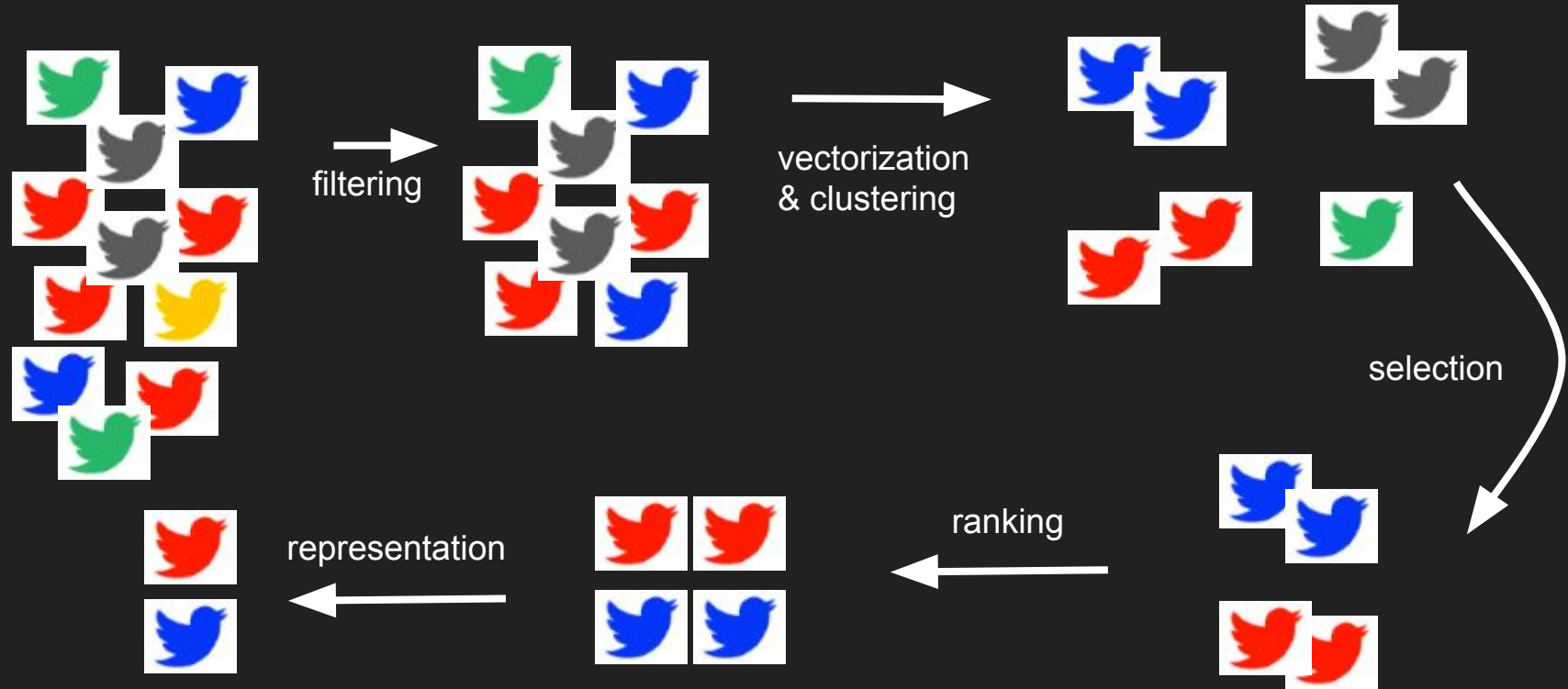


Twitter News Feed

Intermediate presentation

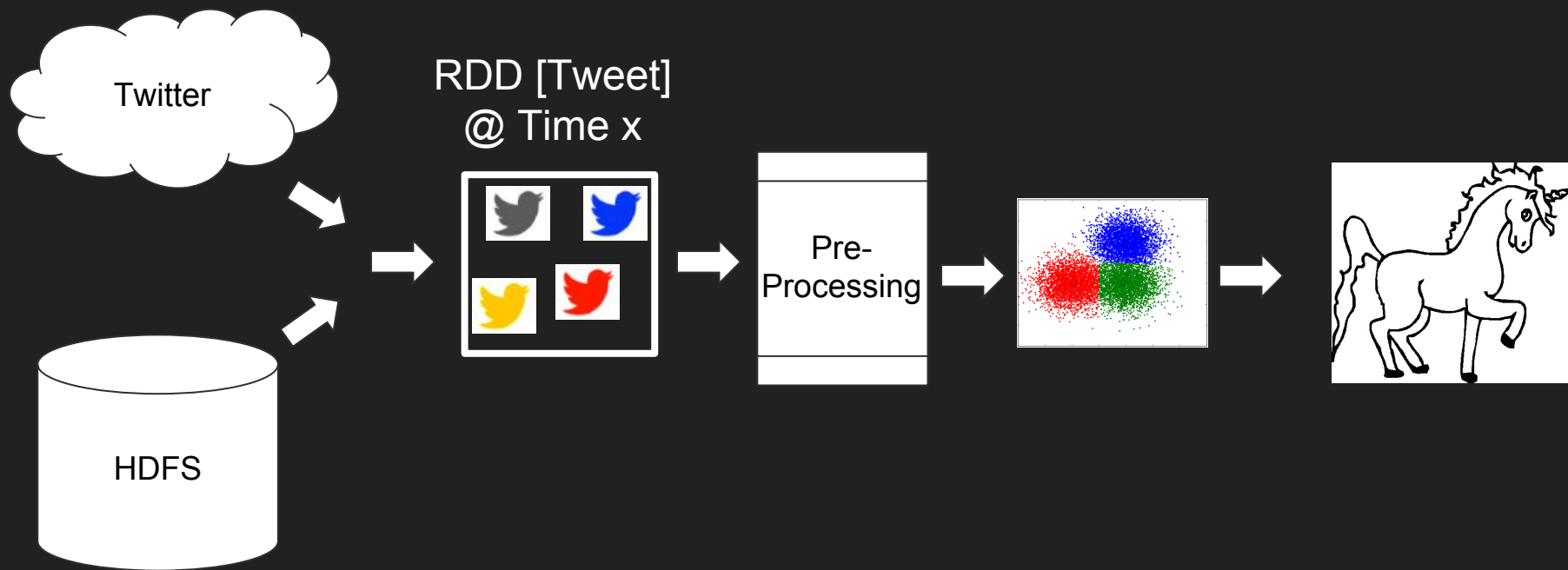
Goal of the project: News feed from Twitter



Apache Spark Streaming



Implementation



Preprocessing



Tokenization

Stop words removal

Stemming / Lemmatizing

Term Frequency in Different Batches

- B1 T1: “Hello world, I love programming!”

	hello	world	I	love	programming
B1 T1	1	1	1	1	1

- B2 T1: “Hello Peter, I love dancing Tango!”

	hello	peter	I	love	dancing	tango
B1 T1	1	1	1	1	1	1

Hashing-TF

- Fixed vector dimension size
- $\text{hash}(\text{word}) \rightarrow \text{Index in vector}$
- Can lead to collisions

	hello, programming	world	peter, love	I, dancing	tango
B1 T1	2	1	1	1	0
B2 T1	1	0	2	2	1

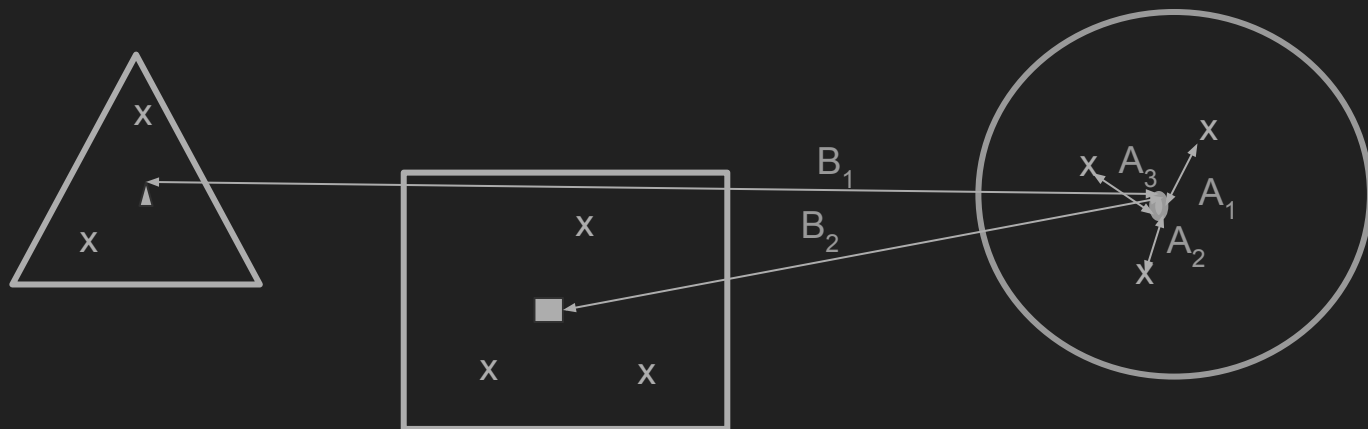
Clustering using Streaming K-Means (1/2)

- Continuous clustering with batches
- Centroids and cluster size are saved
- For every batch, new centroids are calculated
 - Centroids are weighted by their cluster size
 - Old centroids also weighted by decay

Clustering using Streaming K-Means (2/2)

- Important parameters:
 - batch size
 - k
 - decay
 - initial centroids

Cluster Analysis and Selection



- Silhouette = $(B - A) / \max(A, B)$ [range: -1 .. 1]
- Only consider clusters with a minimum number of tweets (relevant) and a minimum silhouette score (likely correct)

Representation

- Representative: Closest Tweet to centroid
- Linked article(s): Most frequently occurring URL(s)

Example result

Cluster size: 14, simplified silhouette: 0.170



Mr Drippy
@simon198

Nigel Farage: £350m NHS pledge 'a mistake'
bbc.co.uk/news/uk-366246...



Jo S
@Lacerslife

Oh convenient you admit that now you tosser
Nigel Farage: £350m NHS pledge 'a mistake'
bbc.co.uk/news/uk-366246...



Patrick Tony Lynch
@tonyakapat

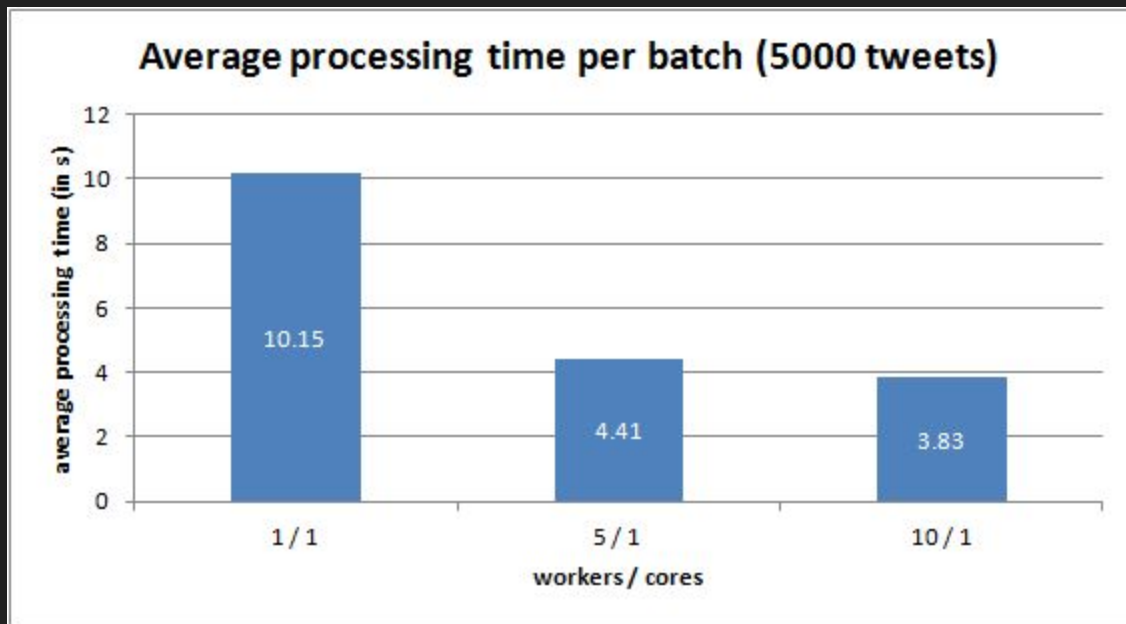
Nigel Farage: £350 million pledge to fund the
NHS was 'a mistake' | via [@telegraphnews](#)



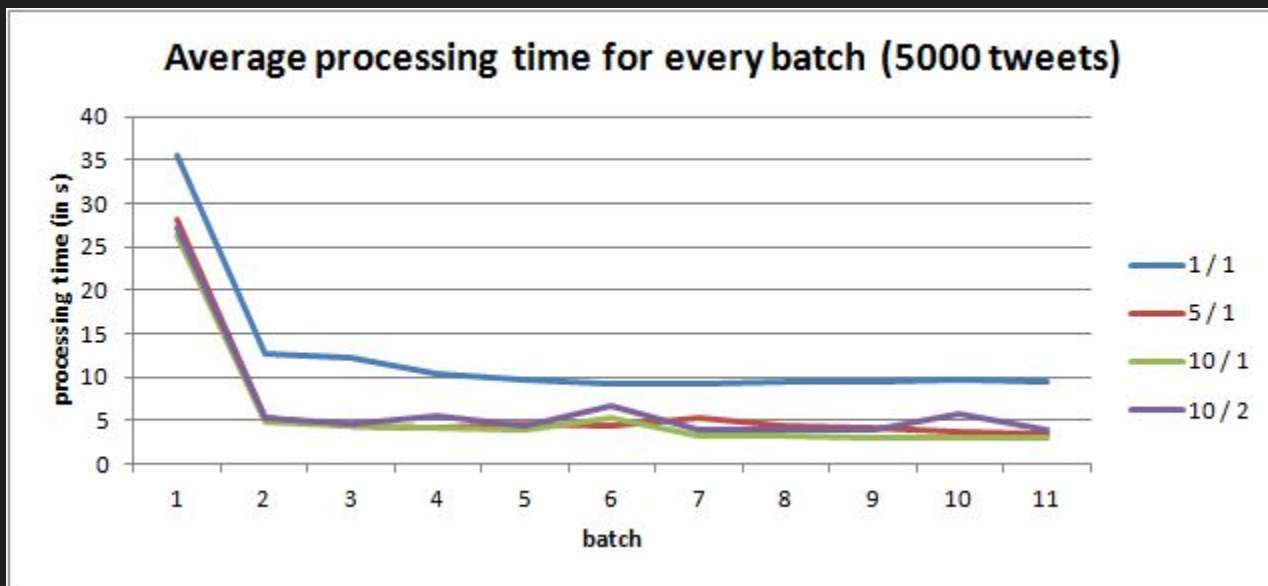
Diana Coman
@DianaComan

Nigel Farage disowns Vote Leave '£350m for the
NHS pledge' hours after EU referendum result

Performance



Performance



Outlook

- Track results across badges
- Better initial centroids
- Dynamically adjust parameters based on cluster results
- Sentiment analysis for interesting clusters

→ Improve the quality of the clustering