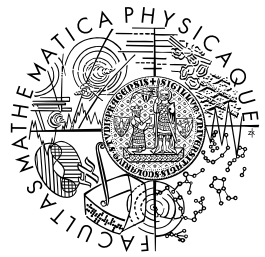

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

BACHELOR THESIS

Václav Stibor

Financial News Sentiment Analysis

Department of Software Engineering

Supervisor of the bachelor thesis: Supername Supersurname
Study programme: study programme
Study branch: study branch

Prague YEAR

I declare that I carried out this bachelor thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

Dedication. It is nice to say thanks to supervisors, friends, family, book authors and food providers.

Title: Financial News Sentiment Analysis

Author: Václav Stibor

Department: Department of Software Engineering

Supervisor: Supername Supersurname, department

Abstract: One key skill required to make good investments in the stock market is being able to correctly analyze news related to the finance and the business sector. Which company is diversifying its sectors or which company is showing signs of heading towards bankruptcy? You need to keep yourself updated with every little deal and fallout happening in the market. Financial news can be a little tricky to understand especially for those who are new to the financial world.

Keywords: key words

Contents

Introduction

In today's era of information explosion and constant flow of information, it becomes more time-consuming to keep track of associations and understand the published content through media and online news, primarily when investing in a specific area. For instance, investing in a company like Apple acquiring and processing a wide range of available information requires significant effort and time investment in studying articles and other sources.

Sentiment analysis, the ability to identify and evaluate the emotional charge of content, has evolved into a crucial instrument for comprehending opinions, attitudes, and the general atmosphere surrounding various topics. This work focuses on developing an application that allows users to visualize and analyze news sentiment using a knowledge graph network, even in real-time.

Knowledge graphs are becoming an increasingly popular tool for representing and linking information. In our application, they are used to allow users not only to track the sentiment of individual reports but also to link and visualize relationships between companies through articles. In this way, users can gain a holistic view of world events and understand the interactions between entities. This thesis will discuss the technical aspects of sentiment analysis and designing and implementing an application that conveys this information to users as efficiently as possible. The aim is to provide users with a tool that allows them to actively monitor and analyze the flow of information about emotional overtones as one of the key identifiers in trading decisions. The thesis will be structured as follows. Chapter 1 will discuss the theoretical background behind the stock market. Chapter 2 will give an overview of data sources and the data itself. Chapter 3 will discuss the sentiment analysis and design of the application. Chapter 4 will discuss the implementation of the application. Chapter 5 will discuss the evaluation of the application. Chapter 6 will discuss the conclusion and future work.

Zeptat se - použití 1. osoby množného čísla nebo pasivní formy (it is required)

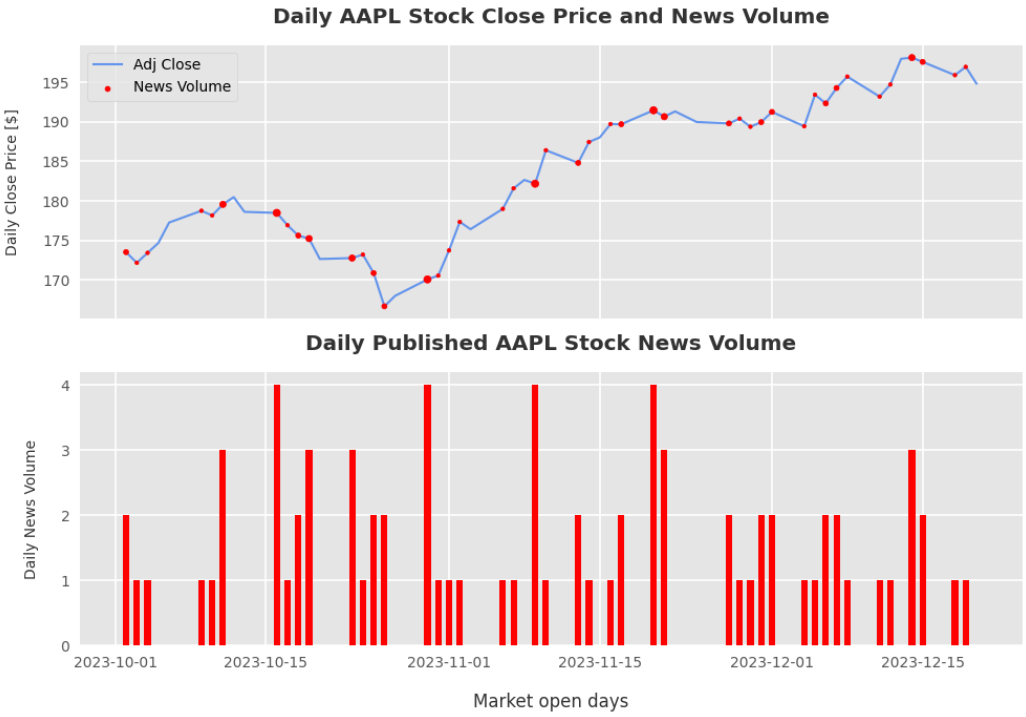

Figure 1 This caption is a friendly reminder to never insert figures "in text," without a floating environment, unless explicitly needed for maintaining the text flow (e.g., the figure is small and developing with the text, like some of the centered equations, as in). All figures *must* be referenced by number from the text (so that the readers can find them when they read the text) and properly captioned (so that the readers can interpret the figure even if they look at it before reading the text — reviewers love to do that).

Motivation

Related work

How to write Introduction

Introduction should answer the following questions, ideally in this order:

1. What is the nature of the problem the thesis is addressing?
2. What is the common approach for solving that problem now?
3. How this thesis approaches the problem?
4. What are the results? Did something improve?
5. What can the reader expect in the individual chapters of the thesis?

Expected length of the introduction is between 1–4 pages. Longer introductions may require sub-sectioning with appropriate headings — use `\section*` to avoid numbering (with section names like 'Motivation' and 'Related work'), but try to avoid lengthy discussion of anything specific. Any "real science" (definitions, theorems, methods, data) should go into other chapters.

It is very advisable to skim through a book about scientific English writing before starting the thesis. I can recommend '*Science research writing for non-native speakers of English*' by Glasman-Deal [1].

You may notice that this paragraph briefly shows different "types" of 'quotes' in TeX, and the usage difference between a hyphen (-), en-dash (–) and em-dash (—).

1. Data

Newspaper articles play a key role in our web application. We must consider several essential aspects to integrate these data into our web application to ensure a smooth and effective implementation. The following chapter will discuss these aspects from different perspectives, including the programmer's viewpoint and legislative considerations.

In section X.X. we will give an overview...

For our entity-level sentiment proposes, we need a body of each article as we discuss in Chapter **??** . When selecting data source of news article, it is essential to consider several main aspects.

refer to chapter Related Works, where we discuss maybe why others work only with titles

Reliability Expresses the degree to which a source can be trusted based on its history and reputation.

Availability Expresses the degree to which a source is available to the public.

Accessibility Refers to the ease with which the data source can be accessed. Consider factors such as API availability, data retrieval methods, and any restrictions on accessing the news articles.

Relevance Ensuring that the selected news articles align with the interests and requirements of your web application and its users.

Timeliness The freshness of the news data is crucial. Choose a source that provides timely updates to keep your application's content current.

Consistency Look for a data source that maintains a consistent format and structure, facilitating easier integration into your web application.

Licensing and Copyright Ensure compliance with legal considerations. Verify the licensing terms and copyright issues associated with using the news articles in your application.

1.1 Third party data providers

API

Application programming interface (API) třetích stran jsou dostupná v různých cenových plánech. Každý plán poskytuje odlišný rozsah přístupu k datům, který typicky spočívá v rozsahu dat, jenž jsou v rámci daného plánu dostupná. Dalším

Aspect	Description
Reliability	Expresses the degree to which a source can be trusted based on its history and reputation.
Availability	Expresses the degree to which a source is available to the public.
Accessibility	Refers to the ease with which the data source can be accessed. Consider factors such as API availability, data retrieval methods, and any restrictions on accessing the news articles.
Relevance	Ensuring that the selected news articles align with the interests and requirements of your web application and its users.
Timeliness	The freshness of the news data is crucial. Choose a source that provides timely updates to keep your application's content current.
Consistency	Look for a data source that maintains a consistent format and structure, facilitating easier integration into your web application.
Licensing and Copyright	Ensure compliance with legal considerations. Verify the licensing terms and copyright issues associated with using the news articles in your application.

Table 1.1 Considerations for Selecting News Article Data Source

nejběžnějším omezení je maximálním počtem dotazů v rámci specifikované časové periody. Drtivá většina poskytovatelů nabízí bezlpatné plány, díky kterým může vývojář otestovat různé endpointy a ověřit, zda odpovídají požadavkům jeho aplikace.

There is always some compromise at the expense of something else (, a proto bychom naší apilikaci dokázali omezit na počet dotazů tak). V naší aplikaci bychom se dokázali omezit na počet dotazů tak, abychom mohli uvažovat i bezplatného plánu anižbychom přišli o endpointy, jenž jsou pro naši aplikaci důležité. Avšak data obsahují častokrát velké mezery. Například Alpha Ventage poskytuje vyhledávání článků na základě tickeru a možností uvedení time range, ve kterém byly články vydány.

Vypadá to, že se můžeme dotazovat pouze na články v intervvalu 5 dní, avšak tento fakt není nikde v API zaznamenán.

Vždy jě něco na úkor něčeho jiného.

1.2 The Guardian

The Guardian is a British daily newspaper that covers American and international news for an online, global audience.

1.3 Dostupnost

2. Natural Language Processing

2.1 Model

There are many models for sentiment analysis of financial news, and each has its advantages and disadvantages. Some of these models are based on traditional statistical methods, while others are based on deep learning.

Among the latest and best models are those based on the Transformer architecture, which are capable of high accuracy and can process large amounts of data. Some of the most well-known ones include BERT (Bidirectional Encoder Representations from Transformers) by Google or GPT (Generative Pre-trained Transformer) by OpenAI.

2.1.1 BERT

Bidirectional Encoder Representation from Transformer is one of the most popular state of the art text embedding model published by Google. One of the reasons BERT is more successful is that it uses a context based embedding model. Without context, the word would have the same meaning in both sentences.

BERT looks at the sentence and figures out what words is related to in the sentence, and will create embedding of the word based on the context. BERT does this by using transformers, which is a state of the art deep learning architecture, that is mostly used for Natural language processing. The architecture uses encoder-decoder paradigm.

2.1.2 GPT

3. Architecture

Conclusion

In the conclusion, you should summarize what was achieved by the thesis. In a few paragraphs, try to answer the following:

- Was the problem stated in the introduction solved? (Ideally include a list of successfully achieved goals.)
- What is the quality of the result? Is the problem solved for good and the mankind does not need to ever think about it again, or just partially improved upon? (Is the incompleteness caused by overwhelming problem complexity that would be out of thesis scope, or any theoretical reasons, such as computational hardness?)
- Does the result have any practical applications that improve upon something realistic?
- Is there any good future development or research direction that could further improve the results of this thesis? (This is often summarized in a separate subsection called 'Future work'.)

This is quite common.

Bibliography

[1] Hilary Glasman-Deal. *Science research writing for non-native speakers of English*. World Scientific, 2010.

[2] Donald Ervin Knuth. *TEX and METAFONT: New directions in typesetting*. American Mathematical Society, 1979.

[3] Leslie Lamport. *LATEX: a document preparation system: user's guide and reference manual*. Addison-Wesley, 1994.

[4] Don Sparling. *English or Czenglish? Jak se vyhnout čechismům v angličtině*. Státní pedagogické nakladatelství, 1989.

[5] Edward R Tufte, Nora Hillman Goeler, and Richard Benson. *Envisioning information*. Graphics press Cheshire, CT, 1990.

[6] Edward R Tufte. *Visual display of quantitative information*. Graphics press Cheshire, CT, 1983.

[7] Claus O Wilke. *Fundamentals of Data Visualization*. O'Reilly Media, Inc., 2019. ISBN: 9781492031086. URL: https://clauswilke.com/dataviz/.

[8] Till Tantau. *The TikZ and PGF Packages (Manual for version 3.1.8b)*. Tech. rep. Institut für Theoretische Informatik Universität zu Lübeck, 2020. URL: http://mirrors.ctan.org/graphics/pgf/base/doc/pgfmanual.pdf.

4. Important first chapter

irst chapter usually builds the theoretical background necessary for readers to understand the rest of the thesis. You should summarize and reference a lot of existing literature and research.

You should use the standard *citations*.

Use \emph command like this, to highlight the first occurrence of an important word or term. Reader will notice it, and hopefully remember the importance.

This footnote is an acceptable way to 'cite' webpages or URLs. Documents without proper titles, authors and publishers generally do not form citations. For this reason, avoid citations of wikipedia pages.

Obtaining bibTeX citation Go to Google Scholar[1], find the relevant literature, click the tiny double-quote button below the link, and copy the bibTeX entry.

Saving the citation Insert the bibTeX entry to the file `refs.bib`. On the first line of the entry you should see the short reference name — from Scholar, it usually looks like `author2015title` — you will use that to refer to the citation.

Using the citation Use the `\cite` command to typeset the citation number correctly in the text; a long citation description will be automaticaly added to the bibliography at the end of the thesis. Always use a non-breakable space before the citing parenthesis to avoid unacceptable line breaks:

```
Trees utilize gravity to invade ye
noble sires~\cite{newton1666apple}.
```

Why should I bother with citations at all? For two main reasons:

- You do not have to explain everything in the thesis; instead you send the reader to refer to details in some other literature. Use citations to simplify the detailed explanations.
- If you describe something that already exists without using a citation, the reviewer may think that you *claim* to have invented it. Expectably, they will demand academic correctness, and, from your perspective, being accused of plagiarism is not a good starting point for a successful defense. Use citations to identify the people who invented the ideas that you build upon.

How many citations should I use? Cite any non-trivial building block or assumption that you use, if it is published in the literature. You do not have to cite trivia, such as the basic definitions taught in the introductory courses.

[1] `https://scholar.google.com`

The rule of thumb is that you should read, understand and briefly review at least around 4 scientific papers. A thesis that contains less than 3 sound citations will spark doubt in reviewers.

There are several main commands for inserting citations, used as follows:

- Knuth [2] described a great system for typesetting theses.
- We are typesetting this thesis with LaTeX, which is based on TeX and METAFONT [2].
- TeX was expanded to LaTeX by Lamport [3], hence the name.
- Revered are the authors of these systems! [2, 3]

4.1 Some extra assorted hints before you start writing English

Strictly adhere to the English word order rules. The sentences follow a fixed structure with a subject followed by a verb and an object (in this order). Exceptions to this rule must be handled specially, and usually separated by commas.

Mind the rules for placing commas:

- Use the *Oxford comma* before 'and' and 'or' at the end of a longer, comma-separated list of items. Certainly use it to disambiguate any possible mixtures of conjunctions: *'The car is available in red, red and green, and green versions.'*
- Do not use the comma before subordinate clauses that begin with 'that' (like this one). English does not use subordinate clauses as often as Slavic languages because the lack of a suitable word inflection method makes them hard to understand. In scientific English, try to avoid them as much as possible. Ask doubtfully whether each 'which' and 'when' is necessary — most of these helper conjunctions can be removed by converting the clause to non-subordinate.

  As an usual example, *'The sentence, which I wrote, seemed ugly.'* is perfectly bad; slightly improved by *'The sentence that I wrote seemed ugly.'*, which can be easily reduced to *'The sentence I wrote seemed ugly.'*. A final version with added storytelling value could say *'I wrote a sentence but it seemed ugly.'*
- Consider placing extra commas around any parts of the sentence that break the usual word order, especially if they are longer than a single word.

Do not write long sentences. One sentence should contain exactly one fact. Multiple facts should be grouped in a paragraph to communicate one coherent idea. Paragraphs are grouped in labeled sections for a sole purpose of making the navigation in the thesis easier. Do not use the headings as 'names for paragraphs' — the text should make perfect sense even if all headings are removed. If a section of your text contains one paragraph per heading, you might have wanted to write an explicit list instead.

Every noun needs a determiner ('a', 'the', 'my', 'some', …); the exceptions to this rule, such as non-adjectivized names and indeterminate plural, are relatively scarce. Without a determiner, a noun can be easily mistaken for something completely different, such as an adjective or a verb.

Consult the books by Glasman-Deal [1] and Sparling [4] for more useful details.

5. Results and discussion

You should have a separate chapter for presenting your results (generated by the stuff described previously, in our case in **??**). Remember that your work needs to be validated rigorously, and no one will believe you if you just say that 'it worked well for you'.

Instead, try some of the following:

- State a hypothesis and prove it statistically
- Show plots with measurements that you did to prove your results (e.g. speedup). Use either `R` and `ggplot`, or Python with `matplotlib` to generate the plots.[1] Save them as PDF to avoid printing pixels (as in figure 5.1).
- Compare with other similar software/theses/authors/results, if possible
- Show example source code (e.g. for demonstrating how easily your results can be used)
- Include a 'toy problem' for demonstrating the basic functionality of your approach and detail all important properties and results on that
- Include clear pictures of 'inputs' and 'outputs' of all your algorithms, if applicable

It is sometimes convenient (even recommended by some journals, including Cell) to name the results sub-sections so that they state what exactly has been achieved. Examples follow.

[1] Honestly, the plots from `ggplot` look much better.

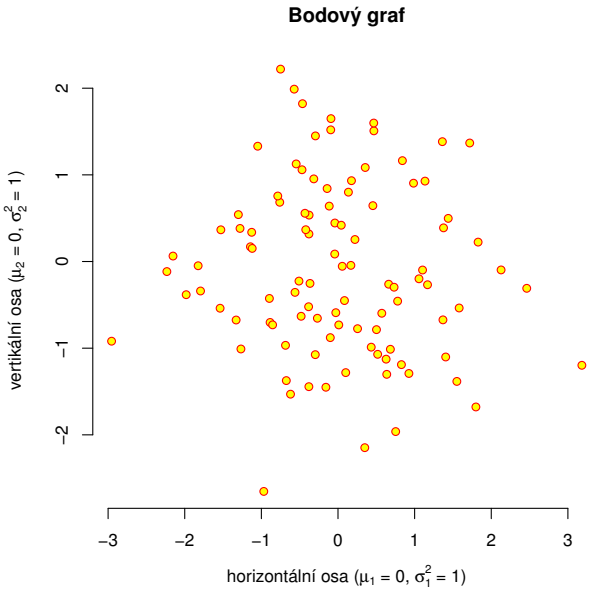

Figure 5.1 This caption is a friendly reminder to never insert figures "in text," without a floating environment, unless explicitly needed for maintaining the text flow (e.g., the figure is small and developing with the text, like some of the centered equations, as in ??). All figures *must* be referenced by number from the text (so that the readers can find them when they read the text) and properly captioned (so that the readers can interpret the figure even if they look at it before reading the text — reviewers love to do that).

5.1 SuperProgram is faster than OldAlgorithm

5.1.1 Scalability estimation

5.1.2 Precision of the results

5.2 Weird theorem is proven by induction

5.3 Amount of code reduced by CodeRedTool

5.3.1 Example

5.3.2 Performance on real codebases

5.4 NeuroticHelper improves neural network learning

5.5 Graphics and figure quality

No matter how great the text content of your thesis is, the pictures will always catch the attention first. This creates the very important first impression of the thesis contents and general quality. Crucially, that also decides whether the thesis is later read with joy, or carefully examined with suspicion.

Preparing your thesis in a way such that this first impression gets communicated smoothly and precisely helps both the reviewer and you: the reviewer will not have a hard time understanding what exactly you wanted to convey, and you will get a better grade.

Making the graphics ‘work for you’ involves doing some extra work that is often unexpected. At the same time, you will need to fit into graphics quality constraints and guidelines that are rarely understood before you actually see a bad example. As a rule of thumb, you should allocate at least the same amount of time and effort for making the figures look good as you would for writing, editing and correcting the same page area of paragraph text.

5.5.1 Visualize all important ideas

The set of figures in your thesis should be comprehensive and complete. For all important ideas, constructions, complicated setups and results there should be a visualization that the reader can refer to in case the text does not paint the

'mental image' sufficiently well. At the bare minimum, you should have at least 3 figures (roughly corresponding to the 3 chapters) that clearly and unambiguously show:

1. the context of the problem you are solving, optionally with e.g. question marks and exclamation marks placed to highlight the problems and research questions
2. the overall architecture of your solution (usually as a diagram with arrows, such as in **??**, ideally with tiny toy examples of the inputs and outputs of each box),
3. the advancement or the distinctive property of your solution, usually in a benchmark plot, or as a clear demonstration and comparison of your results.

5.5.2 Make the figures comprehensible

The figures should be easily comprehensible. Surprisingly, that requires you to follow some common "standards" in figure design and processing. People are often used to a certain form of the visualizations, and (unless you have a very good reason) deviating from the standard is going to make the comprehension much more complicated. The common standards include the following:

- caption everything correctly, place the caption at an expectable position
- systematically label the plots with 'main' titles (usually in boldface, above the plot), plot axes, axis units and ticks, and legends
- lay out the diagrams systematically, ideally follow a structure of a bottom-up tree, a left-to-right pipeline, a top-down layered architecture, or a center-to-borders mindmap
- use colors that convey the required information correctly

  Although many people carry some intuition for color use, achieving a really correct utilization of colors is often very hard without previous experience in color science and typesetting. Always remember that everyone perceives color hues differently, therefore the best distinction between the colors is done by varying lightness of the graphics elements (i.e., separating the data by dark vs. light) rather than by using hues (i.e., forcing people to guess which one of salmon and olive colors means "better"). Almost 10% of the population have their vision impaired by some form of color vision deficiency, most frequently by deuteranomaly that prevents interpretation of even the most 'obvious' hue differences, such as green vs. red. Finally, printed colors look surprisingly different from the on-screen

colors. You can prevent much of these problems by using standardized palettes and well-tested color gradients, such as the ones from ColorBrewer[2] and ViridisLite[3]. Check if your pictures still look good if converted to greyscale, and use a color deficiency simulator to check how the colors are perceived with deuteranomaly.

Avoid large areas of over-saturated and dark colors:

- under no circumstances use dark backgrounds for any graphical elements, such as diagram boxes and tables — use very light, slightly desaturated colors instead
- avoid using figures that contain lots of dark color (as a common example, heatmaps rendered with the 'magma' color palette often look like huge black slabs that are visible even through the paper sheet, thus making a dark smudge on the neighboring page)
- increase the brightness of any photos to match the average brightness of the text around the figure

Remember to test your figures on other people — usually, just asking 'What do you think the figure should show?' can help you debug many mistakes in your graphics. If they think that the figure says something different than what you planned, then most likely it is your figure what is wrong, not the understanding of others.

Finally, there are many magnificent resources that help you arrange your graphics correctly. The two books by Tufte [5, 6] are arguably classics in the area. Additionally, you may find many interesting resources to help you with technical aspects of plotting, such as the `ggplot`-style 'Fundamentals' book by Wilke [7], and a wonderful manual for the TikZ/PGF graphics system by Tantau [8] that will help you draw high-quality diagrams (like the one in **??**).

5.6 What is a discussion?

After you present the results and show that your contributions work, it is important to *interpret* them, showing what they mean in the wider context of the thesis topic, for the researchers who work in the area, and for the more general public, such as for the users.

Separate discussion sections are therefore common in life sciences where some ambiguity in result interpretation is common, and the carefully developed

[2] `https://colorbrewer2.org`
[3] `https://sjmgarnier.github.io/viridisLite/`

intuition about the wider context is sometimes the only thing that the authors have. Exact sciences and mathematicians do not need to use the discussion sections as often. Despite of that, it is nice to position your output into the previously existing environment, answering:

- What is the potential application of the result?
- Does the result solve a problem that other people encountered?
- Did the results point to any new (surprising) facts?
- How (and why) is the approach you chose different from what the others have done previously?
- Why is the result important for your future work (or work of anyone other)?
- Can the results be used to replace (and improve) anything that is used currently?

If you do not know the answers, you may want to ask the supervisor. Also, do not worry if the discussion section is half-empty or completely pointless; you may remove it completely without much consequence. It is just a bachelor thesis, not a world-saving avenger thesis.

A. Using CoolThesisSoftware

Use this appendix to tell the readers (specifically the reviewer) how to use your software. A very reduced example follows; expand as necessary. Description of the program usage (e.g., how to process some example data) should be included as well.

To compile and run the software, you need dependencies XXX and YYY and a C compiler. On Debian-based Linux systems (such as Ubuntu), you may install these dependencies with APT:

```
apt-get install \
  libsuperdependency-dev \
  libanotherdependency-dev \
  build-essential
```

To unpack and compile the software, proceed as follows:

```
unzip coolsoft.zip
cd coolsoft
./configure
make
```

The program can be used as a C++ library, the simplest use is demonstrated in listing 1. A demonstration program that processes demonstration data is available in directory `demo/`, you can run the program on a demonstration dataset as follows:

```
cd demo/
./bin/cool_process_data data/demo1
```

After the program starts, control the data avenger with standard `WSAD` controls.

Listing 1 Example program.

```
#include <CoolSoft.h>
#include <iostream>

int main() {
  int i;
  if(i = cool::ProcessAllData()) // returns 0 on error
    std::cout << i << std::endl;
  else
    std::cerr << "error!" << std::endl;
  return 0;
}
```