Rocksteady – A text analytics system

Khurshid Ahmad, Professor of Computer Science

Barry Redmond, Research Officer and Software Architect

Peter Dunne, Research Officer and Software Engineer

Document History

Version	Date	Person	Comments
1	27 July 2011	Khurshid Ahmad	Initial Document Creation
2	11 July 2011	Barry Redmond	Modified Formatting. Inserted this table. Added to 'How to use Rocksteady' section.

Table of Contents

About *Rocksteady*

Rocksteady: A text analytics system.

The repetition of names of persons, places, ideas and events is used sometimes for emphasizing the importance of what a name represents. The same appears to be true of the affect words repeated preferentially to show negative/positive sentiment. Change in management of an enterprise, or change of government, is often characterised by the rapidity of citations of names and in the citations of affect terms. Our ***Rocksteady*** system is designed to track such changes.

- ***Rocksteady*** is the first t**ext analytic systems** which can collect, analyse and aggregate news and blogs **at different scales of time**.
- ***Rocksteady*** produces a **unique profile** of changes in news related to **people, topics, and sentiment.**
- ***Rocksteady*** is the first system to c**orrelate details of the changing world** – stock market prices, opinion poll numbers, sales figures– with text analytics – **news flow, sentiment and topics**.
- ***Rocksteady*** has been used in the analysis of changes in the citation of political parties and their leaders during an **election campaign**; in the analysis of news and prices of c**ommodities like oil**.
- ***Rocksteady*** **uniquely identifies the sentiment** articulated in individual texts, the strength of the sentiment, and the orientation of the sentiment.
- ***Rocksteady*** can compute an **overall affect score** for a collection of documents using a proprietary technique.

Example Uses of *Rocksteady*

Rocksteady: The analysis of the Irish General Election, 2011.

Rocksteady was used to collate, analyse and aggregate the 2011 Irish General election news within Ireland. This consisted of **forty newspapers** comprising **two million words**, on topics related to politics. This analysis was run for the entire Election Campaign between December 2010 and February 2011. The full analysis can be found at http://electionnewsanalysisireland.wordpress.com/.
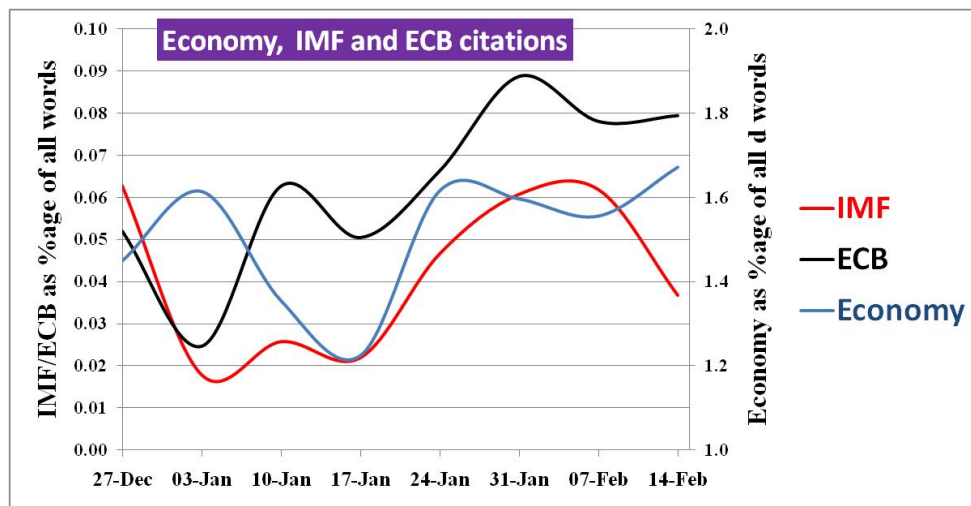
The citations of the individual party leaders was charted by *Rocksteady*:
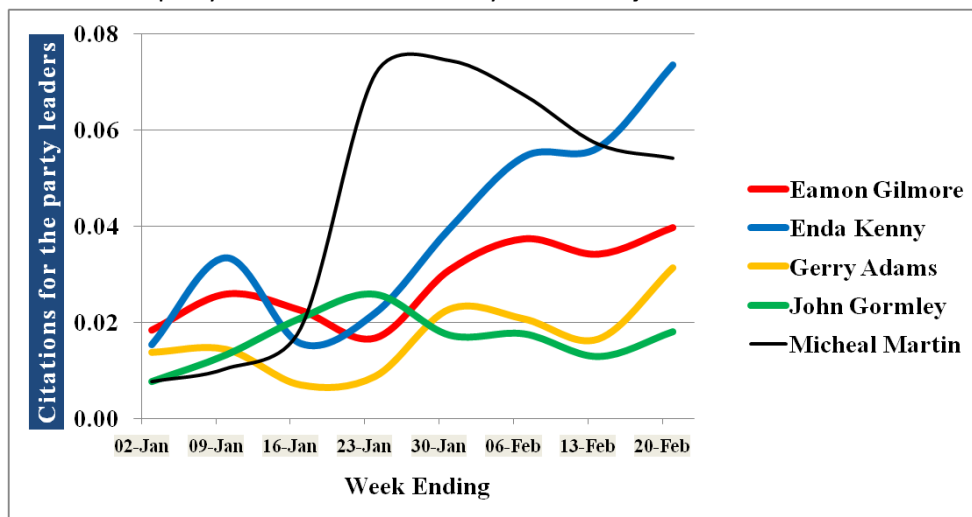
The change in the overall citations of the five main parties follows a similar pattern to their leaders:
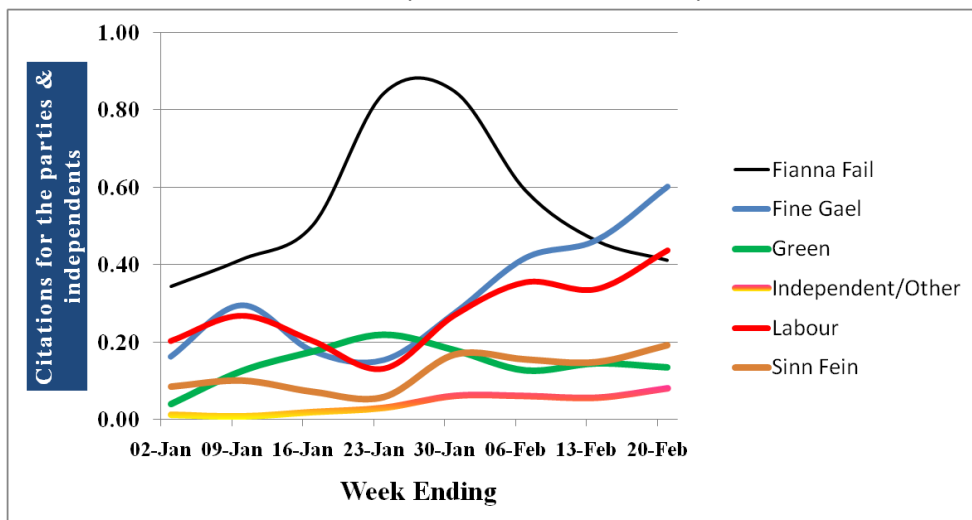

The economy was a key topic with two key change agents, the IMF and the European Central Bank

Economy, IMF and ECB citations

Rocksteady: The Arab Spring, oil companies and the price of oil.

The onset of Arab Spring in March 2011 led to renewed pressure on the price of oil. *Rocksteady* was deployed to collect, analyse and aggregate citations of 69 oil companies, fields, and refineries together with the citations of major oil producers and consumers. *Rocksteady* collected over 2240 news and blog articles (c. 1.5 million words) from 30 news sites and blogs dedicated to the oil industry during March 2011. The full analysis can be found at http://expertinoil.wordpress.com/2011/03/11/week-of-the-21st-february-to-the-27th/.

British Petroleum had the highest citations partially due to their problems in the Gulf of Mexico and in Libya:

News about Oil Companies - 1-31 March 2011
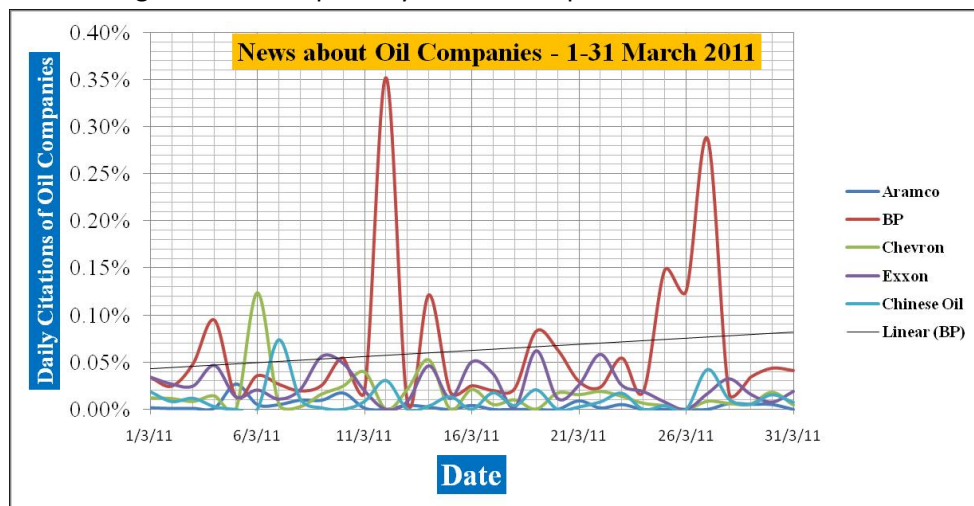
Rocksteady showed that citation of the names of the 69 companies led and lagged the price of oil future contracts and usually had an inverse relationship.

Variation in news about oil companies and oil forward contracts
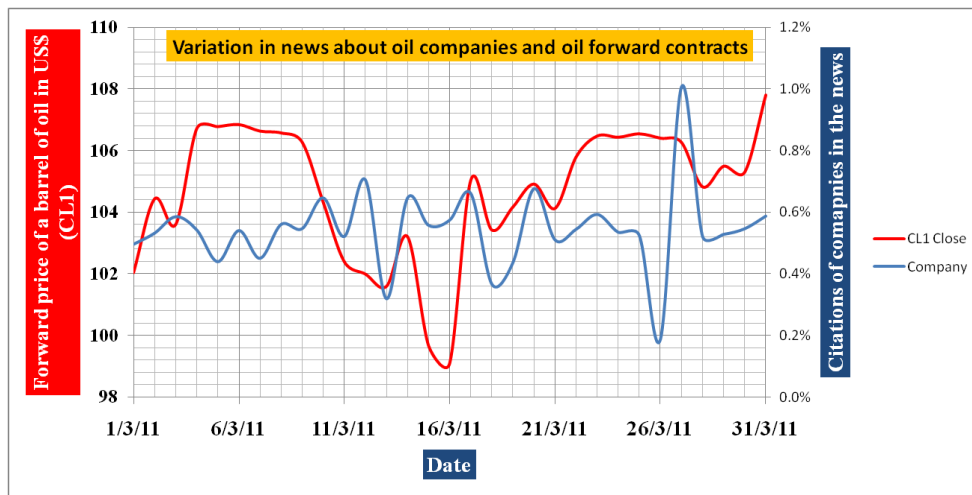
Rocksteady detected sentiment in the oil news was subdued punctuated by three spikes in negative sentiment:

Sentiment Change in Oil News
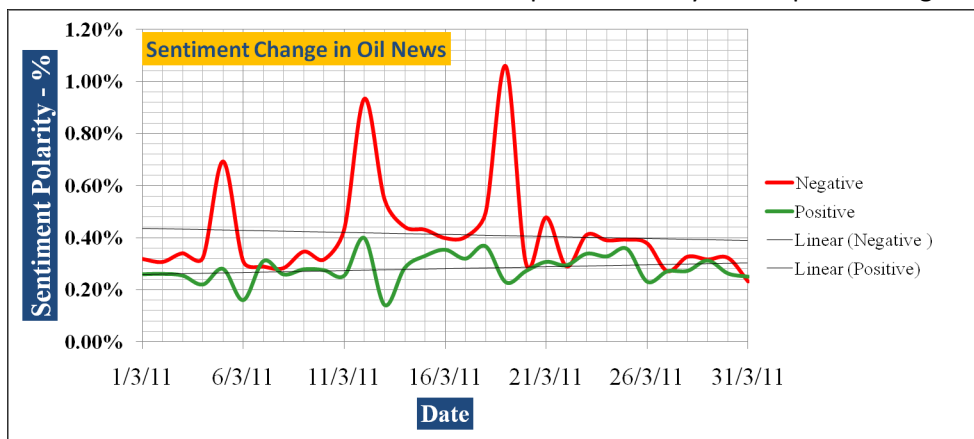
The sentiment spike was related to developments in Libya.

The citations to the country in 2240 news items at the beginning of the Libyan civil war almost equal to the citations to the G8:

G8 and Libya in Oil News
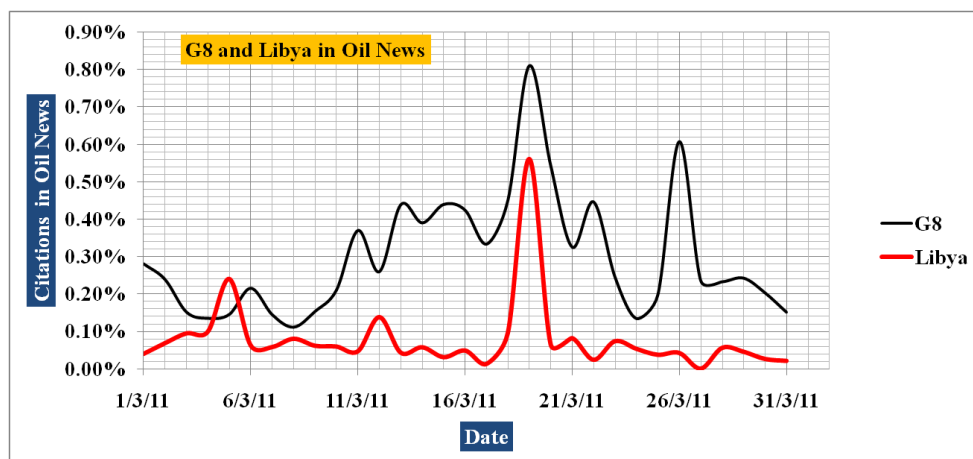
How to use Rocksteady

Rocksteady: Setup

When **Rocksteady** is run for the first time, you will be asked to select some basic settings before you can do any analysis:

- Select a Data and Corpus directory
- Select a dictionary
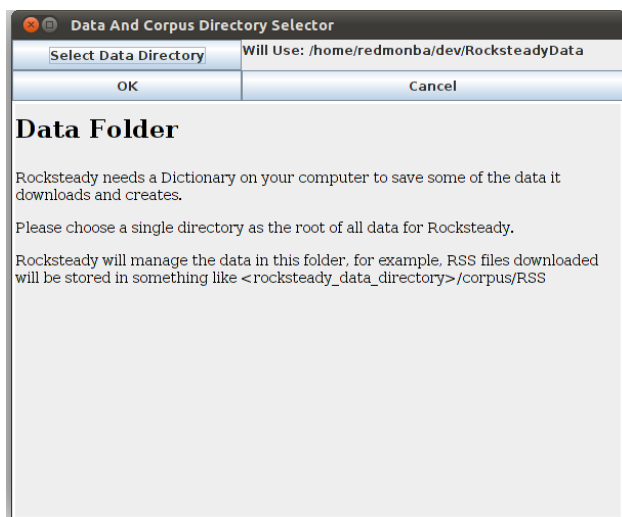- Select an Articles Analyser.

You may also want to:

- Set your proxy settings - if you access the Internet via a proxy.

All of these settings can be accessed and changed later from the *Options* menu in the menubar:
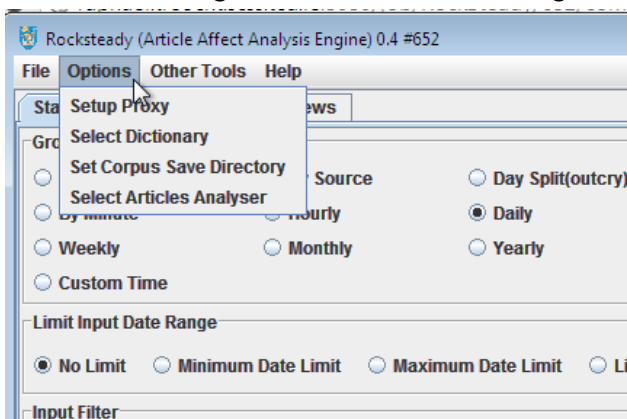
Select a Data and Corpus Directory

Rocksteady can create many files and folders, e.g. it can save all the articles it scrapes and analyses from the Internet. You should tell **Rocksteady** where you want it to save these.

Click 'Select Data Directory' to open a file chooser dialog.
Chose the folder where you want any folders and files that **Rocksteady** creates to be saved. You should create a new folder for this (e.g. a folder called 'Rocksteady' in '*My Documents*').
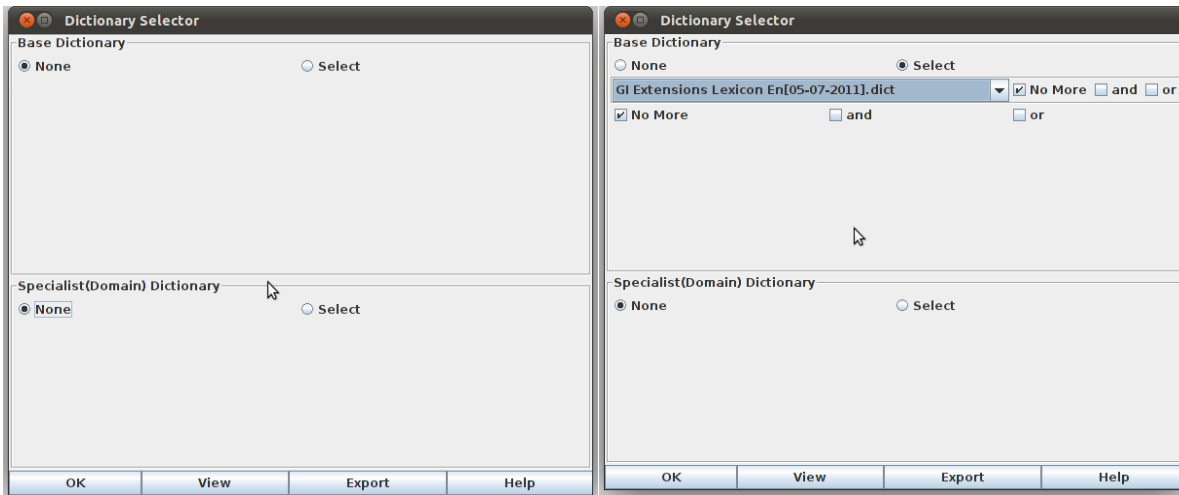Click '*OK*'.

Select a dictionary

At the core of **Rocksteady** is a dictionary. **Rocksteady** comes with a number of built in, proprietary, dictionaries. You can also build your own dictionaries, or merge dictionaries using the dictionary selection dialog tool.

Unless otherwise advised, you should chose one 'Base Dictionary' and 'None' for the Specialist Dictionary.

The 'Base Dictionary' you should chose is called 'GI Extensions Lexicon En[05-07-2011].dict'. The text within the square brackets is a date, in the format: [dd-mm-yyyy]. If you see a later dictionary with the same name, it is probably a better revision of that dictionary, and you should chose that.
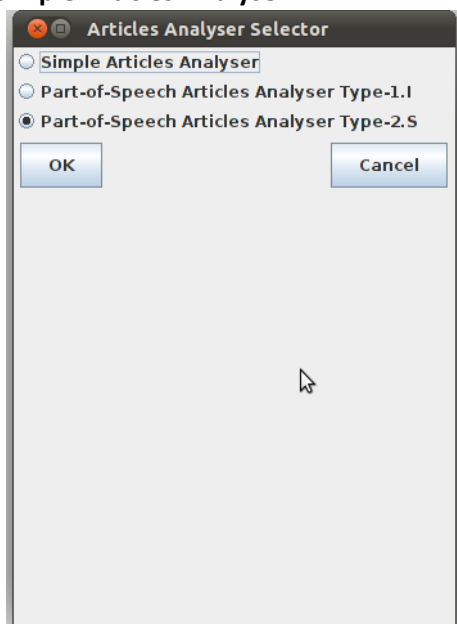

Once you've set up your dictionaries as shown, click 'OK' to save your dictionary preferences.

Select an articles analyser

Rocksteady can analyse the articles using a number of different techniques and algorithms. These algorithms have trade offs in terms of speed, accuracy and features.

Simple Articles Analyser

This is the default.
This can handle multi-word-terms in the dictionaries.
This can NOT find Named Entities.
This is the fastest.

Part-of-Speech Articles Analyser Type-1.I

This is NOT recommended.

Part-of-Speech Articles Analyser Type-2.S
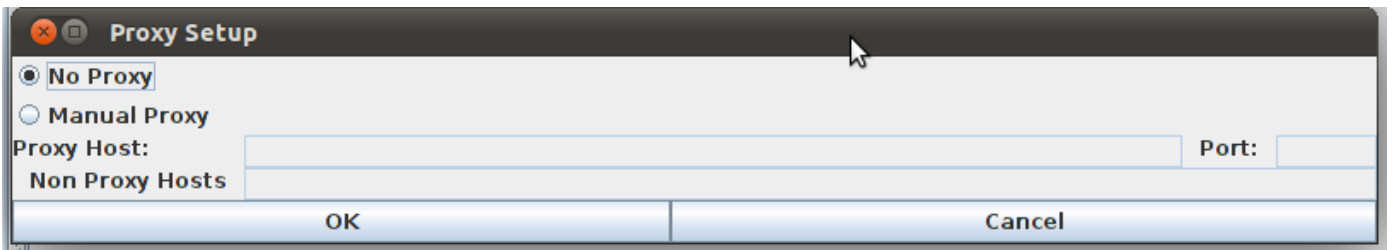
This has improved disambiguation.
This can find Named Entities.
This can NOT handle multi-word-terms.

Once you have chosen the analyser you wish to use, click 'OK' to save your preference.

Set your proxy settings

Rocksteady can run it's RSS feeds and other Internet tools through a proxy, however it currently only supports a simple http proxy, without authentication.

If you are **not** behind a proxy, select "No Proxy".

If you are behind a proxy, select "Manual Proxy". In the 'Proxy Host' box, enter the host address (e.g. 'my-proxy.myCompany.com/proxy' or '192.168.5.6'), and in the 'Port' box enter the port your proxy listens on (e.g. 8080). Additionally, you can optionally specify any hosts (addresses) which it should NOT use the proxy for (e.g. 'localhost').

Rocksteady: Overview

The main screen of **Rocksteady** is shown below. It is split into two sections: on the left is the *Control Panel*, and on the right is the *Data Panel*. The menu-bar is located at the top.
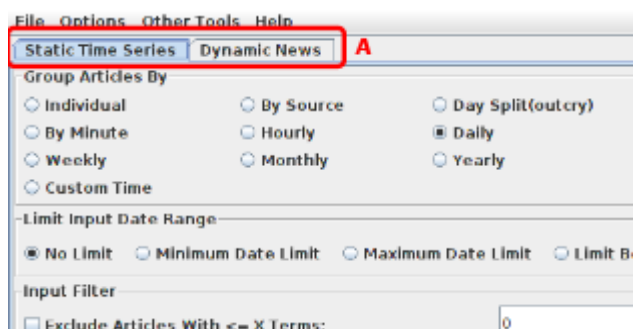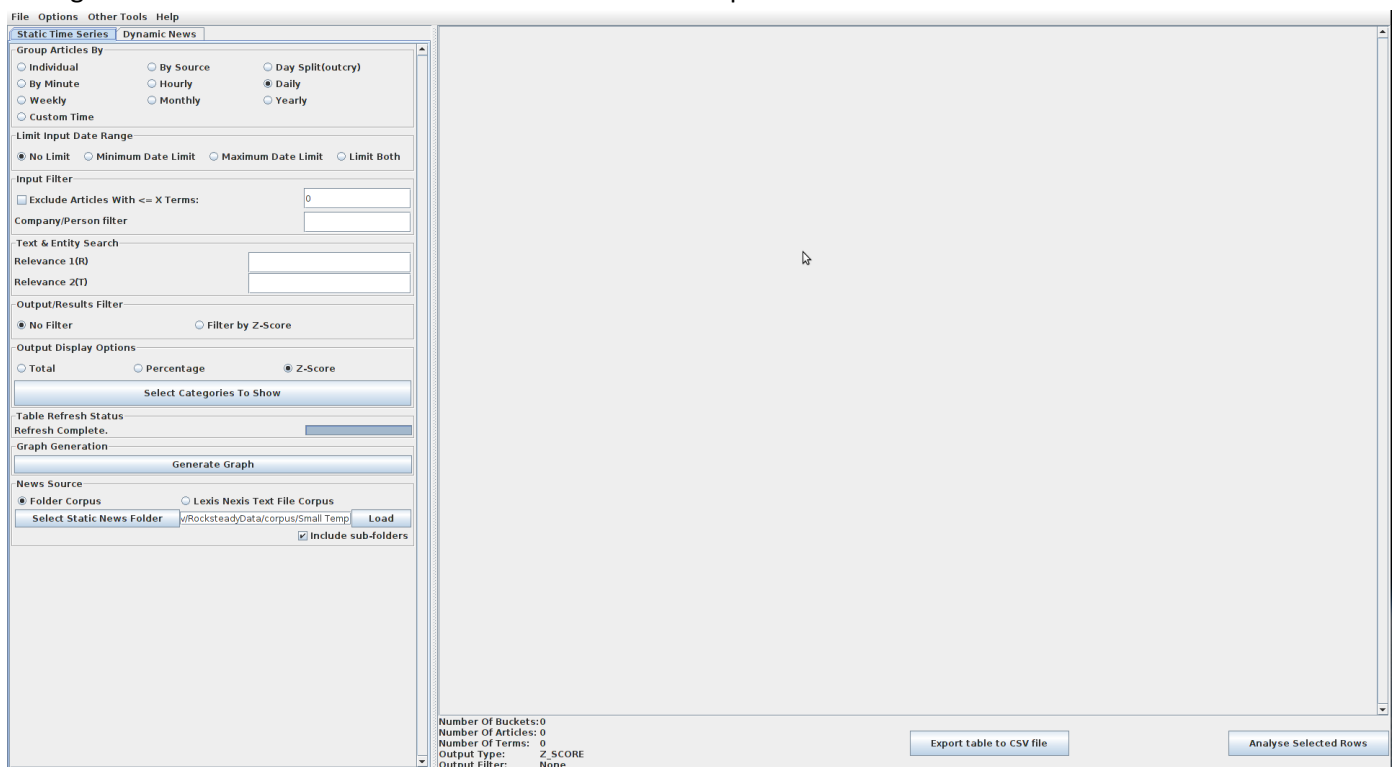
The **Rocksteady** system allows the analysis of either online texts (Dynamic News) or pre-stored news texts (Static Time Series). You can select between these two options using the tab at the top of the Control Panel (marked 'A').

To start with, it is unlikely you will have any texts in the correct format for **Rocksteady** to analyse, so this guide will

start with the **Dynamic News**.
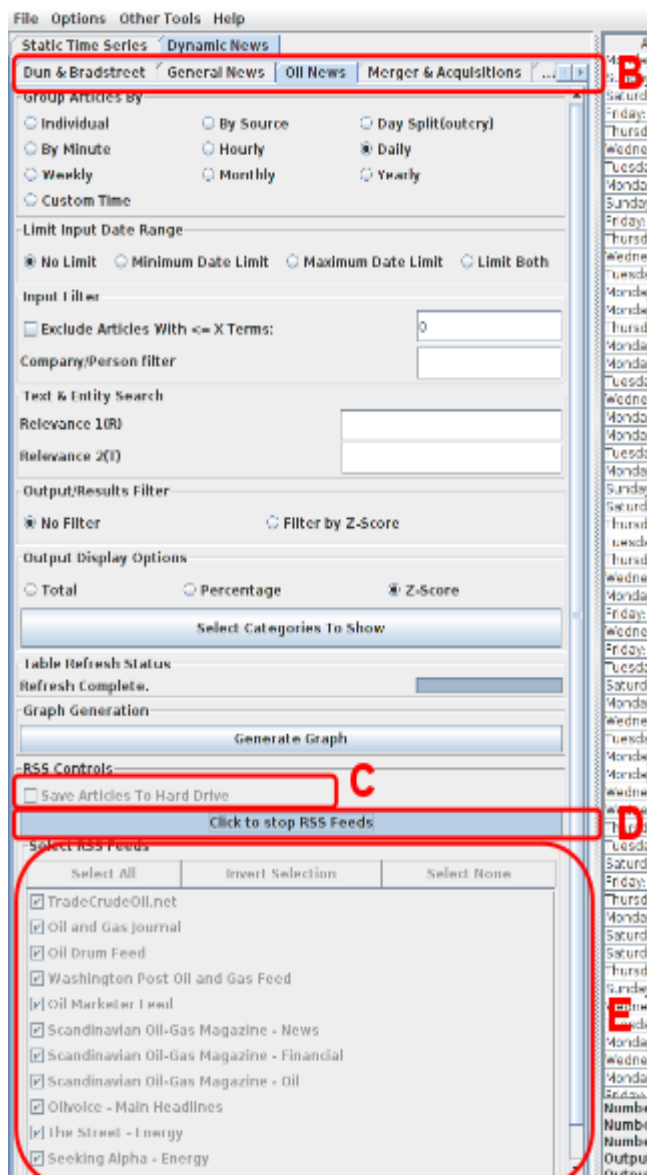
Dynamic News

- Once you've selected the *Dynamic News* tab, you'll see sub-tabs (marked 'B'), which are RSS feeds aggregated by some topic or area.
- "General News" contains **all** the RSS feeds, so if you're unsure, this is a good choice.
- Once you've selected your set of feeds, you can customise exactly which feeds in the set will be run, this is done by ticking the boxes beside their names in the "Select RSS Feeds" section at the bottom (marked 'E'). The buttons at the top of this section are a convenient way of selecting or deselecting multiple feeds quickly.
- The button marked '**D**' will say "Run Selected RSS Feeds" before you click it. You can start the selected feeds by clicking it.
- Click '**D**' again to stop running the feeds.
- Data should start filling the Data Window on the right after you start the feeds. (This may take up to 30 seconds, depending on your computer).
- Finally, you can save all the articles which *Rocksteady* finds while running the RSS feeds for future additional analysis (in the *Static Time Series* tab) by checking the box called "Save Articles To Hard Drive" (marked 'C'). These will be saved in the 'Data and Corpus' folder you specified during setup.
- You cannot change '**C**' or '**E**' while the feeds are running.

Static News

If you have a corpus of articles stored on your computer, *Rocksteady* can analyse these articles using the *Static Time Series* tab.

- The controls to select the file or folder to analyse using the *Static Time Series* tab are highlighted at **G**.
- This guide will assume you're analysing a corpus in a folder. As such, ensure that 'Folder Corpus' is selected.
- If you wish to include sub-folders, and not just the article files in the specific folder you will specify, ensure that 'Include sub-folders' is ticked.
- You can select the folder to analyse using the 'Select Static News Folder' button.
- Alternatively, you can type or paste the location into the text-field, and then click 'Load'.
- After you've selected the folder, or clicked 'Load', the articles will be loaded in, and analysed, and the results will be displayed in the panel on the right.
- This process may take a long time, depending on the size of the corpus, and the speed of your machine.
- You can see the progress in the area marked **F**.

The Data Panel

Once **Rocksteady** has analysed the news, it displays the results in the *Data Panel*, which is the large panel on the right hand side of the main window.

Each row, in the area marked **H** is show the aggregate results for a number of articles. How the articles are

aggregated, and which rows they go into is selected in the "Group Articles By" section of the *Control Panel*. In the image above, 'Daily' is selected, and as such each row represents a single 24 hour period.

The data may be sorted by any column, by clicking on the column header, clicking a second time will reverse the sort order on that column.
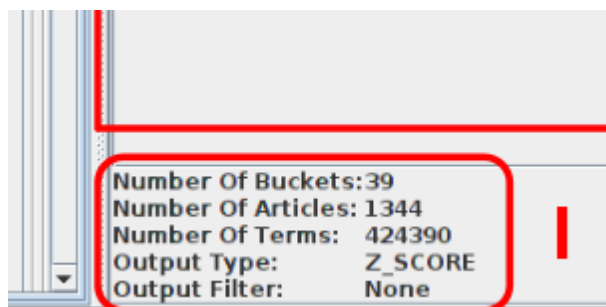
The area marked **I** contains some statistics for the data in the table in **H**. In this example, it shows that there are 39 rows (Buckets), with a total of 1344 articles, and 424,390 terms (words) in those articles. It shows that the values in each of the Affect cells is a Z-Score, and finally that there is no filter on the data in the table.

The button marked **J** says "Export table to CSV file". You can use this button to save the data in the table to a *Comma Separated Values*(CSV) file which can then be read in by a spread sheet (e.g. Excel) or some other application to further analyse the data.

The button marked **K** says "Analyse Selected Rows". This will open a new window where all the articles from any rows which are currently highlighted can be analysed separately.

News and Affect Aggregation

- The system analyses the news data using a bag of words approach using specially constructed dictionaries.

- The system can display the results of the analysis at different scales by the minute, hour, daily, monthly and so on.

- The user can specify his/her own scales. The news is collected from different sources and the results can be displayed accordingly.

- *Rocksteady* can identify names and keywords that are relevant to the topic of texts in a document

- The frequency of names, affect categories can be displayed as total, percentage of words in a text, or as z-score.

- The user can choose any of the 180 affect categories for tabulation.

- The *Rocksteady* system continually gathers data from a large number of feeds and within minutes.

The user can ask to look at aggregate values for the citations of companies, citations related to exploration activities, and the report's of field:

Article Grou...	First Article Date ▼	Article Count	Term Count	Company	Exploration	Field
Thursday: ...	07-07-11 09:48:49+0100	64	37012	*0.46*	-0.55	1.97
Wednesda...	06-07-11 02:22:13+0100	34	17874	*0.16*	*2.47*	-0.26
Tuesday: 0...	05-07-11 01:55:02+0100	35	19378	1.05	-0.55	-0.75
Monday: 04...	04-07-11 03:56:36+0100	40	52559	*0.20*	-0.21	*0.75*
Sunday: 03...	03-07-11 04:23:36+0100	2	2479	1.77	-0.55	-0.75
Friday: 01/0...	01-07-11 03:52:03+0100	42	22677	-0.54	1.04	-0.75
Thursday: ...	30-06-11 07:50:03+0100	12	4282	-0.34	-0.55	1.29
Wednesda...	29-06-11 03:10:24+0100	5	1817	-1.01	-0.55	-0.75
Tuesday: 2...	28-06-11 07:00:50+0100	1	396	-1.75	-0.55	-0.75

The user can look at the variation in citations for a specific company, country, CARTEL or transnational organisation:

Article Grou...	First Article ...	Article Count	Term Count	BP	G20	G8	Libya	OPEC
Thursday: ...	07-07-11 0...	64	37012	1.38	-0.04	0.31	-0.14	-0.41
Wednesda...	06-07-11 0...	34	17874	-0.72	1.24	0.65	-0.59	-0.52
Monday: 04...	04-07-11 0...	40	52559	1.00	0.37	0.01	-0.39	-0.25
Sunday: 03...	03-07-11 0...	2	2479	-1.06	-1.36	-1.42	0.02	-0.32
Friday: 01/0...	01-07-11 0...	42	22677	-0.26	0.39	-0.06	-0.32	0.14
Wednesda...	29-06-11 0...	5	1817	-1.06	1.54	1.98	2.77	2.75
Tuesday: 0...	05-07-11 0...	35	19378	1.43	0.09	0.49	-0.52	-0.49
Thursday: ...	30-06-11 0...	12	4282	0.35	-0.66	-0.55	-0.24	-0.17
Tuesday: 2...	28-06-11 0...	1	396	-1.06	-1.58	-1.42	-0.59	-0.74

Or indeed the user can ask *Rocksteady* to output values of affect and sentiment:

Article Grou...	First Article ...	Article Count	Term Count	Active	Neg	POS	Passive	Strong	Weak
Thursday: ...	07-07-11 0...	64	37012	0.76	0.39	1.85	0.16	0.79	0.87
Wednesda...	06-07-11 0...	34	17874	0.67	0.26	0.26	-0.47	0.65	-0.53
Monday: 04...	04-07-11 0...	40	52559	0.89	2.11	0.67	-0.44	1.21	-0.62
Sunday: 03...	03-07-11 0...	2	2479	-2.12	-1.36	-0.28	0.48	-1.04	0.97
Friday: 01/0...	01-07-11 0...	42	22677	0.22	0.51	0.55	0.22	0.73	-0.12
Wednesda...	29-06-11 0...	5	1817	-0.09	-1.31	-1.48	-0.26	-1.49	1.96
Tuesday: 0...	05-07-11 0...	35	19378	0.91	-0.64	0.36	0.96	0.44	-0.22
Thursday: ...	30-06-11 0...	12	4282	0.10	0.20	-0.44	1.56	0.26	-1.36
Tuesday: 2...	28-06-11 0...	1	396	-1.34	-0.17	-1.48	-2.21	-1.54	-0.95