SPERG: Scalable Political Event Report Geoparsing in Big Data

Aswin Krishna Gunasekaran, Maryam Bahojb Imani,
Latifur Khan, Christan Grant,
Patrick T. Brandt, Jennifer S. Holmes

Computer Science Department
The University of Texas at Dallas

XSEDE
Extreme Science and Engineering Discovery Environment

Introduction

- Digital newspaper archives serve as an easily accessible, rich source of information to conduct analytic studies.
- Extracting geographic coordinates solely from text descriptions is a process known as geoparsing.
- Major challenge is to geoparse massive corpora like newspaper archives.

Objective: Extract the latitude-longitude information of primary focus locations of archived news reports on political events.

Primary Focus Location

- News articles contain stories about people, events and *places*.
- An event can occur only at a single location. We aim to extract this location among focus locations, and call this as *primary focus location*.
- Some applications:
  - determining crime pattern locations;
  - predicting the place of protests and political unrest;
  - identifying the geolocation of natural disasters.

A News Report Example

Primary Focus Location:
Bali

Other locations:
Canada, Indonesia, Earth

Objective: We wish to obtain the lat-lon of Bali.

What we will do in the next two, three years will determine our future. This is the defining challenge. This call to action came last month from Rajendra Pachauri, chair of the Intergovernmental Panel on Climate Change, after that group of global scientists warned humanity's very survival is at stake if we don't apply a cold compress to Earth's increasingly sweaty brow. Tomorrow is the first day of that brief time to get our act together. In Bali, Indonesia, delegates from Canada and about 180 other countries, or parties, will launch the 13th annual United Nations conference on climate change. Judging by the most optimistic predictions of the outcome, it's safe to say the gulf between doomsday rhetoric and diplomatic dithering couldn't be wider. No matter what happens, most observers agree, Canada will no longer be a major player. Just as each conference since 1995 has resulted in an agreement of sorts, this 12-day event will produce something. The UN itself sets the stage by defining the goals in advance. Sometimes, though - and this is one of them - the UN erects such a huge net that delegates can score with even a very weak, misdirected shot.

Related Works

- Based on the availability and performance of various geoparsers, we focus on Profile[1], Cliff-Clavin[2] and mordecai[3] for our study.
- Below table potrays the functionality of these geoparsers.

Geoparsers	NER Extraction	Coordinates Extraction	Focus Location	Primary Focus Location	Multi-Lingual
Cliff-Clavin	Stanford CoreNLP	✓	✓	✗	✗
Mordecai	MITIE	✓	✗	✗	✗
Profile	MITIE	✗	✓	✓	✗
Profile (modified)	MITIE	✗	✓	✓	✓

[1] M. B. Imani, S. Chandra, S. Ma, L. Khan, and B. Thuraisingham, “Focus location extraction from political news reports with bias correction,” in Big Data (Big Data), 2017 IEEE International Conference on. IEEE, 2017, pp. 1956–1964.
[2] C. D'Ignazio, R. Bhargava, E. Zuckerman, and L. Beck, “Cliff-Clavin: Determining geographic focus for news,” NewsKDD: Data Science for News Publishing, at KDD 2014, 2014.
[3] mordecai, “[online],” URL: Available: https://github.com/openeventdata/mordecai.

Cliff-Clavin

- It employs context-based geographic disambiguation over organizations and locations extracted from the text using Stanford CoreNLP.
- Identifies focus places at city, state and country levels along with geo-coordinates.

Drawback: Failure to identify primary focus location.

{"results": {"places": {"mentions": [{"lon": 0, "source": {"string": "Earth"}, "name": "Earth", "lat": 0}, {"lon": 120, "countryGeoNameId": "1643084", "source": {"string": "Indonesia"}, "countryCode": "ID", "name": "Republic of Indonesia", "lat": -5}, {"lon": -113.64258, "countryGeoNameId": "6251999", "source": {"string": "Canada"}, "countryCode": "CA", "name": "Canada", "lat": 60.10867}, {"lon": 115, "countryGeoNameId": "1643084", "source": {"string": "Bali"}, "stateGeoNameId": "1650535", "countryCode": "ID", "name": "Provinsi Bali", "stateCode": "02", "lat": -8.5}], "focus": {"cities": [], "countries": [{"countryCode": "ID", "name": "Republic of Indonesia", "lon": 120, "countryGeoNameId": "1643084", "lat": -5}, {"countryCode": "CA", "name": "Canada", "lon": -113.64258, "countryGeoNameId": "6251999", "lat": 60.10867}], "states": [{"stateGeoNameId": "1650535", "countryCode": "ID", "name": "Provinsi Bali", "lon": 115, "countryGeoNameId": "1643084", "stateCode": "02", "lat": -8.5}]}}}}

Mordecai

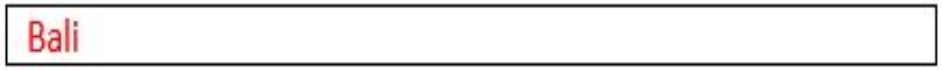
- It uses MITIE's NER tool to extract place names and then uses a gazetteer in an elasticsearch index to identify focus country and other place names.
- Identifies focus places at city, state and country levels along with geo-coordinates.

Drawback: Failure to identify primary focus location.

[{"word":"Rajendra Pachauri", "country_predicted": "BTN", "country_conf": 7.1679338e-12}, {"word": "Earth", "country_predicted": "NA", "country_conf": 0.0000010062237}, {"word": "Bali", "country_predicted": "IDN", "country_conf": 0.97859323, "geo": {"admin1": "Bali", "lat": "-8.5", "lon": "115", "country_code3": "IDN", "place_name": "Provinsi Bali"}}, {"word": "Indonesia", "country_predicted": "IDN", "country_conf": 0.99957937, "geo": {"lat": "-5", "lon": "120", "country_code3": "IDN", "place_name": "Republic of Indonesia"}}, {"word": "Canada", "country_predicted": "CAN", "country_conf": 0.99998522, "geo": {"lat": "60.10867", "lon": "-113.64258", "country_code3": "CAN", "place_name": "Canada"}}]

Profile

- It identifies a primary focus location associated with a document using MITIE's NER tool and with a supervised learning method.



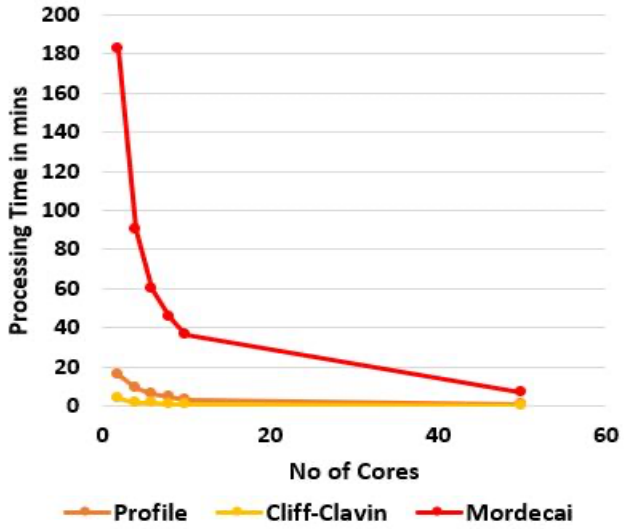
Drawback:

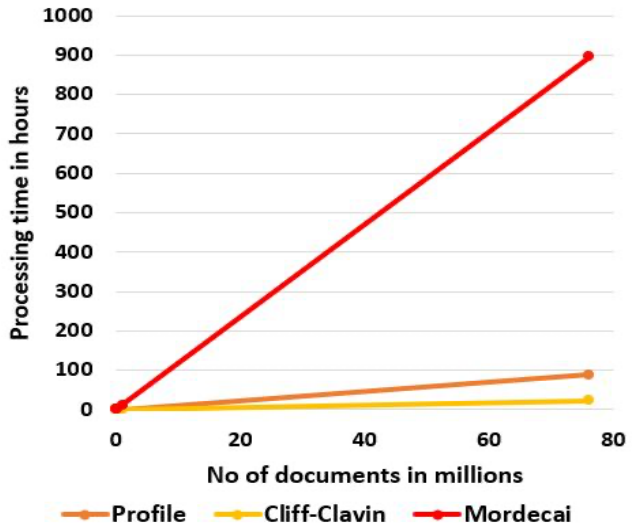
- Failure to extract lat-long information for the identified primary focus location.
- Higher memory consumption.

Comparison

Geoparser	Total Processing Time	Speed
Cliff-Clavin	17.27 Mins	239.1 Docs/Sec
Profile	81 Mins	51.12 Docs/Sec
Mordecai	912 Mins	4.53 Docs/Sec

Cliff-Clavin takes 0.29 hours, Profile takes 1.35 hours and Mordecai takes 15.2 hours to process 250K documents on a workstation with 10 cores.





Processing Speed of Cliff-Clavin >> Profile >> Mordecai

Comparison (.. continued)

Primary focus location accuracy comparison between different methods:

Atrocity Event Data[1]:

Method	**Accuracy** (%)
Profile_s	69.47
Profile_s ($\gamma = 7$)	**71.27**
Cliff-Clavin	63.75
Modified Stanford-CoreNLP	60.83
Modified Mordecai	49.96

New York Times[2]:

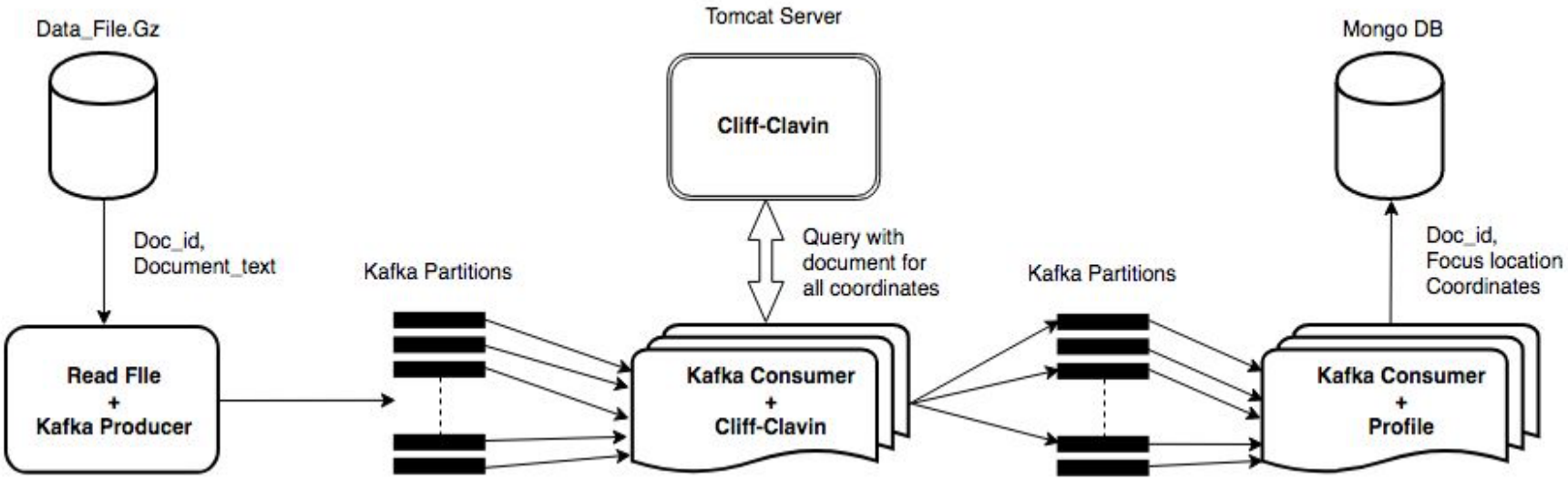
Method	**Accuracy** (%)
Profile_s	54.27
Profile_s ($\gamma = 7$)	**64.21**
Cliff-Clavin	53.65
Modified Stanford-CoreNLP	36.25
Modified Mordecai	22.97

[1] [Online]: http://eventdata.parusanalytics.com/data.dir/atrocities.html/
[2] [Online]: https://catalog.ldc.upenn.edu/ldc2008t19

SPERG

- It's a federated system comprising of Cliff-Clavin and Profile to extract coordinates of the primary focus location.
- Cliff-Clavin gives accurate lat-lon and is faster than mordecai.
- Profile identifies primary focus location accurately.
- This systems three major tasks are:
  - Read and distribute data from source file;
  - Identify coordinates of all the locations using Cliff-Clavin;
  - Identify primary focus location using Profile and map with results of Cliff-Clavin.

SPERG: Architecture

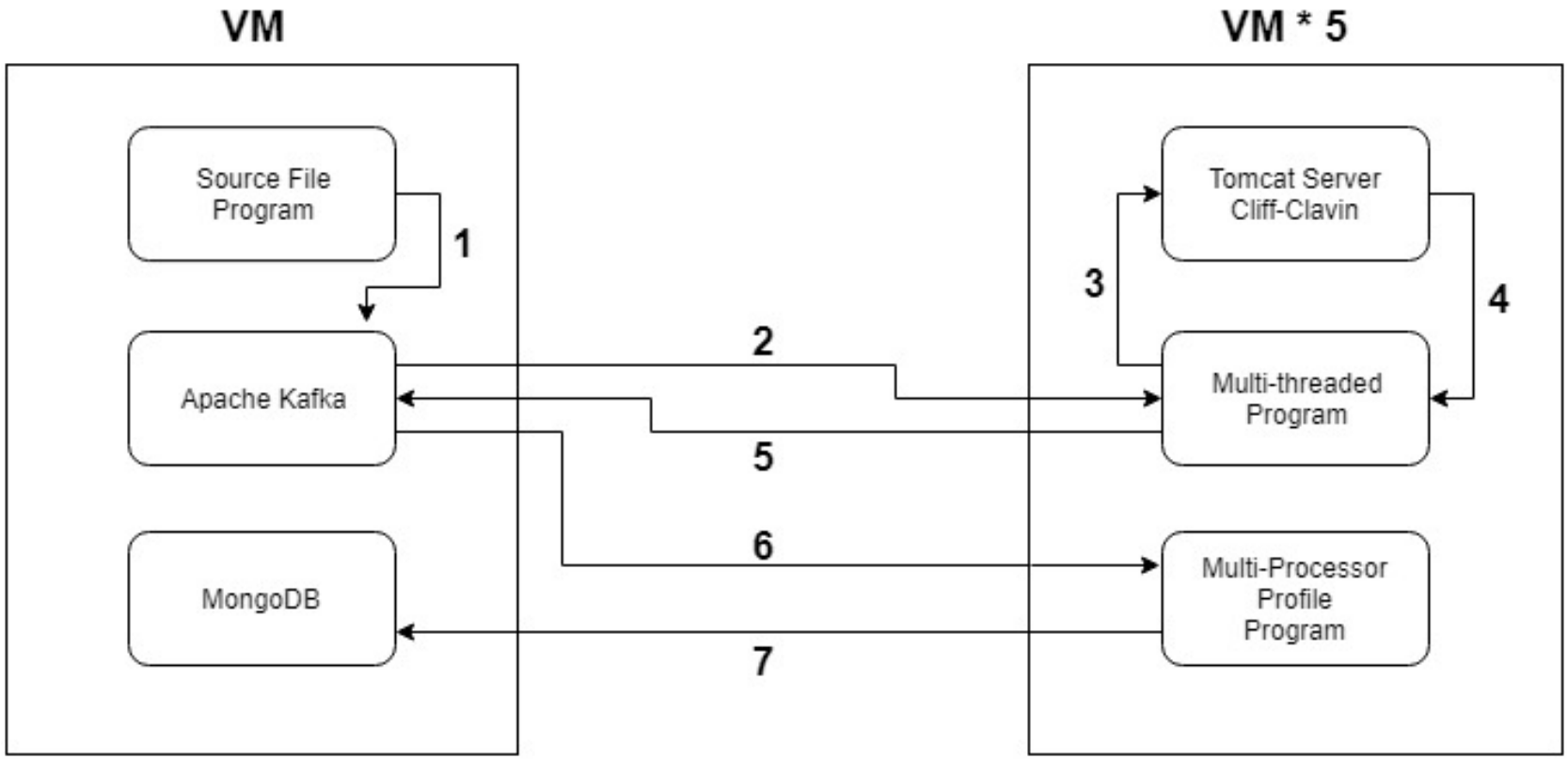


Distributing Data Source Task

Geo-coordinate Extraction Task

Focus Location Coordinate Mapping Task

SPERG: Flow Diagram



Architecture (..continued)

- Message transfers between servers are carried out using Apache Kafka.
- Parallelism is high when the rate of data flow is greater than what the pipeline can process.
- Cliff-Clavin is hosted as a service in Apache Tomcat server, so that it can process hundreds of requests in parallel.
- Profile was utilized in a multiprocessor environment.

Distributing Data Source

- A single 37 GB compressed source file contained 76.1 milllion news reports.
- Efficient geoparsing will rely on a parallel processing approach.
- To asynchronously consume documents from queues, we used message buffer such as Apache Kafka.
- Why kafka? – Highly scalable in nature and fault tolerant.
- Documents were pushed to kafka in batches of size slightly greater than the throughput of Sperg.

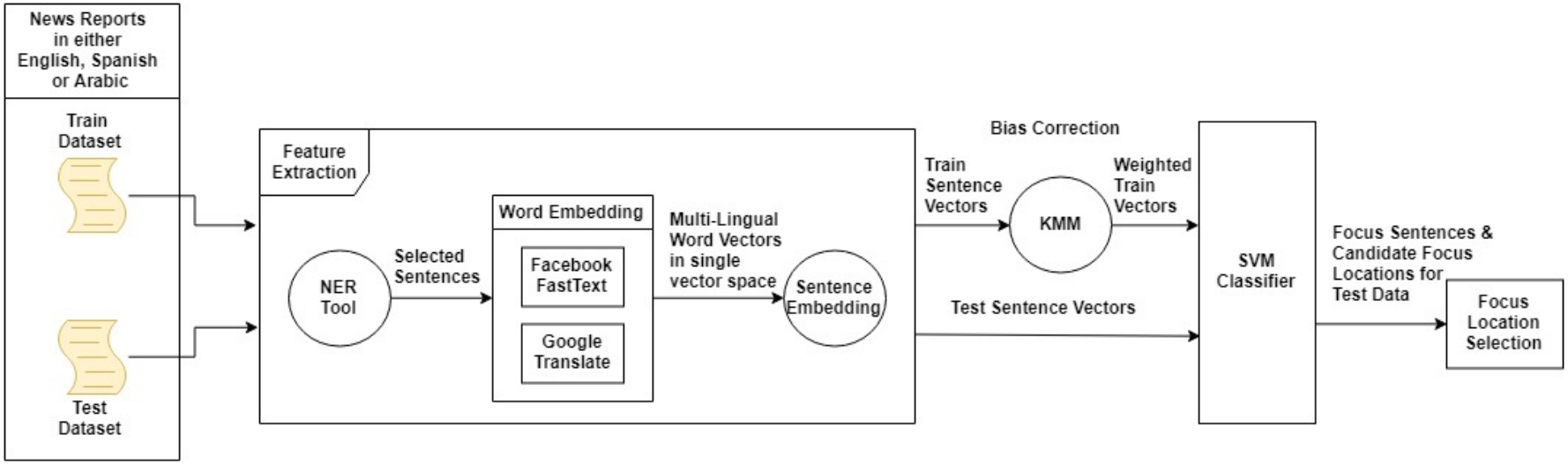
Geo-coordinate Extraction

- It's an I/O intensive process, thus we used a multithreaded program to process documents.
- Each thread places an HTTP request to the Cliff-Clavin server with the document consumed from Kafka.
- Response from the server contains coordinates of all locations mentioned in the document.
- This geographic information is appended onto the original document and pushed to a different kafka topic.

Focus Location Coordinate Mapping

- Profile identifies the primary focus location of a news report and it is the major bottleneck in the pipeline.
- It's a process intensive program, thus we implemented it in a multiprocessor environment.
- Each processor consumes from Kafka and gets the primary focus location using Profile's algorithm.
- This primary focus location is mapped with Cliff-Clavin's result for the coordinates and then written onto a MongoDB.

PROFILE: Architecture



- We modified word embedding technique of Profile to enable support for multiple languages with a single trained model.

PROFILE: Architecture (..continued)

Word Embedding:

- Used Facebook's fastText and Google Translate API to align monolingual vectors from two languages in a single vector space.
- The length of these vectors are 300 and they effectively encode the semantic meaning of the words in context.

Sentence Embedding:

- Sentence vector is computed by taking mean of word vectors in the sentence
- The effectiveness of this approach was empirically compared with another scheme of sentence embedding by assigning different weights to each word.

PROFILE: Bias Correction

Challenges:

- The requirement of suitable amount of labeled data instances for training.
- Traditional supervised learning methods assume that the training and test data sets are generated from the same data distribution.

Solution:

- Leveraging the sampling bias correction by weighting each training data instance using density ratio estimates between the test and training data distributions with Kernel Mean Matching (KMM) [1]

[1] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola, "Correcting sample selection bias by unlabeled data," in Advances in neural information processing systems, 2006, pp. 601–608.

Optimizations

- Processed documents in batches of 6 million so that there would be minimal reprocessing in case of a system failure.
- Maximum parallelism is observed when the number of partitions of a topic in Kafka is equal to the total number of Kafka consumers.
- For instance, a total of 25 consumer threads across 5 servers, requires a topic with 25 partitions in Kafka so that each thread consumes from it’s own dedicated partition.

Optimizations (..continued)

- MongoDB write performances were improved by disabling journaling and creating indices after processing all the documents.
- Profile's word embedding approach loads 6 GB vector file into memory, so forking 10 processes would yield a memory consumption of 60 GB.
- To tackle memory overflow, fasttext vectors are hosted as common service since only read operations were performed.

Setup

- We used XSEDE resources, two s1 xxlarge machines with 44 cores and 120 GB memory, to create 6 virtual machines of Intel E5-2680v3 Hasweell CPU @ 2.50GHz

VM Count	RAM	Disk Space	No. of Cores	Use
5	29 GB	239 GB	10	Geoparsing
1	29 GB	2 TB	10	Host Kafka and MongoDB

Experiments (Datasets)

Dataset	Documents	Size
Terrier Event Data	76.1M	76GB
Randomly Sampled Terrier Event Data	247.8K	290MB
Atrocity Event Data Training data for profile	3.6K	9.3KB

- The Terrier data contained encoded event data for 76.1 million news reports which was prepared with assistance from [1].
- We randomly sampled small amount of terrier data to optimize Sperg's architecture and find it's throughput.
- The Atrocities Event Data is a collection of English news reports on atrocities and mass killings in several locations.

[1] O. Tange, GNU Parallel 2018. Ole Tange, Mar. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1146014

Experiments

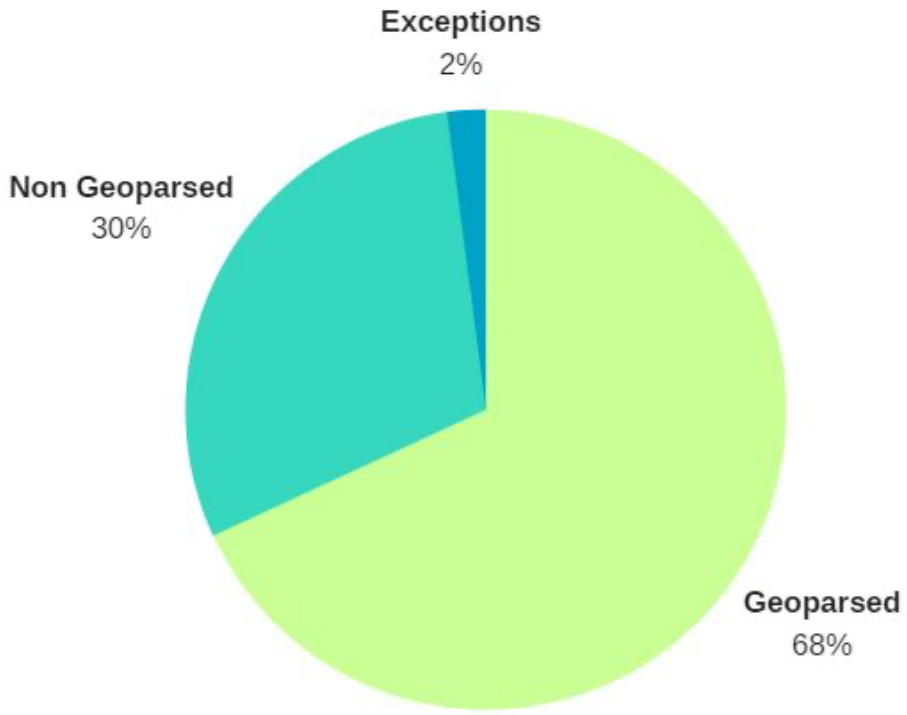
- Two topics with 25 partitions for Cliff-Clavin's consumer and 50 partitions for Profile's consumer were created on Kafka.
- A python script read 6 million documents from source file at fixed interval and published the documents to Cliff-Clavin's consumer in a single thread.
- Cliff-Clavin program had 5 threads consuming documents from kafka, making HTTP requests to Cliff-Clavin server and then publishing the results onto Profile's consumer asynchornously.

Experiments (..continued)

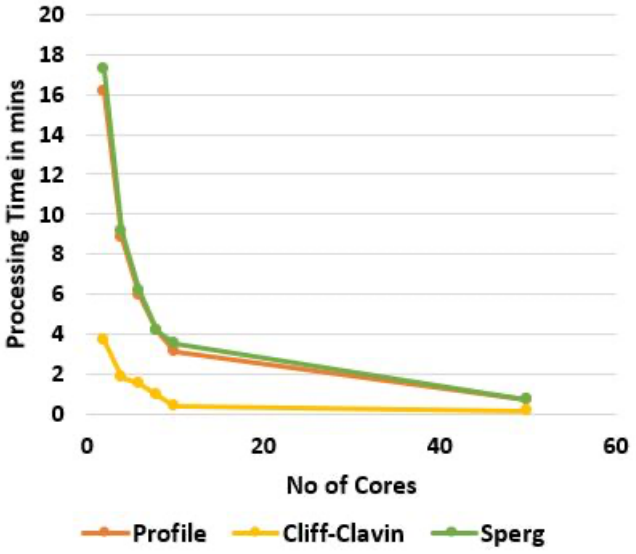
- To achieve parallelism, Profile was hosted in a multiprocessor environment with each processor responsible for:
  - Consuming text and coordinates from Kafka;
  - Identifying the primary focus location from the given text;
  - Matching the primary focus location with the result from Cliff-Clavin for the coordinates;
  - Writing the results onto a MongoDB collection.
- Indexing was performed last, since insertion are slower for an indexed collection as the indices have to be re-calibrated after every insertion.
- Sperg processed 76.1 million documents in 5.29 days inclusive of two server failures that demanded reprocessing of two batches.

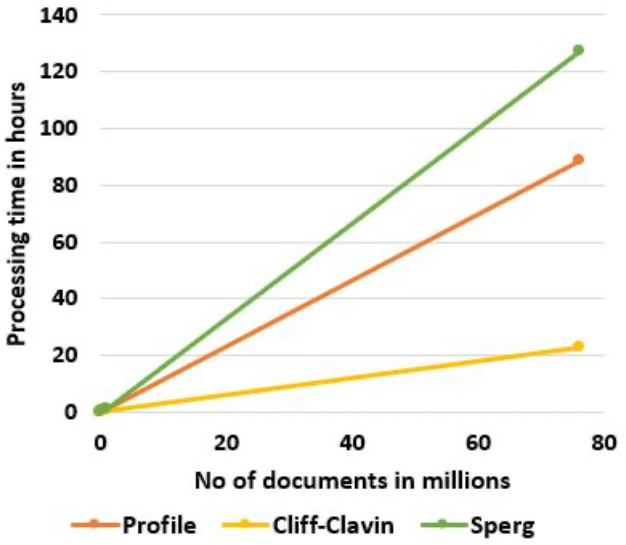
Results

- 76.1 million documents were processed by Sperg in 5.29 days.
- 51.83 million documents were successfully geoparsed.
- 22.68 million documents were failed to be geoparsed by Sperg.
- 1.59 million documents ran into exception while processing them with the available geoparsers.



Performance of Sperg





- We observed that Cliff-Clavin can parse 76.1 million documents within 24 hours while Profile would complete the same task in 4 days.
- Therefore, Sperg is configured to produce throughput equal to that of Profile, but observed throughput is lower because of network latency and system failures.

Evaluation of Geoparsed Docs

Case	Cliff Clavin Result	Profile Result	Documents count (in Millions)
I	✓	✓	47.74
II	✓	✗	4.09

- Case I is when results of Profile matched with one of the location from the results of Cliff-Clavin.
- Case II is when Cliff-Clavin identifies a single location while profile didn't, which is taken as the primary focus location for the document.

Evaluation of Non-Geoparsed Docs

Case	Cliff Clavin Result	Profile Result	Documents count (in Millions)	Possibility to extract coordinates ?
I	✓	✓	11.26	Yes
II	✗	✓	2.81	Yes
III	✓	✗	2.87	Maybe
IV	✗	✗	5.74	No

- In Case I and II, Profile gave result but didn't match with Cliff-Clavin's result, so a different geoparser can be used to combine with Profile.
- In Case III, possibility to identify coordinates of primary focus location by utilizing a frequency based algorithm over Cliff-Clavin's result.
- In Case IV, both geoparsers failed to identify any location and used different NER tools, So they may not have any locations mentioned.

Conclusion and Future Work

- **Conclusions**
  - Implemented a scalable distributed framework to extract geo-coordinates of the primary focus location for the events from political news reports;
  - Sperg successfully geoparsed 51.83 million documents;
  - Sperg is flexible due to the capability to add or remove servers dynamically as per the requirement or problem size.

- **Future Work**
  - Build a real time system to process live news reports on the go;
  - Support for multilingual report geoparsing using the capabilities of profile;
  - Add more geoparsers to our federated system to process the remainder of the documents which Sperg failed to geoparse.

Thank you!