# DISCLAIMER

This project is of R&D nature and hence an accurate estimate of times cannot be predicted. However, ballpark figures are arrived at from previous experiences and estimations of coding requirements. This is an indicative costing based on limited testing of some methods and algorithms. Having not seen the samples of actual data, these algorithms may be way off course and hence the efforts to develop alternative methods may cost more time. Every effort will be made to stay within the estimates but cannot guarantee it. As with any R&D project there is a risk, albeit low, that the expected outcome may not be reached precisely and hence some adjustments may become necessary.

11/26/2019

Arapaut Sivaprasad, Ph.D.

WebGenie Software Pty Ltd.

[avs2904@webgenie.com](mailto:avs2904@webgenie.com); [avs\_webgenie\_com@me.com](mailto:avs_webgenie_com@me.com)

<https://www.linkedin.com/in/dr-siva/>

m: 0411 132 646

**STLP Project Costing**

HLD, Software, Hosting, Maintenance and Documentation

# INTRODUCTION

This document is an indicative costing based on the following general categories. The number of hours for each task is estimated along with the approximate time-frame in which to finish those tasks. A summary (Ref. email from Gavin) is in **Table 1**, and details are given below it.

* Project proposal converted to a high-level design document
* Research and Development of the best algorithm
  + Natural Language Processing (“NLP”)
  + Heuristics-based classification (“Heuristics”)
* Coding and testing
* Tech deliverables including:
  + source code
  + method/tech to host the solution
  + tech maintenance document documenting the source code/system and the components of the solution

# SUMMARY

## Table 1 – Cost and time-frame

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **Item** | **Cost\***  **(hours)** | **Time-frame**  **(day or week)** | **Comments** |
| 1 | High Level Design Document (HLD) | 5 | 1 day | This is the first step before commencing the project. The HLD must be reviewed and signed-off. |
| 2 | R&D | 20 | 1 week | Use sample data to determine the best approach. Include studying NLP and heuristics methods. |
| 3 | Coding and Testing | 80 | 4 weeks | Applying the algorithms in a coding and testing environment. It will involve testing and re-coding to reach optimum level of confidence. It will not include the setup and usage of third-party software (e.g. PDF-to-text converter, image analyser, etc.) which are assumed to be in place already. |
| 4 | Source code delivery | 20 | 1 week | Commenting the code, packaging and delivery for installation on production server. It will not include porting the code into another language. |
| 5 | Tech documentation | 40 | 2 weeks | External documentation of code, installation and maintenance instructions. |
|  | **TOTAL\*\*** | **165** | **8 weeks** | Can only commit 20 hours per week but will try to put in more time and finish the project in 4-5 weeks. |
| \* Estimated times are not cut and dry and may carry over to previous or next steps.  \*\* Significant changes to the HLD may alter these estimates. Total time estimate is only a best guess at this time and may have to be revised with mutual consent. | | | | |

# DETAILS

Given below are details of the intended methodology and procedures but these are neither exhaustive nor unchangeable. Some are just thoughts and may be modified by using feedback from the stakeholders. Feedback based on experience may help to reduce the overall times.

I am happy to **reconsider if the quotes**, which are based on my experience with comparable projects, appear to be much more than anticipated.

Unlike a normal commercial R&D project, it is expected that the current project will not consume several hundred hours. But, at the same time, it is not a standard project like building a web site or server‑side software using code base already developed.

## Project proposal converted to a high-level design document

The **HLD** will form the basis of the project and will be as detailed as possible to begin with. It must be **studied and approved** before the work begins. While small changes and additions are normal during later stages, major changes in strategy or features will result in a re-visit of the time costs.

Will attempt to stick to the **milestones**, but it cannot be guaranteed. At times it may become necessary to back track or go sideways to study/fix something. It can affect the milestones. Similarly, the total times given in **Table 1** are also indicative. Small variations up or down must be expected. **Total costs will not be changed without mutual agreement.**

## Research and Development of the best algorithm

It is expected that **Python’s NLP** methods will form the major technology to compare documents against each other or against a reference. Preliminary testing has shown that two related web pages can be matched at **90-98%** accuracy. However, unrelated documents were giving matches between 40-91%. It is therefore necessary to use other criteria to diverge the scores. Significant **R&D** is required to achieve a consistently **high confidence score**.

The above, however, may not apply with the documents in store. For example, two documents may significantly vary in content but are still in the same category (e.g. security incident reporting). NLP will not categorise them into one. In such cases we will have to use other methods to compare them. Heuristics based on other characteristics (e.g. title, source, logo, images, etc.) may help. Arriving at appropriate **heuristics** will require testing with several hundred documents in different categories.

## Coding and testing

Like in all major developments, coding, testing and re-coding will take the bulk of the time. Since we are dealing with a wide variety of documents, a method suitable for one type may not be suitable for another. We may have to **empirically arrive at the best method** and then test it extensively to prove the hypothesis.

Coding will mostly be on a laptop and hence may be inherently slower. Will try to do the coding on a server, where both the speed and subsequent deployment on production could be faster.

## Tech deliverables

Depending on the language used the source code may have to be compiled (e.g. **C++, GO**) or interpreted (e.g. **Python, Perl**). For maintainability it is planned to use interpreted languages. Of these, Python is the language of choice due to its robust **NLP capabilities**, and Perl or GO may be used for **parallelisation**. Every effort will be made to create easy-to-understand and maintainable code that are platform-independent.

Delivery of the code will be **in the language it is written**. This may cause issues for maintenance if there is no in-house expertise in those languages. Porting the code to a language (e.g. **C#**) or framework (e.g. **Dot-NET**) may be considered but I am not experienced in these yet.

There are two ways to go if it is essential to port the code, viz., (1) **assist a Dot-NET expert** to understand the Python/Perl code or (2) learn C# and Dot-NET myself. Option1 could be faster and cost-effective than option2, but I am happy to **learn a new language/framework**. A major hindrance to porting will be the **unavailability of NLP** libraries in other languages as in Python. It is, however, possible to retain the NLP functions in Python and use a different language for the main program.

## Method/tech to host the solution

It is imperative that the code is installed on a **remotely accessible server**. I do not expect major issues with it if there are Linux servers available. At present, however, I have hit some issues with installing the required Python libraries on my Linux server which is a few years old.

There will be two methods of operation on the servers: (1) interactively via a **web interface** and (2) **command-line** or **batch** execution from shell. Both methods will be installed. The command-line interface will not require external expertise to install. Neither will a **functional web interface**, but a fancy web interface will require the services of a graphic and/or web designer. My graphic skills are next to nothing. The cost of an external designer is **not included** in the estimate.

## Tech maintenance document

Detailed documentation in **Word/PDF, PowerPoint, HTML** or **Confluence** (in only one format, not all) shall be prepared within the time quoted.

# FEEDBACK AND SIGN-OFFS

It is vital that periodic review of progress and sign-offs at milestones are conducted. While this may pose a problem with the busy work schedule of the stakeholders, absence of it may result in an unsatisfactory product and/or conflict.

I shall try to minimise the amount of material to be reviewed and plan the work such that I will not be waiting for the feedback before proceeding. There may still be occasions when a feedback and/or sign-off is necessary before proceeding. We must therefore have a mechanism in place. I suggest that **two working days** are given for an essential **feedback/sign-off**. In its absence I shall proceed as if it has been signed off. Alternatively, I shall suspend the main work until the feedback/sign-off arrives. During this period, I shall go over the completed work for corrections, documentations, etc. This will, obviously, extend the time cost estimates in Table 1.

# PAYMENTS

I am flexible with any payment plan suitable to the stakeholders. Payment, at the rate previously intimated by the agency, for the total hours spent on each week is preferred. In some weeks it could be more than 20 hours and in others (unlikely) less. If a fixed payment for 20 hours per week is preferred by the stakeholder, it is fine too. I will, however, try to finish the work ahead of the time-frames quoted in Table 1 in which case the payments may have to continue for a time even after the work is finished.

# NOTE ADDED FOR CLARITY

This costing is based on my trials using sample data from web pages and a specific algorithm. Even if the data comes from other sources, if it is not too different in structure and content, modified versions of the same algorithm will work. If, however, the data is very different structurally and content-wise it will require different algorithms and the estimated times will vary.

Also, to be noted is that the costing is based on producing the first working version. The resulting software will be functional but not necessarily market-ready. Additional enhancements and fine tuning will be required. Such work could be done by an in-house person who is technically competent or by an external programmer. It is my hope that I will be the one doing those as well.

\_\_\_\_\_\_\_\_\_\_\_END OF DOCUMENT\_\_\_\_\_\_\_\_\_\_\_