Assignment Report

Online Clustering for Topic Detection
in Social Data Streams

Submitted By

Monisha Nair 2017H1120241P

Sonali Sharma 2017H1120239P

Contents

Description 3

Cluster Centroid Object 3

Tweet Object 3

Feature Selector 3

Fading Function 3

Similarity Function 3

Dynamic Measures 4

Parameters Used 4

IMPROVEMENTS 4

DATASETS: 4

BLOCK DIAGRAMS 5

SYSTEM BLOCK DIAGRAM 5

FLOW CHART : 6

FEATURE SELECTION: 7

EVALUATION 8

EVALUATION MEASURE FOR REFINED CLUSTERS: 8

EVAULATION MEASURE FOR CLUSTER SIMILARITY: 8

ANALYSIS: 9

GRAPHS: 9

SNAPSHOTS OF TOPIC DETECTED 11

Description

Cluster Centroid Object

- Cluster Label
- Start Time
- Last Updated Time
- List of keywords
- Buffer
- Child array
- Height

Tweet Object

- Tweet Identifier
- Start Time
- List of keywords for the tweet

Feature Selector

In order to get relevant keywords from tweets Unigrams and bi-grams of tweets are considered.

A score is given to each segment Based on its weight in Local and Global context.

1. For Global Context we are using two services:--
    a. Microsoft N-gram services Stanford NER services
2. To assign weight in local context:--
    a. A local corpus of the keywords is maintained which is used to calculate TF-IDF of segments from the tweets
    b. A local corpus of the NER-keywords is maintained which is used to calculate weight of NER-segments from the tweets

Using the above final score to keywords is allotted and list of top Keywords are selected.

Fading Function

Used to decay old to decay old cluster centroids

$$f(t_c) = 2^{-\lambda(t_{so} - t_c)}$$

- t_{so} : publication time of the post tw that could be added to C.
- λ is the decay rate of a cluster.

Similarity Function

$$sim(so, CC) = \frac{\sum_{j=1}^{|I^i|} sgn_C . ff_I(j)}{\sum_{j=1}^{|U^i|} sgn_C . ff_U(j)} \times f(t_c)$$

- sgnC.ffI(j) ,sgnC.ffU(j) are the sum of frequencies of the terms appearing in the intersection and union, respectively
- f(tc) is the fading function

Dynamic Measures

Dynamic similarity measures=10**(height_of_descend_centroid)*similarity_of_parent_centroid

Dynamic Buffer count=0.125*(height_of_descend_centroid)*buffer_count_of_parent_centroid

Parameters Used

S.no.	Parameters
1	Initialisation tweets
2	Final_weight_for_segment_to_be_a_keyword
3	Threshold_for_active_cluster_pruning
4	Threshold_for_similarity
5	Threshold_to_merge_cluster

IMPROVEMENTS

- Select Tweets to initialize NER-corpus and corpus
- Obtain Tweet from stream
- Preprocess the tweets text
- Extract keywords from tweets as shown in figure 3
- Make tweet Object
- On getting interruption pass the list of collected tweet objects along with parameters Threshold_for_active_cluster_pruning , Threshold_for_similarity in order to obtain centroid objects or current clusters.
- If time is available pass the list of centroid objects to a global queue
- Pop centroid object from queue , extract it's buffer which contains list of tweet objects
- Calculate the dynamic similarity measure , dynamic buffer count and further pass them to calculate list of child clusters. Following cases could occur.
    - If no clusters are formed : Buffer of current centroid needs to be populated in order to refine it. However if the parent cluster has child then the children are added to the queue for future refinement else the parent cluster is added to the list of final topic.
    - If clusters are formed: if parent clusters has child clusters then newly formed clusters are compared with these child clusters and merged with the most similar one and added to the queue .If no merging happens then new clusters ae added as children of the parent cluster and added to the queue for refinement.

DATASETS:

- Twitter Datasets

BLOCK DIAGRAMS

SYSTEM BLOCK DIAGRAM

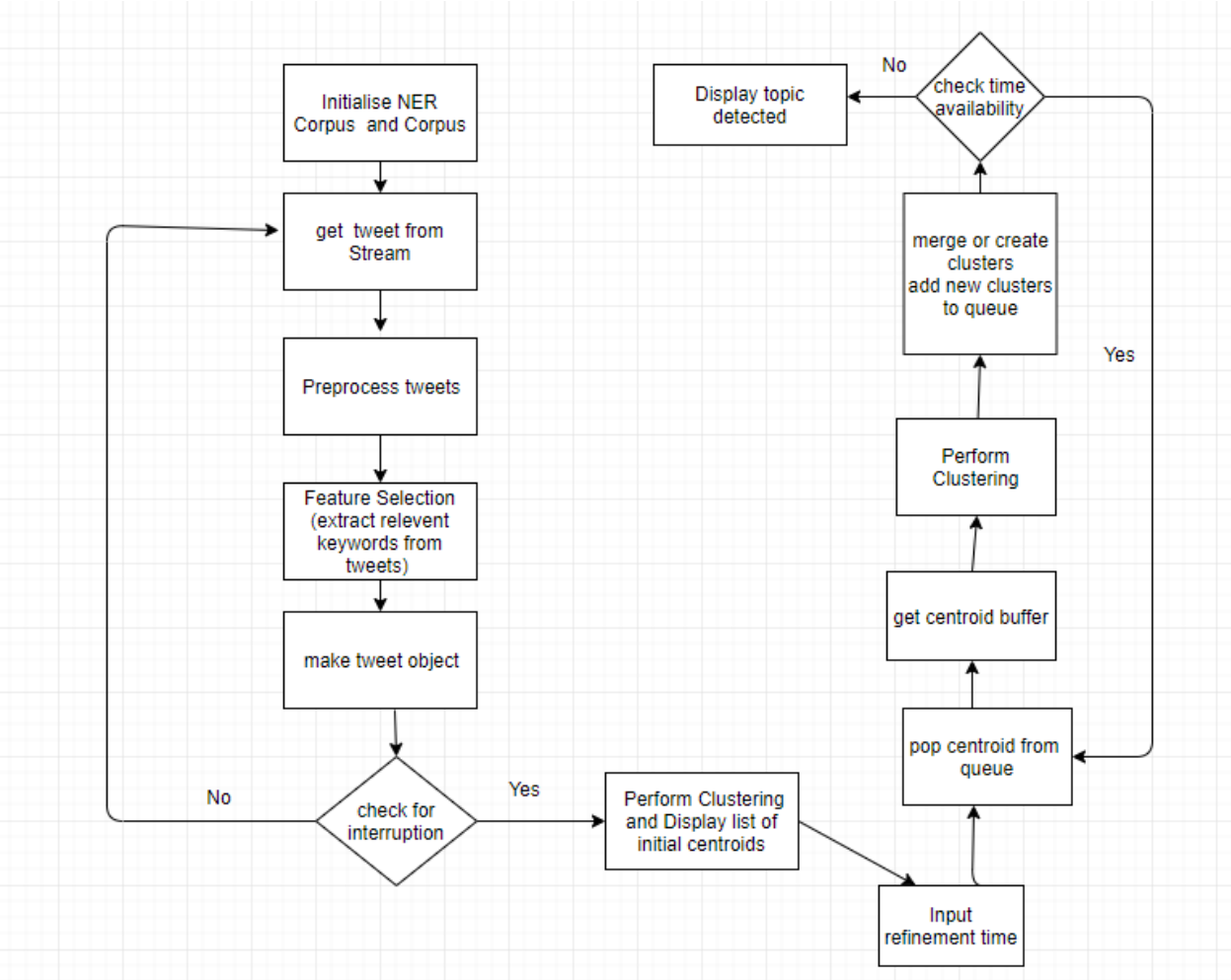


Figure 1

FLOW CHART :

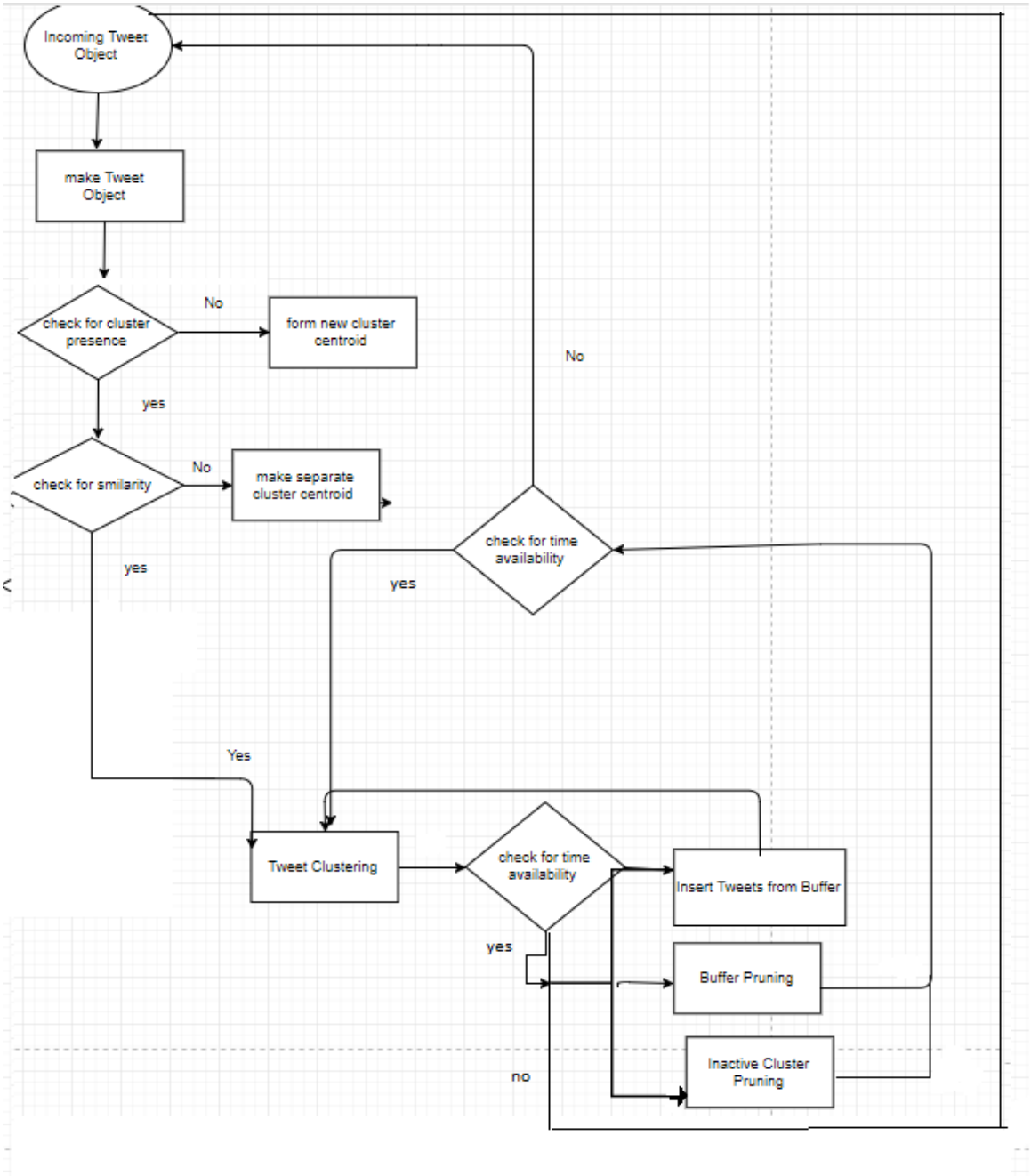


Figure 2

FEATURE SELECTION:

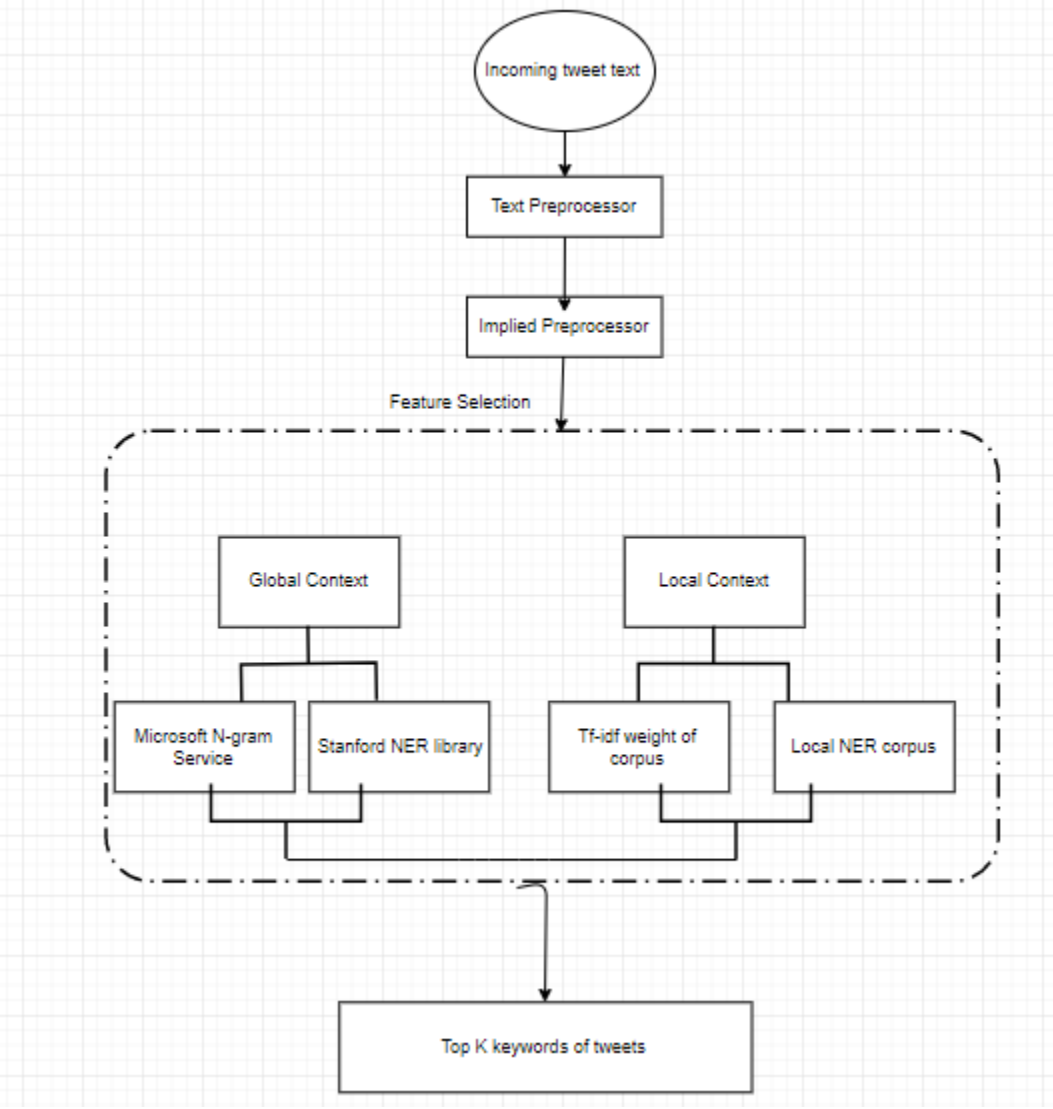


Figure 3

EVALUATION

EVALUATION MEASURE FOR REFINED CLUSTERS:

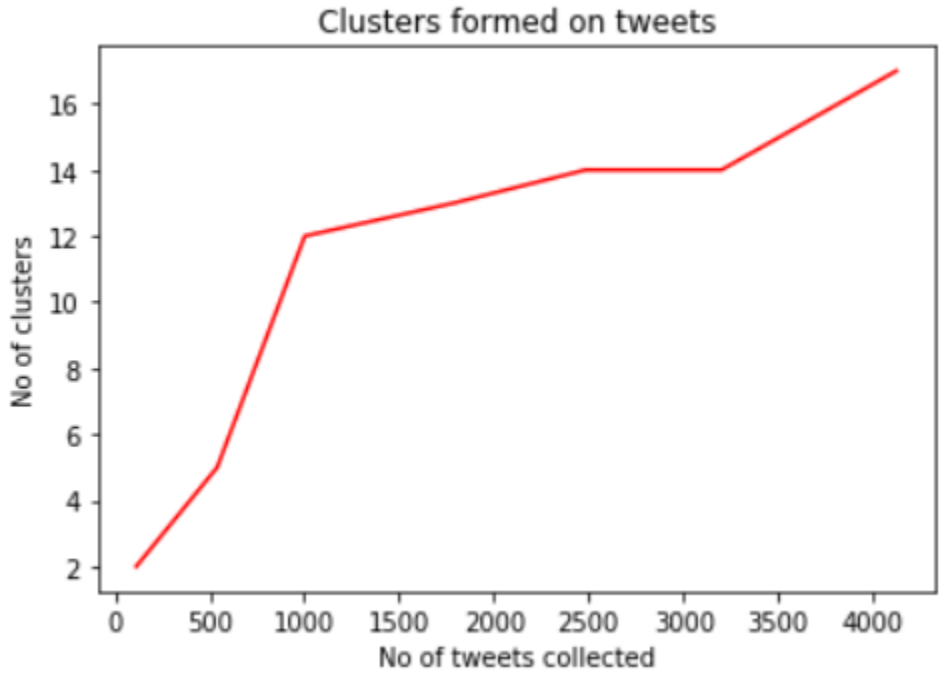
fetched_tweet	Assignd_tweet	level_1_cluster	refinedcluster	final_cluster_	time	similarity_	buffer_
113	90	2	0	2	12	0.002	5
540	329	5	0	5	12	0.002	5
1789	1049	8	5	12	12	0.002	15
1002	713	11	5	13	12	0.002	15
2482	2110	11	7	14	12	0.002	25
3200	2875	14	5	14	12	0.002	25
4123	3698	17	11	17	12	0.002	25

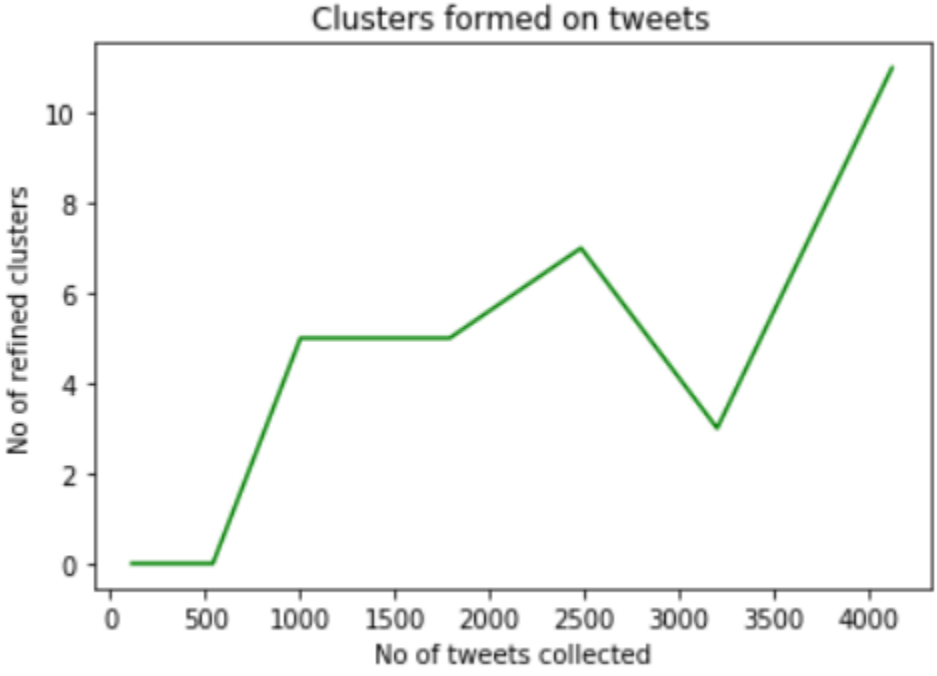
EVAULATION MEASURE FOR CLUSTER SIMILARITY:

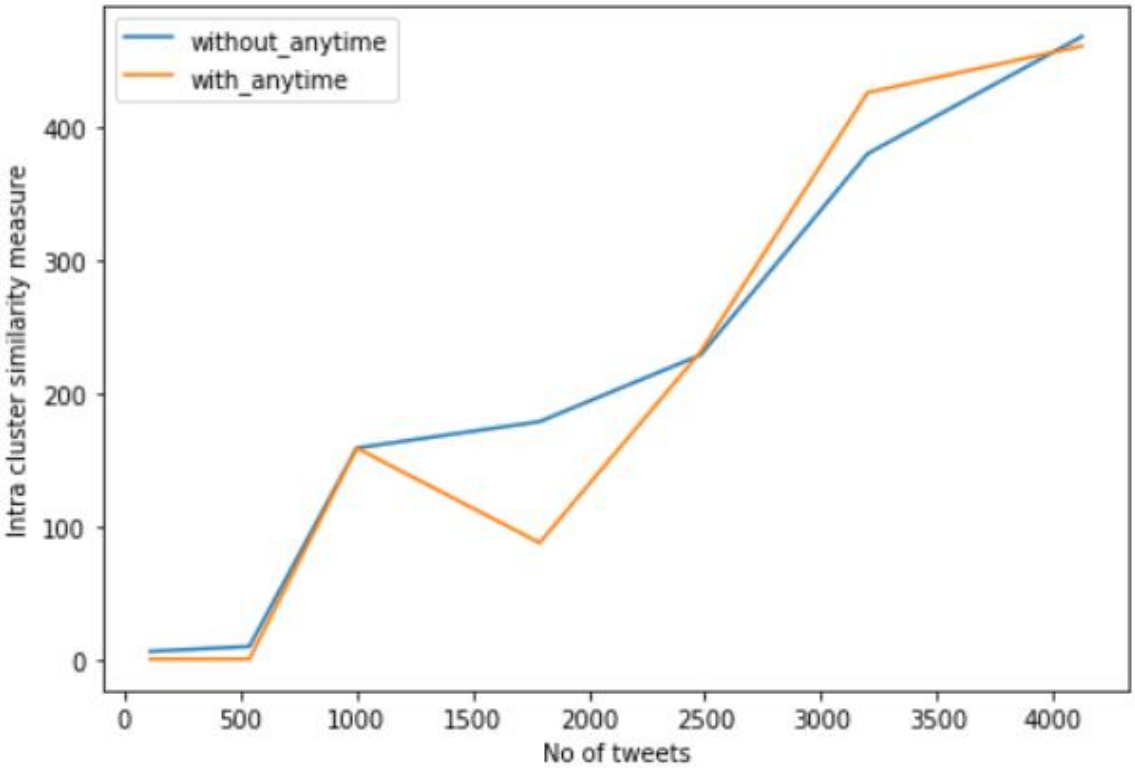
No of tweets	**Cluster_similarity_without_anytime**	**Cluster_similarity_with_anytime**	**difference**
113	5.78	0	-5.78
540	9.4567	0	-9.4567
1002	158.498	158.498	0
1789	178.2995	87.2995	-91
2842	228.3789	230.7896	+2.4107
3200	379.4621	425.2998	+45.83
4123	467.547	460.3456	-7.20

ANALYSIS:

GRAPHS:






without_anytime
with_anytime
Intra cluster similarity measure
400
300
200
100
0
0
500
1000
1500
2000
2500
3000
3500
4000
No of tweets

SNAPSHOTS OF TOPIC DETECTED

tweet_topics_without_anytime	tweet-topics_after_anytime
['health coverage', 'coverage Need'], ['open enrollment', 'begin today'], ['tech company', 'Senate'], ['question today', 'today'], ['merit based', 'Diversity', 'Visa'], ['best family', 'work best', 'care plan', 'health care']['working family']	['care coverage', 'next year', 'health care'], ['liberty week', 'week', 'Senate', 'abortion week'], ['working family'], ['meet need', 'coverage Need'], ['enrollment start', 'open enrollment', 'begin today'], ['medium election', 'social medium', 'Facebook', 'Senate']
['York', 'citizen parent'], ['Bank', 'American', 'community', 'begin today', 'today'] ,['people', 'Senate', 'Circuit'] ,['economy'],['Congress', 'Immigration'],['right'] ,['start today'] ,['plan'],['local economy', 'friend coworkers'] ,['vote', 'Legislative'] ,['rule', 'billion decade']	['community', 'begin today', 'today'] ,['make', 'Senate', 'Larsen', 'Circuit'],['people'] ,['economy'],['health care', 'Congress', 'Immigration'] ,['right'],['start today'],['plan'] ,['cost', 'recipient integral', 'integral part', 'cost employer'] ,['friend coworkers'],['vote', 'Legislative'] ,['rule', 'begun minute', 'billion decade'] ,['York', 'citizen parent']
['York', 'citizen parent'], ['people', 'Circuit'] ,['need', 'able discriminate', 'right Businesses'] ,['Congress', 'Immigration'], ['right', 'fast', 'reform markup'] ,['responder', 'support', 'Program', 'Diversity', 'Lottery'] ,['start today'], ['local economy', 'friend coworkers'] ,['group', 'Resilient'],['vote', 'Legislative'] ,['rule', 'begun minute']	['make', 'Senate', 'Joan', 'Larsen', 'Circuit'],['people'],['protect', 'Nebr'] ,['able discriminate', 'right Businesses'] ,['health care', 'Congress', 'Immigration'] ,['right', 'fast', 'reform markup'],['based', 'including away', 'Diversity', 'Lottery'] ,['start today'],['cost', 'recipient integral', 'integral part', 'cost employer'] ,['friend coworkers'],['support', 'Resilient'] ,['vote', 'Legislative'],['rule', 'begun minute'] ['York', 'citizen parent']