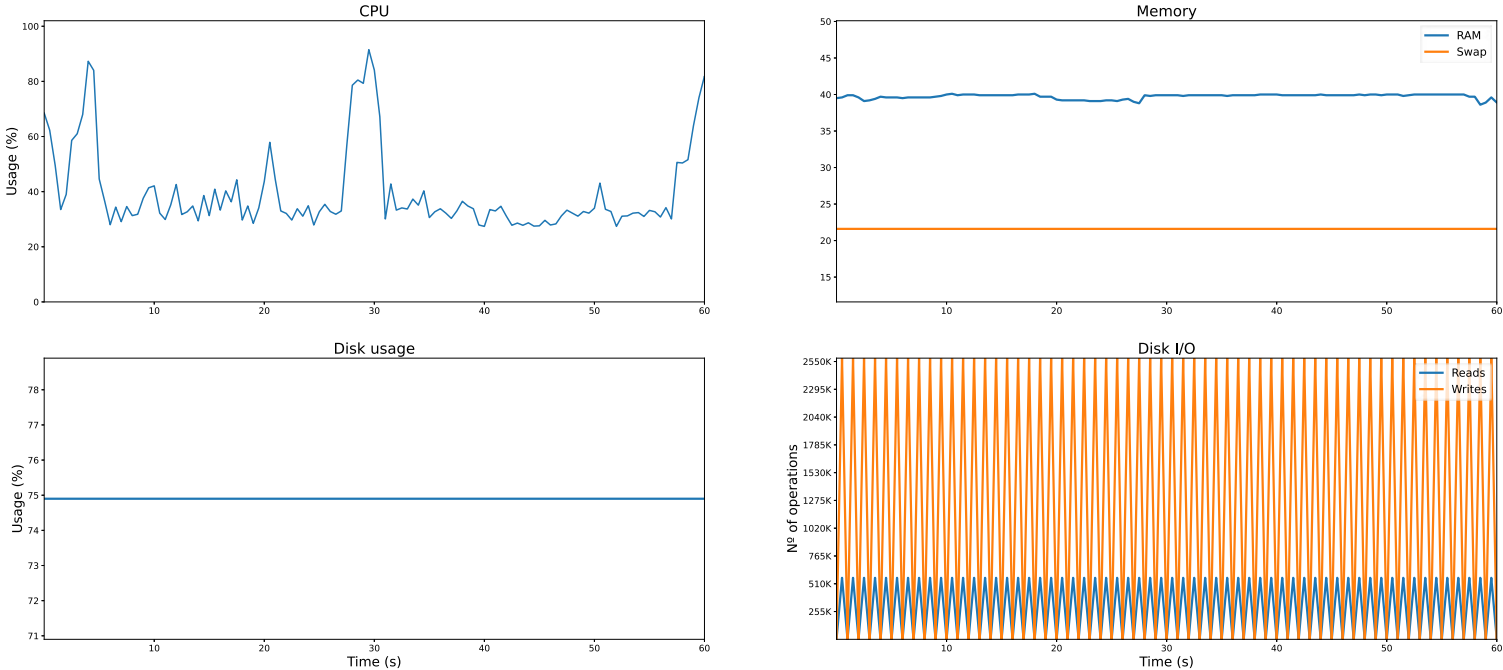# Execution report

This report provides of a brief summary of how the execution of the pipeline went.

Resources used

The CPU, memory and disk usage during the execution of the pipeline are reflected in the following graphs:

The average CPU, memory and disk usages during the execution are the following:

Resource	Average use (%)
CPU	41.51
RAM memory	39.72
Swap memory	21.6
Disk	74.9

The total number of reads and writes to the disk during the execution are the following:

I/O action	Total done
Reads	36697293
Writes	167762055

Pipeline run time

The run times for the different stages (in seconds) are the following:

Stage	Run time (s)
Preprocessing	5.08
Obtain text processors	48.75
Train the models	61.77
Run the predictions	1.05
Evaluate the results	5.65
Analyze the features	0.98
Total time needed	123.27

Size of the data used

Size of the train, validation and test datasets used in bytes.

Data set	Size (MB)
Train	6.86
Validation	2.07
Test	0.99

Metrics obtained for the models

Values for the different metrics obtained for the two models tested.

Metric	Bag of words	TF-IDF
Accuracy	0.36	0.37
F1 Score	0.66	0.65
Precision	0.36	0.35
ROC	0.93	0.93

Feature analisis

The model that performed the best according to most metrics is the Bag of Words.

The most seen words in Bag of Words were the following:

- objective
- php
- template
- swift
- begin

and the least seen were:

- gtk
- fscanf
- scanf
- malloc
- c

For the TF-IDF model, the top positive words were:

- fscanf
- printf
- malloc
- scanf
- c

and the top negative were:

- php
- javascript
- java
- python
- objective c