Scaling NewSum - Big data Text Clustering & Summarization using N-Gram graphs

GOALS

Business related goals

- improve the quality of the solutions our product offers
- allow NewSum technology to expand to new domains/markets

Technical goals

Measure and evaluate:

- the accuracy of candidate clustering components,
- the effectiveness (summary quality) of alternative summarization components
- the overall scalability of the system

CHALLENGES

Challenges related to business

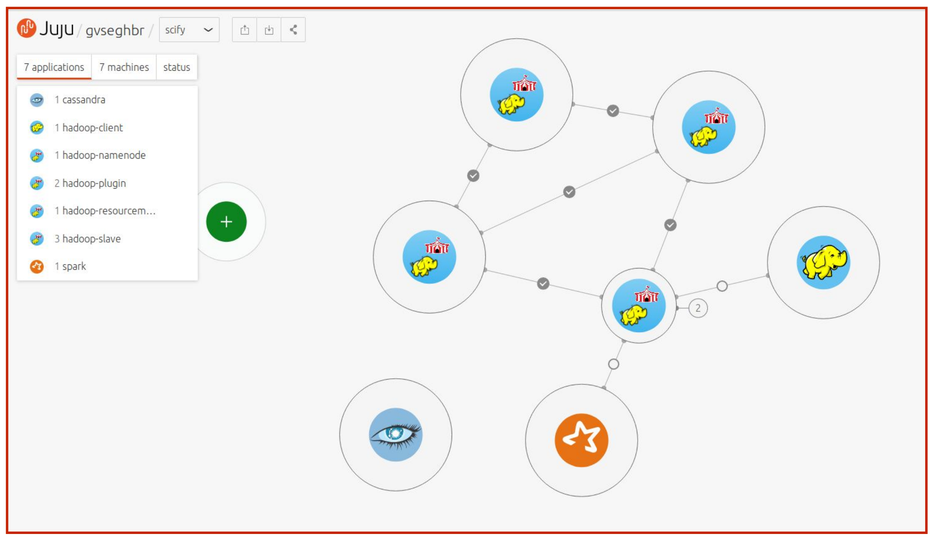
- Expansion to new markets should take domain specific characteristics into account as system parameters
- A product manager is not able to configure the product-related settings appropriate for each domain, so a semi-supervised process would be invaluable

Technical Challenges

- Define a process for evaluating different clustering and summarization components
- Scale the algorithms to process thousand of sources/articles

DEMO SETUP

The experiments run at the Tengu testbed
with the support of IMEC
- Cassandra - Hadoop - Spark



RESULTS

Experiment set 1:
Measure effectiveness of NewSum's candidate clustering implementations

Related datasets:
Multiling (articles with clustering information)
6GB database of news articles

Methodology:
Run clustering on 2 different clustering implementations and measure recall and precision.

Automatic evaluation for MultiLing dataset
Manual process for news articles dataset

Results:
Selected the algorithm with higher precision & recall

MORE RESULTS

Experiment set 2:
Measure scalability

Related dataset:
6GB database of news articles

Methodology:
Run the clustering pipeline using as input a) the algorithm from experiment set 1 b) a variable number of articles.
Measure speed

Results:
Increased 5 times the speed of the clustering pipeline!
Identified areas of improvement

Experiment set 3:
Measure effectiveness of NewSum's candidate summarization implementations

Related datasets:
6GB database of news articles

Methodology:
Run the summarization pipeline using as input a) configuration/parameter setting b) a number of clusters to be summarized.
Recall and precision were measured through a manual process.

Results:
Implemented/Identified the process for selecting the algorithm appropriate for each scenario

CONCLUSIONS

What we achieved:

- Defined a process for evaluating clustering algorithms
- Defined a process for evaluating summarization components
- Increased 5 times the speed of the clustering pipeline!
- Measured scalability and identified bottlenecks

How Fed4Fire+ helped us

- Patron's support was crucial to the success of the experiments
- Provided a quick way to start experimenting with big data without having to worry about the underlying technologies
- Funding allowed us allocate time to implement the algorithms and analyze next steps

POST MORTEM

Next steps:

- Continue working on algorithm implementations
  Distributed N-gram graphs
  Improve clustering speed using blocking methodology

- Automate the set up of a pipeline in a cloud environment to be used in production.

- Release a domain specific product related to Blockchain news.