just the part of the page most relevant to your query. The sentiment plus keywords help you figure out if the article is what you want. Keywords help you verify that the link is the correct topic, and the bins will allow you to refine the results if you find that there are to many options, and many are irrelevent.

2.2 Break It Down

If a finer level of detail is wanted, press 'Break It Down'. This mode causes all the data to be organized into good and bad results. This allows you to gather a quick overview of the system, and get an overall idea of how positive and negative the results are.

2.2.1 How to read the breakdown

The following screen helps us determine how the breakdown works in order to get a quick overview of the system.



Search: undefined in: All

Search

	Good	Bad
All(198) • Technology (120) • Pop Culture (22)	G1 Phone May Not Be the Best Phone Ever But It Is Right Now I've had the G1 for 3 days now and without a doubt this is the best phone I've ever laid hands on, with every passing moment I learn new things and see new	The G1 Sucks Yes, the G1 sucks! We all know it, it's just that most of us don't want to have to admit it. http://g1-sucks.com Sentiment: Bad Keywords: Technology

The breakdown consists of two components:

1. Visualization
2. Result separation.

The visualization component consists of a simple pie graph that will allow you to quickly determine if the overall sentiment is for or against the given topic. The separation component allows us to isolate one set of opinionated results, and by clicking a sentiment we can display only the results that match the sentiment.

2.3 Help Mode

Users can get help with our product by clicking a 'help' link on any of our pages. This link redirects to help on the page they just came from; if clicked on our main page, it will explain the purpose of our product and how to use it. This ‘main’ help page will also be accessible from all other help pages. If ‘help’ is clicked on a results page, the opened page will also explain how to interpret the results presented.

TODO: screenshot

The above screenshot shows the main help page. As described previously, it has basic information about our product and how to use it. It additionally links to this document, or an updated user manual. The help navigation tree is on the left, allowing the user to easily access help on any feature they desire.

TODO: mock-up screenshots of help page

The above screenshot shows a sample help web page for an analysis page. As you can see, each feature on the page has an explanation—the visualization explains how the data was broken down, each column of data has an explanation describing it, etc. This page describes to the user precisely what they are looking at and what they can do in order to clear any confusion they may have.

3 Advanced Features

For the more advanced user, we provide APIs for making sentiment search queries, and for getting access to our raw data. We will write and make available language-dependent APIs for this. Currently a Java API is available and a language independent API using HTTP requests will be made available soon.

3.1 Sentiment Search API

The Sentiment Search API has the exact same function as soblogitsgood’s web interface but can be accessed programmatically. This is useful for advanced users that wish to retrieve blog post information and do their own analysis on top of to existing sentiment analysis. The Java API for this supports the following methods:

```
BlogResults sentimentSearch(String query);
BlogResults sentimentSearch(String query, SearchFilter filter);
```

BlogResults support methods to retrieve blog posts, and SearchFilter can be used to specify restictions on the results like posting data, or category.

Here is an example of how to use the API:

```
SearchFilter filter = new SearchFilter();
filter.setCategory(SentimentSearch.Catergories.TECHNOLOGY);
BlogResults results = SentimentSearch.sentimentSearch("iPhone", filter);
while(results.hasMoreResults()) {
  System.out.println(results.getNextBlogPost().getURL());
}
```

3.2 Data Access API

The Data Access API allows developers to get direct access to soblogitsgood’s database of blog information. This API will have access restrictions to avoid the general public from flooding our system with requests for huge numbers of results. The Data Access API is very similar to the sentiment search one, but has less limitations. The supported Java API methods are:

```
BlogResults getBlogs(SearchFilter filter);
BlogResults getRandomBlog();
BlogResults getRandomBlog(SearchFilter filter);
BlogPost getRandomBlogPost();
BlogPost getRandomBlogPost(SearchFilter filter);
```

We allow users to request all blog posts that match the specified filter. This is useful if they want to do an analysis over the large data set. We also support getting random blogs or blog posts so that developers can do an analyis on a smaller sample of our data.

4 Glossary

API API stands for Application Programming Interface and is an interface that a piece of software implements that allows other software to interact with it.

Bin Bins are categorizations that [The Product] uses to classify articles. These categories mirror the structure of data within our servers, and thus searches within a bin are much faster and more accurate.

Blog A blog is a ”web log” that consists of the authors thoughts, sentiments or general ideas. They can express anything from the day to day activities to a highly technical tutorial about a given topic.

Blogosphere The blogosphere is made up of all blogs and their interconnections. The term implies that blogs exist together as a connected community (or as a collection of connected communities) or as a social network in which everyday authors can publish their opinions.

Corpus A corpus is a collection of writing, that may be preprocessed and annotated to enhance automatic processing, testing, or verification of algorithms.

Crawl A crawler explores the web in some kind of automated manner in order to discover sites that match a certain criteria. In soblogitsgood’s use case, all sites are explored but only blogs are stored and processed.

Natural Language Natural languages are languages that are not formally defined and instead evolved over time through every day human interaction. Currently, English is the main natural language that we search using soblogitsgood.

Opinion In our case opinion is basically any kind of natural discussion about a given topic. Although opinions should be based on facts, in general they do not have to be, and in general an opinionated article should not be affiliated with the topic is is discussing, to avoid bias.

Sentiment Sentiment is the overall emotional content within an article. Some articles for example will be angry and others may be gleeful. Some sentiments can be classified into positive and negative allowing for refinement.