Pipeline effectiveness in Sketch Engine

Matúš Kostka

Lexical Computing CZ
Botanická 554, 602 00 Brno-Královo Pole

December 10, 2022

Overview

1. Reasons and goals
2. Measurement
3. Results (charts in different documents)
4. Conclusion
5. Conclusion

Reasons and goals

- Measure the most used pipelines in Sketch Engine.
- Measured parameters: **Execution time**, **CPU usage**, **Max RSS**.
- Create a tool for the future measurements of pipelines.
- Originally was to measure pipelines with 1 (initialization time), 10,000; 100,000; 1,000,000 tokens (later remeasured with more sizes).
- Analyze the result and calculate linear regression.

Tool and data used for measurements

- Bash.
- Compressed prevertical files from **wikipedia** measured in 2020 and 2021.
- Measured on machine with **32** cores and **256** GB RAM.

Overall result for 10,000 tokens

	Min value	Max value	Average	Median
Execution time (min)	0.04	12.71	1.30	0.90
CPU usage (%)	0	100	26	18
RAM usage (GB)	0.007	2.326	0.252	0.141

Minimim: Hebrew (tok1), Hebrew (yap), Thai.
Maximum: Tagalog, Japanese, Tagalog.

Overall result for 100,000 tokens

	Min value	Max value	Average	Median
Execution time (min)	0.07	55.00	4.51	1.81
CPU usage (%)	0	127	40	38
RAM usage (GB)	0.008	5.443	0.388	0.187

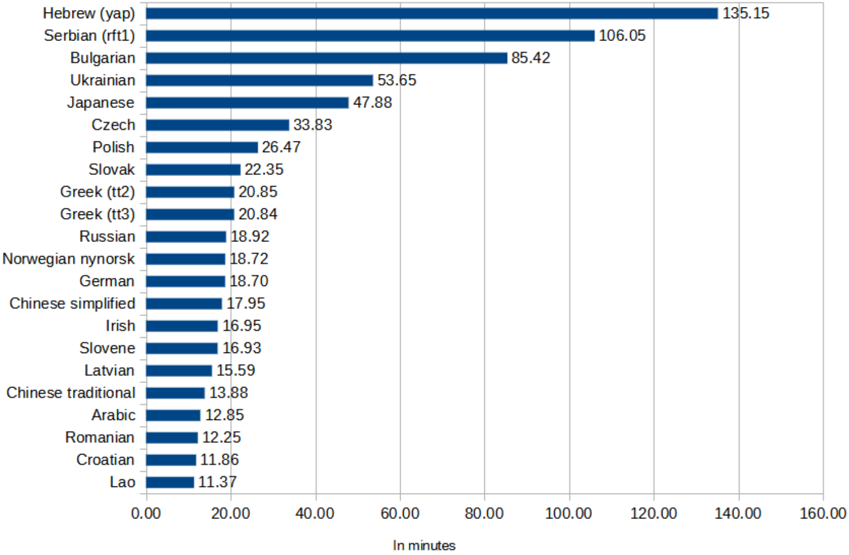
Minimim: universal, Hebrew (yap), Thai.
Maximum: Tagalog, Bulgarian, Tagalog.

Overall result for 1,000,000 tokens

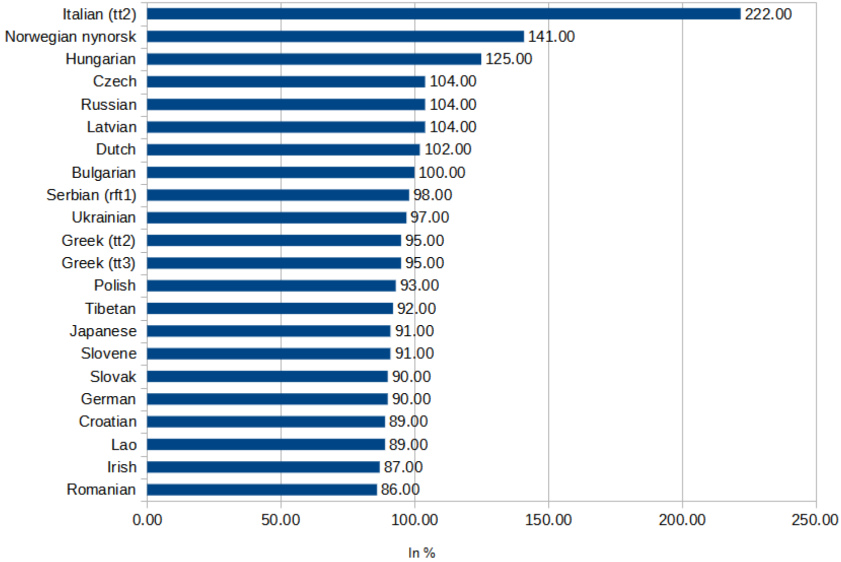
	Min value	Max value	Average	Median
Execution time (min)	0.39	135.15	17.00	6.59
CPU usage (%)	0	222	75	77
RAM usage (GB)	0.008	5.629	0.733	0.209

Minimim: universal, Hebrew (yap), Thai.
Maximum: Hebrew (yap), Italian (tt2), Japanese.

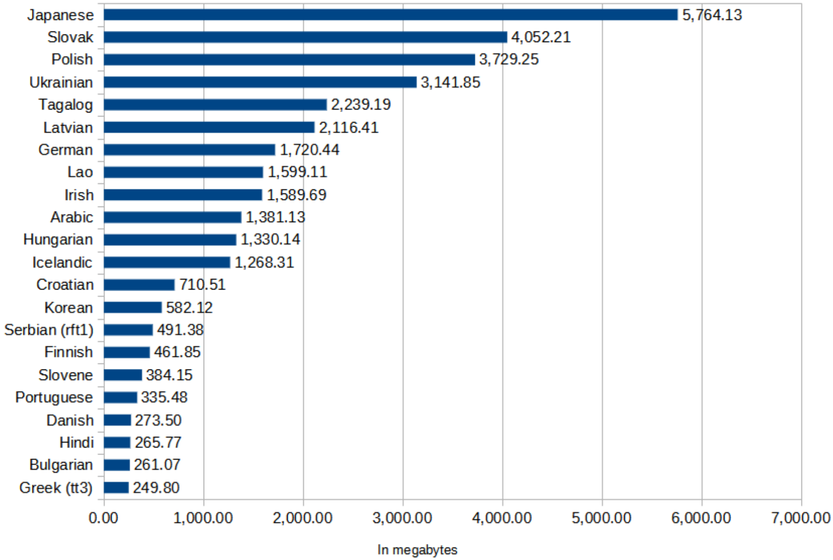
Execution time 1,000,000 tokens



CPU usage 1,000,000 tokens

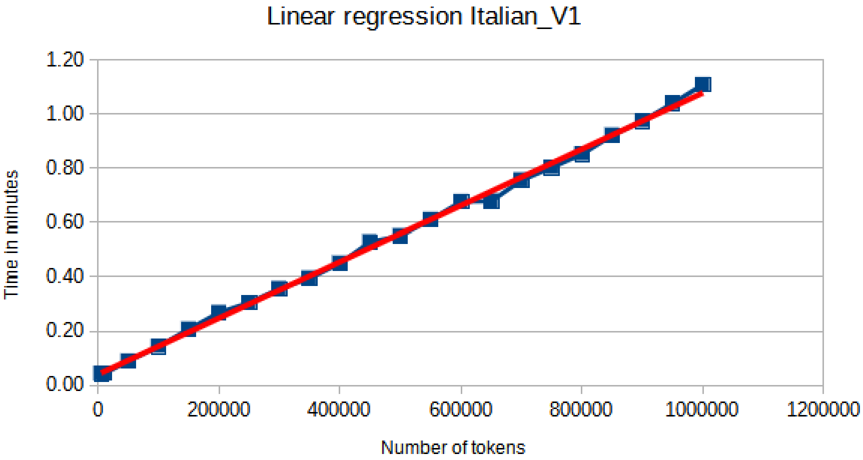


RAM usage 1,000,000 tokens



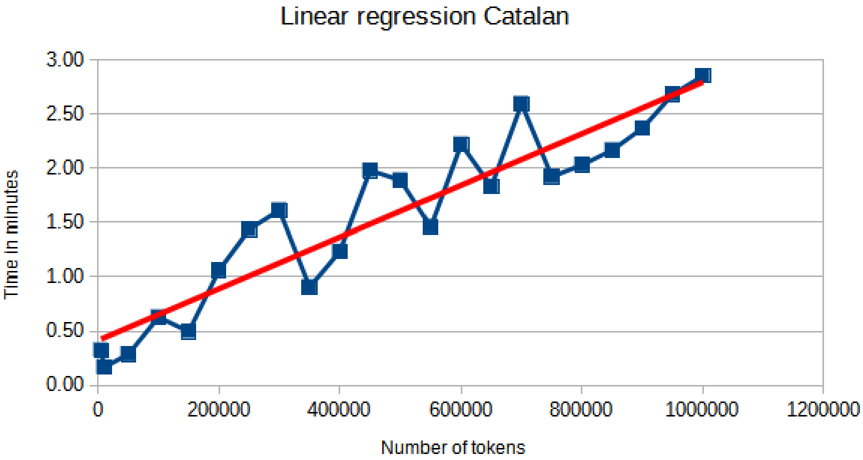
Linear regression

Slope: 0.000001039 **Intersection:** 0.039905863 **Error rate:** 0.91182 sec



Linear regression

Slope: 0.000002383 **Intersection:** 0.41385 **Error rate:** 18.47 sec



Conclusion

- Tagalog pipeline. (most problematic in 1M)
- Results depend on supported features by pipeline. (uninorm, unitok, lemmatizer, treetagger).
- LR Average error rate is **58,47** seconds.
- LR Median error rate is **24,39** seconds.
- Pipelines with differ alphabet from Latin are slower (in most cases).
- In 1,000,000 measure, **43%** of are slower than 10 minutes.

Future work

- Remeasure suspicious or failed pipelines.
- Keep data up to date.

Thank you for your attention.