

Solving a research problem with open source software - A Story of Collaboration

Gabriel Couture¹
gabriel.couture@dti.ulaval.ca

¹Université Laval



canarie

Our Team

- We make calls for scientific software ideas
- Researchers submit their idea and we choose the bests
- Ideas with the greatest ratio of impact/ work win



Florent Parent

Responsable administratif



Philippe Després

Chercheur pivot



Félix-Antoine Fortin

Chef d'équipe de développement



Gabriel Couture

Développeur de logiciels scientifiques



Olivier Chouinard-Barville

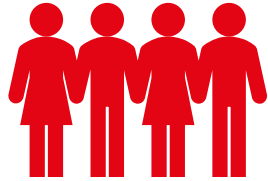
Développeur de logiciels scientifiques



Bruno Lavoie

Conseiller technologique - VALERIA

Context



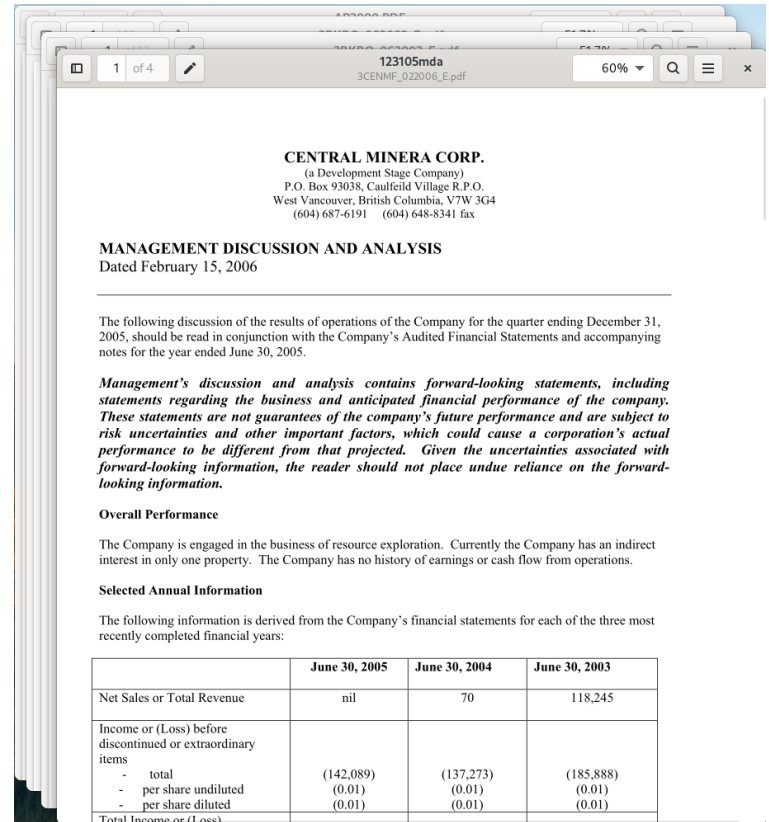
A Researcher from **Université Laval's**
Faculty of administration contacted us



Need to analyse **large corpora** of
financial documents

The Research Project

- Compare the quality of financial reports of federal corporations
 - In particular , between the English and french version
- Study the evolution of prevalence of terms related to the environment



CENTRAL MINERA CORP.
(a Development Stage Company)
P.O. Box 93038, Caulfield Village R.P.O.
West Vancouver, British Columbia, V7W 3G4
(604) 687-6191 (604) 648-8341 fax

MANAGEMENT DISCUSSION AND ANALYSIS
Dated February 15, 2006

The following discussion of the results of operations of the Company for the quarter ending December 31, 2005, should be read in conjunction with the Company's Audited Financial Statements and accompanying notes for the year ended June 30, 2005.

Management's discussion and analysis contains forward-looking statements, including statements regarding the business and anticipated financial performance of the company. These statements are not guarantees of the company's future performance and are subject to risk uncertainties and other important factors, which could cause a corporation's actual performance to be different from that projected. Given the uncertainties associated with forward-looking information, the reader should not place undue reliance on the forward-looking information.

Overall Performance

The Company is engaged in the business of resource exploration. Currently the Company has an indirect interest in only one property. The Company has no history of earnings or cash flow from operations.

Selected Annual Information

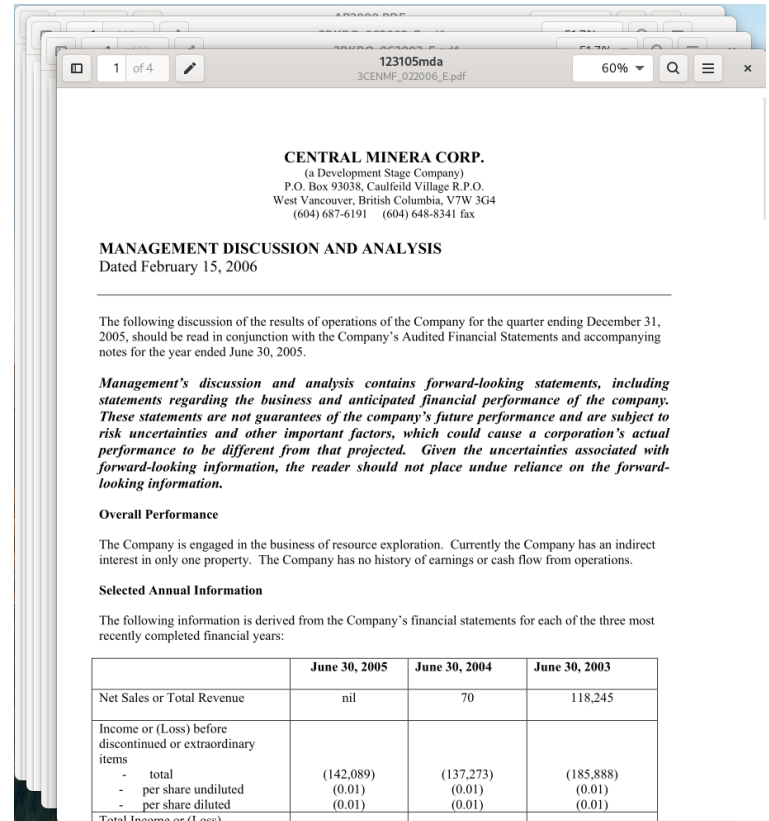
The following information is derived from the Company's financial statements for each of the three most recently completed financial years:

	June 30, 2005	June 30, 2004	June 30, 2003
Net Sales or Total Revenue	nil	70	118,245
Income or (Loss) before discontinued or extraordinary items			
- total	(142,089)	(137,273)	(185,888)
- per share undiluted	(0.01)	(0.01)	(0.01)
- per share diluted	(0.01)	(0.01)	(0.01)
Total Income or (Loss)			

Technical needs

- Find the frequency of common terms
- Compute a set of readability scores
- Involve 80,000 financial PDF documents
 - Ranging from one to a few hundred pages
- E.g. Dale-Chall score formula

$$0.1579 \left(\frac{\text{difficult words}}{\text{words}} \times 100 \right) + 0.0496 \left(\frac{\text{words}}{\text{sentences}} \right)$$



The screenshot shows a PDF document titled "123105mda" from "3CENMF_022006_E.pdf". The document is for "CENTRAL MINERA CORP. (a Development Stage Company)" with contact information: "P.O. Box 93038, Caulfield Village R.P.O. West Vancouver, British Columbia, V7W 3G4 (604) 687-6191 (604) 648-8341 fax". The section is "MANAGEMENT DISCUSSION AND ANALYSIS" dated February 15, 2006. The text discusses the results of operations for the quarter ending December 31, 2005, and includes a disclaimer about forward-looking statements. It also has sections for "Overall Performance" and "Selected Annual Information".

CENTRAL MINERA CORP.
(a Development Stage Company)
P.O. Box 93038, Caulfield Village R.P.O.
West Vancouver, British Columbia, V7W 3G4
(604) 687-6191 (604) 648-8341 fax

MANAGEMENT DISCUSSION AND ANALYSIS
Dated February 15, 2006

The following discussion of the results of operations of the Company for the quarter ending December 31, 2005, should be read in conjunction with the Company's Audited Financial Statements and accompanying notes for the year ended June 30, 2005.

Management's discussion and analysis contains forward-looking statements, including statements regarding the business and anticipated financial performance of the company. These statements are not guarantees of the company's future performance and are subject to risk uncertainties and other important factors, which could cause a corporation's actual performance to be different from that projected. Given the uncertainties associated with forward-looking information, the reader should not place undue reliance on the forward-looking information.

Overall Performance

The Company is engaged in the business of resource exploration. Currently the Company has an indirect interest in only one property. The Company has no history of earnings or cash flow from operations.

Selected Annual Information

The following information is derived from the Company's financial statements for each of the three most recently completed financial years:

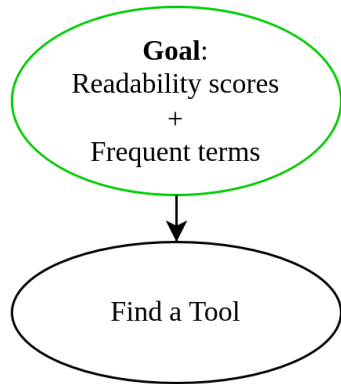
	June 30, 2005	June 30, 2004	June 30, 2003
Net Sales or Total Revenue	nil	70	118,245
Income or (Loss) before discontinued or extraordinary items			
- total	(142,089)	(137,273)	(185,888)
- per share undiluted	(0.01)	(0.01)	(0.01)
- per share diluted	(0.01)	(0.01)	(0.01)
Total Income or (Loss)			

Our Goals

Goal:
Readability scores
+
Frequent terms

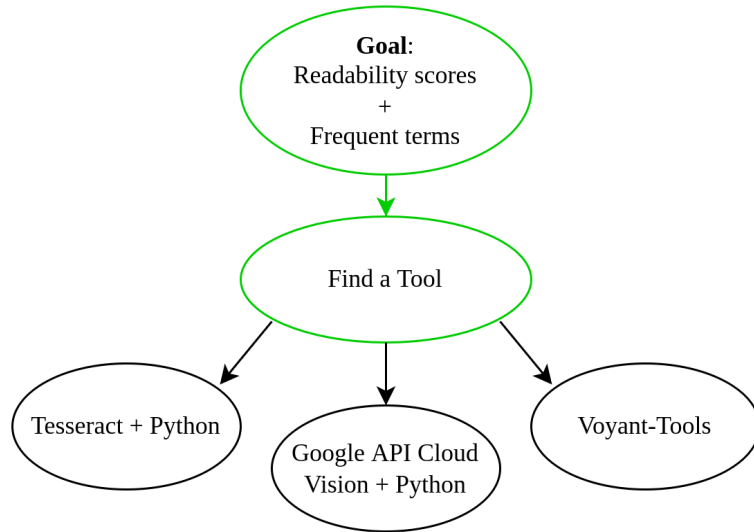
- Get a set of readability scores of PDF documents
- Get Frequent Terms of each PDF document
- (Bonus) Allow researchers to easily run the analyzes themselves

Find a software (or build one?)



- We didn't want to build one ourselves
 - We want to avoid maintaining it
 - Probably involve more work
 - Will likely **never be re-used**
- Find/ contribute to a pre-existing software is preferable
 - No maintenance needed
 - Maximize **re-use**
 - (Probably) less work

Weigh the pros and cons of each method



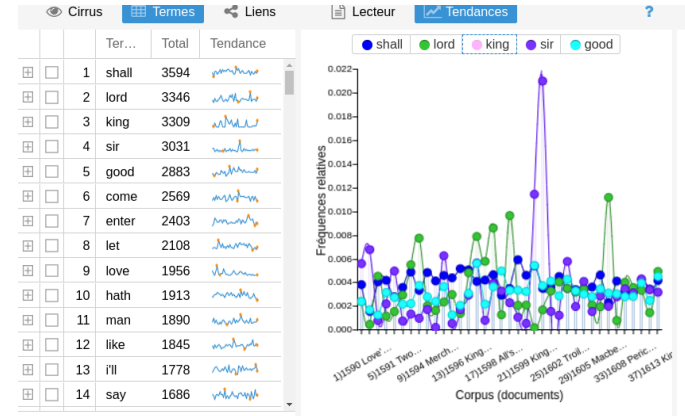
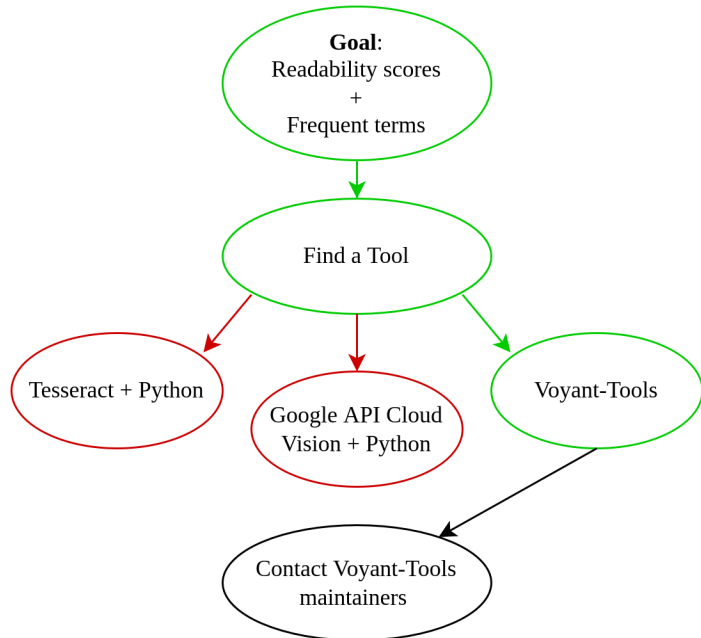
3 tools/ methods that could solve our problem

- Tesseract + Python
- Google API Cloud Vision + Python
- Voyant-Tools
 - [https://voyant -tools.org/](https://voyant-tools.org/)

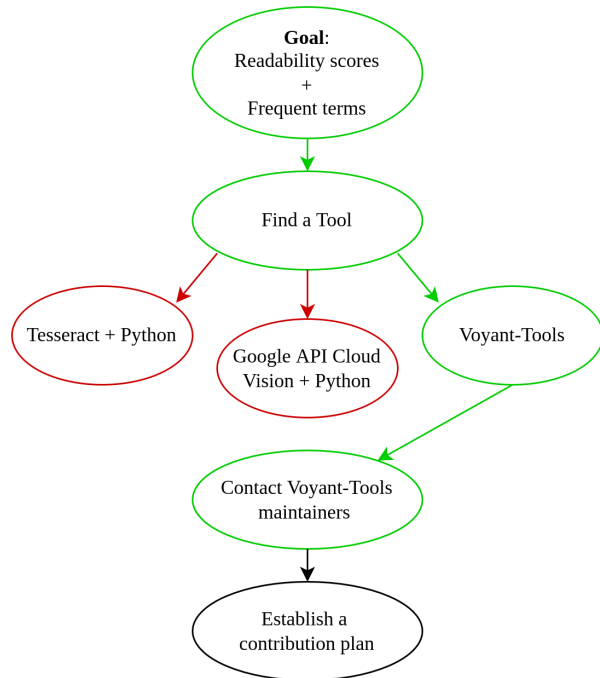
We choose Voyant-Tools



- Has a nice UI and is open source
- It nearly has all the functionalities we need
 - Just missed the readability scores
- However: likely to encounter scalability issues

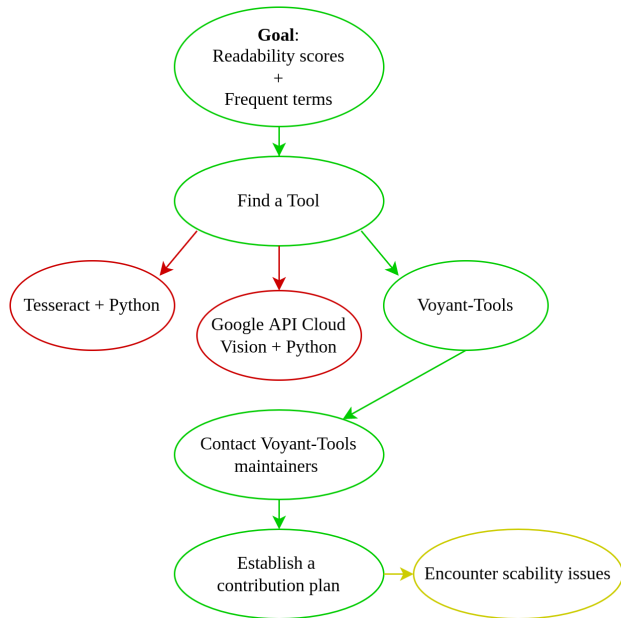


Establishing a contribution plan



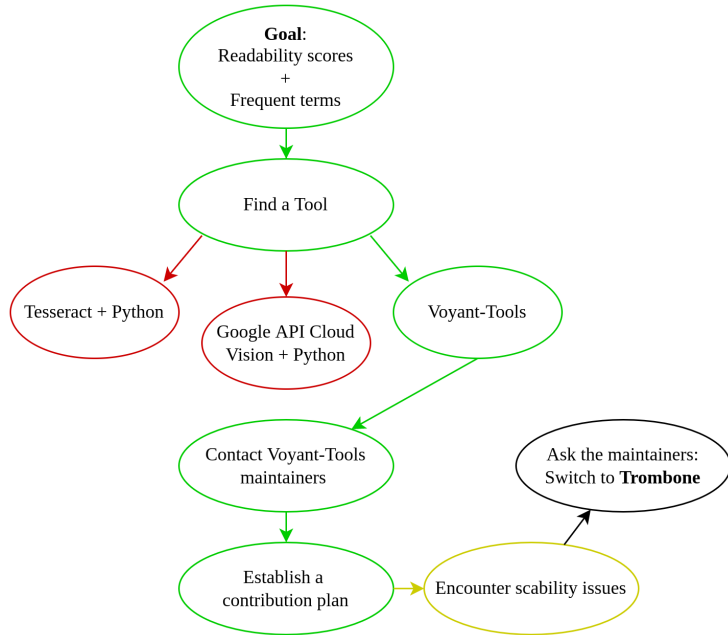
- We contacted the maintainers to validate their interest in our additions
- Basically: we would like to add several readability scores to Voyant-Tools
- They agreed :)
 - We made a plan

Encounter scability issues



- We did some load tests
 - Can't handle more than ~3000 PDF
 - (Not bad actually!)
- It was not a big surprise
 - We expected to use Voyant-Tools backend rather than use the web page

Switching to Trombone



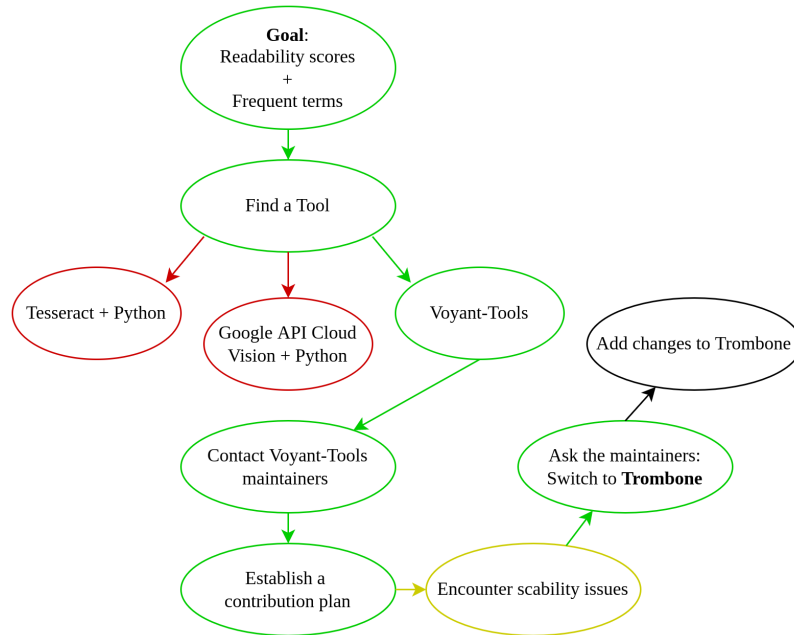
- We ask the maintainers:
 - They suggest we use Trombone
- Trombone is used by Voyant-Tools backend
 - <https://github.com/voyanttools/trombone>
- We contribute a way to run Trombone as a CLI

Adding instructions to build an executable jar and related documentation



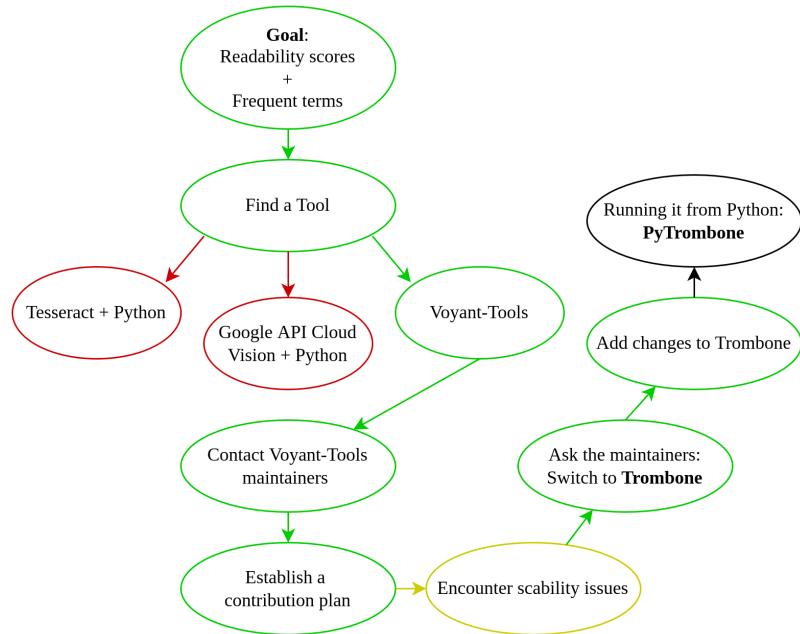
gacou54 committed on Sep 2, 2021

Our contributions



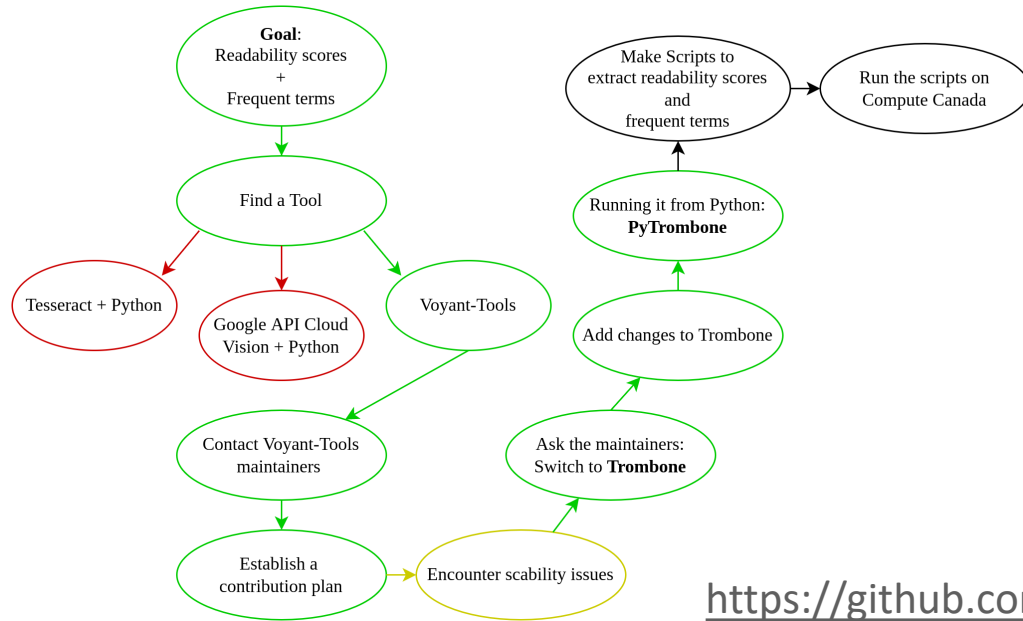
- We added the following readability scores
 - Dale-Chall Index
 - SMOG Index
 - Gunning FOG Index
 - Coleman-Liau Index
 - Automated Readability Score
 - LIX Score
- Each of them was added and validated through a PR onGithub

Access Trombone from Python



- We have software with the desired functionalities
 - We now want to use it in batches
- We know Python, the researcher knows Python, let's use Python!
 - We make **PyTrombone**, a Python wrapper for Trombone
 - <https://pypi.org/project/pytrombone/>

Running the analyzes

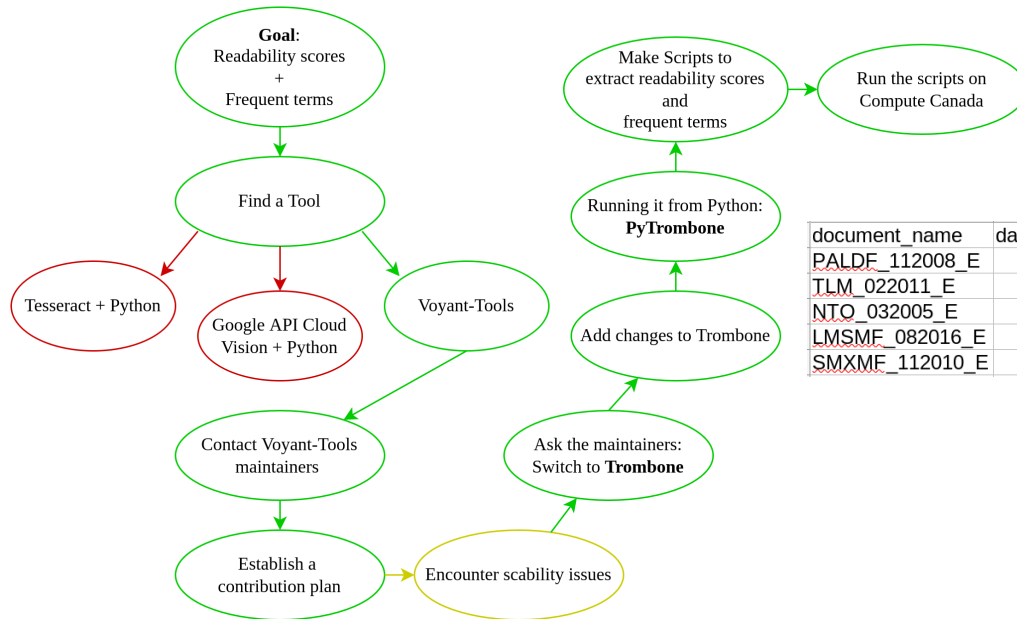


- Now we have all the tools we need
- We made scripts to
 - Calculate the readability scores
 - Get the frequent terms
- We used Compute Canada
 - To have a reusable method for researchers

<https://github.com/ulaval-rs/trombone-scripts>

We have the Readability scores

And the frequent terms



We finally have the researcher's results!



document_name	daleChallIndex	colemanLiauIndex	smogIndex	lixIndex	automatedReadabilityIndex	fogIndex
PALDF_112008_E	11.4	15.4	18.3	65.5	20.2	21.6
TLM_022011_E	12.2	16.0	19.4	67.4	20.3	22.8
NTQ_032005_E	10.8	13.4	17.3	58.6	17.1	20.1
LMSMF_082016_E	11.0	12.9	17.2	57.6	17.3	20.1
SMXMF_112010_E	11.9	13.9	17.3	62.7	18.9	20.5

Review of our goals

- Get a set of readability scores for PDF documents
 - Check
- Get Frequent Terms of each PDF document
 - Check
- (Bonus) Allow researchers to easily run the analyzes themselves
 - Meh, PyTrombone is fine

Conclusion

While solving the researcher's problem:

- We contributed to an open-source project rather than start our own
 - This maximizes the reuse of our work
 - No maintenance
 - The maintainers really help over many parts of the project

