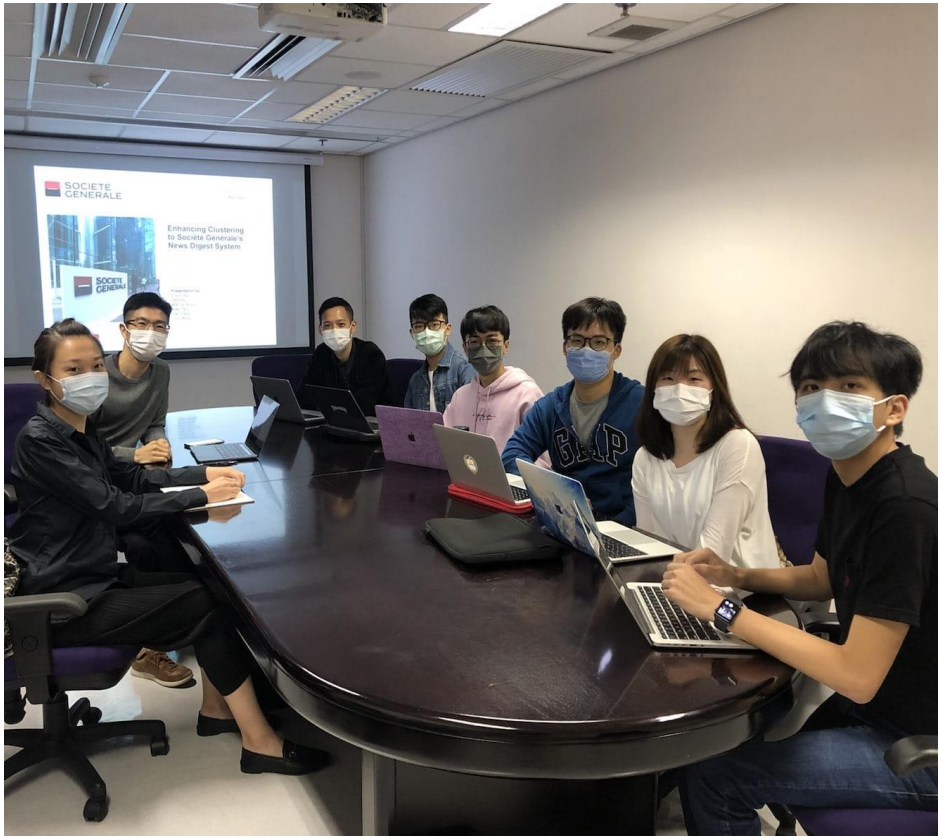
April 2021

Enhancing Clustering for Société Générale's News Digest System

Presented by
Enoch Ng
Felix Ng
Marcus Wong
Riley Tsoi
Sam Pang
Tony Wong

Background

About our project



WHO IS OUR CLIENT

We are very pleased to have the Risk Management (Market Risk) Department of Société Générale (Asia) to be our project client.

WHAT HAVE WE BEEN DOING

Our group is conducting a research project on enhancing SocGen's NLP algorithms used in the news digest system.

What are the current limitations?

1. "Not able to group similar articles into a group"
2. "Cannot identify the most popular news over a specified time period"
3. "Lack of systematic and numerical ways to evaluate the grouping"

Business Analyst

How do we address the limitations?

Grouping similar news into one cluster

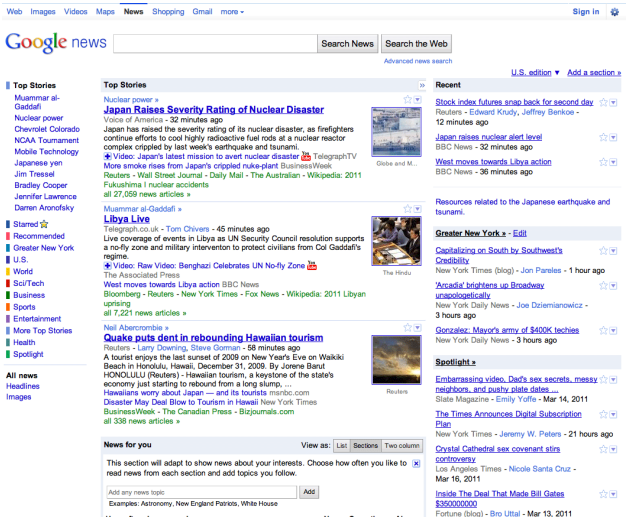
Identifying popular keywords in the clusters

Providing selection of news based on time period

Exploring different evaluation metrics

Where do we take reference from and why?

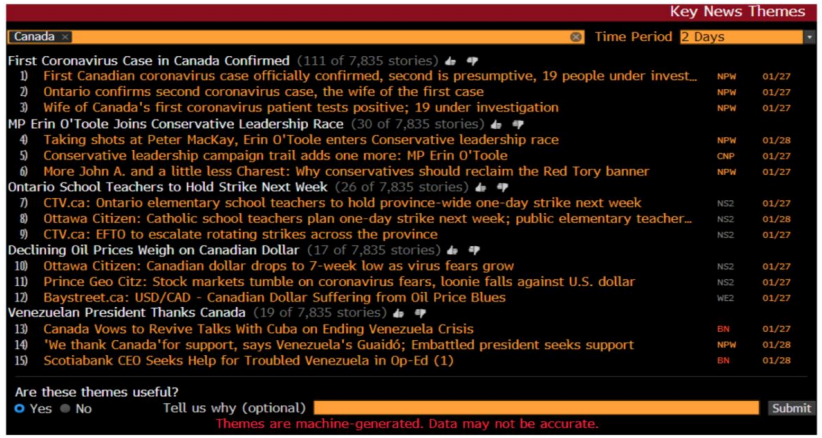
Google News



Uses methods trained on large text corpus

Google's news word embedding methods are trained with approximately **100 billion words**, which are also proven academically to have higher accuracy**.**

Bloomberg



Similar use-case

The functionalities and features of Bloomberg's NSTM are the **closest to our desired deliverable**.

How can our project bring value to SG?

Time-saving

Able to identify top news at a glance

A certain degree of customization

Read clustered news within different timespan or from different news datasets

Systematic measurement of clustered results

Better evaluate the word embedding and clustering methods objectively

Workflow
Recap & Updates

Workflow Recap

- Non-custom dataset
- Custom dataset

- Comparison of different clustering methods
- Keyword and Named Entity Recognition
- Recursive clustering

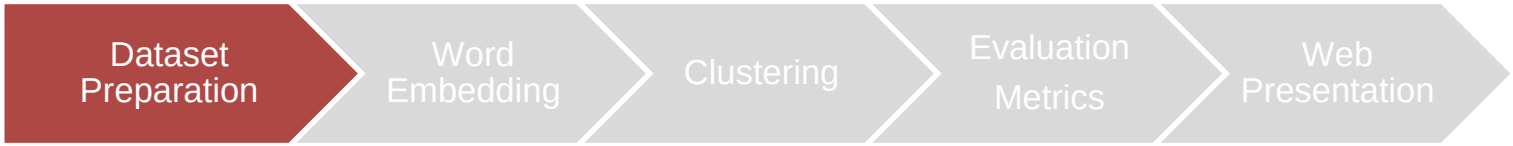
- Website demonstration
- Architecture
- Result evaluation

Dataset Preparation → Word Embedding → Clustering → Evaluation Metrics → Web Presentation

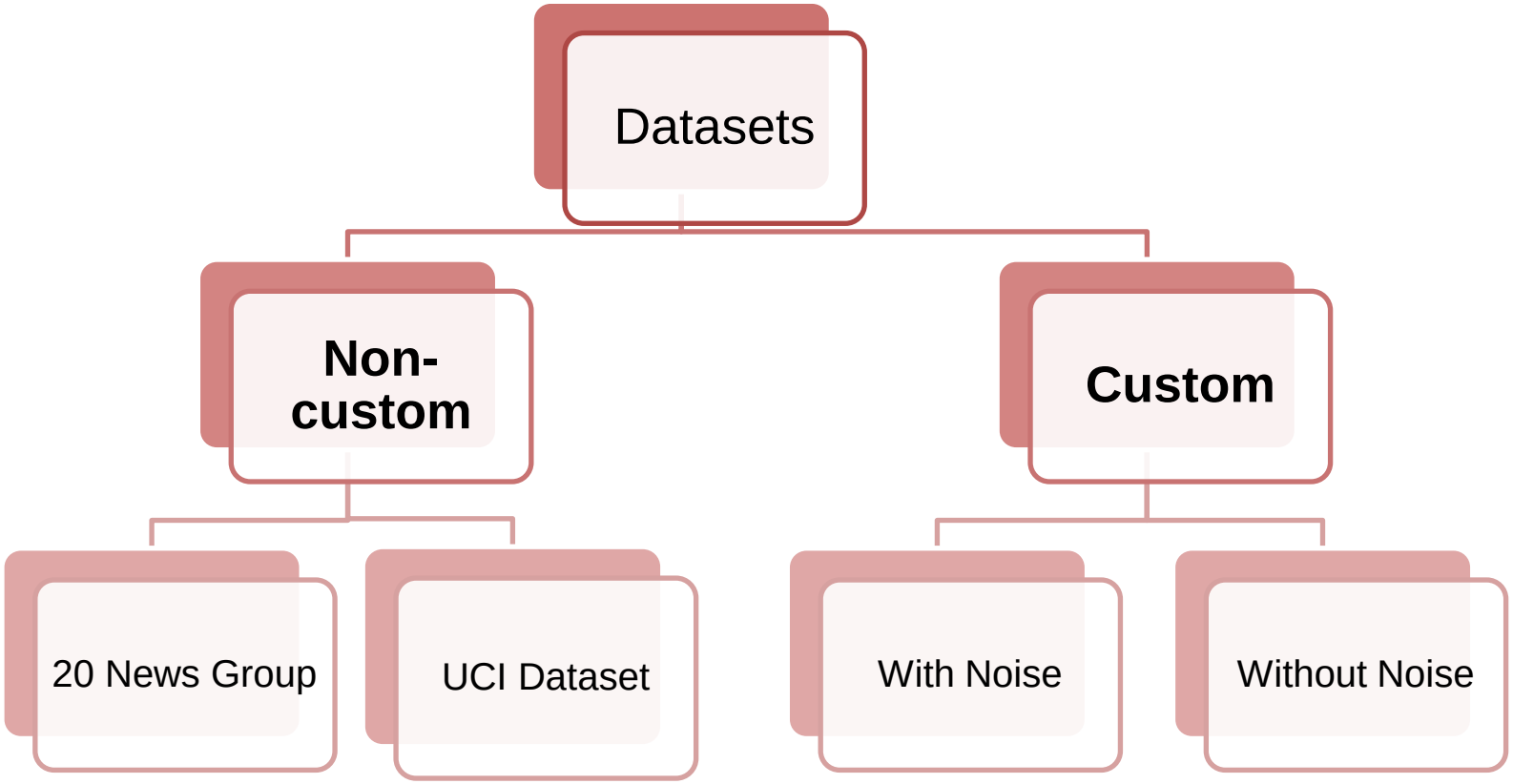
- Comparison of different word embedding methods

- Extrinsic metrics
- Intrinsic metrics
- Clustering evaluation

Dataset


Dataset Preparation
Word Embedding
Clustering
Evaluation Metrics
Web Presentation

Dataset Classification



Non-custom Dataset

Dataset	Size	Category	Attribute
20 News Group	~1000 news is chosen from 20,000 data	20 newsgroups, e.g.: • comp.graphics • rec.sport.baseball • sci.space • talk.politics.guns • soc.religion.christian	Subject; Date; Text; etc
UCI Dataset	~420,000 news stories	4 categories: • Business • Science and Technology • Entertainment • health	Headlines; URL; Category; Publisher; etc

Custom Dataset

Selected Tags:
Brexit, Cryptocurrency, Electric Vehicle, Hong Kong, US-China Relations

Noisy Data

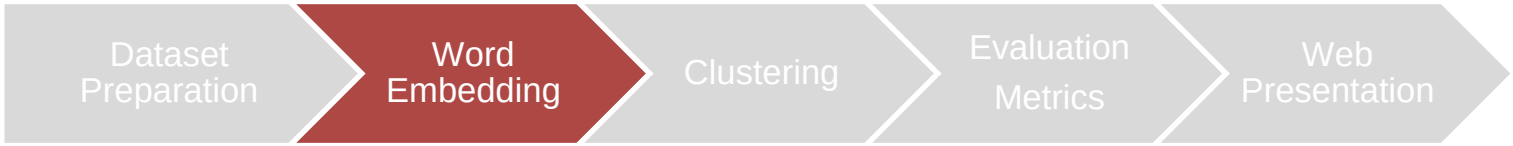
Any data that is

- Meaningless; or
- Unstructured; or
- Cannot be understood and interpreted correctly by machines

Adding noisy data to our custom dataset:

Spelling Errors	Industry Abbreviations	Slangs

Word Embedding



Word Embedding Methods

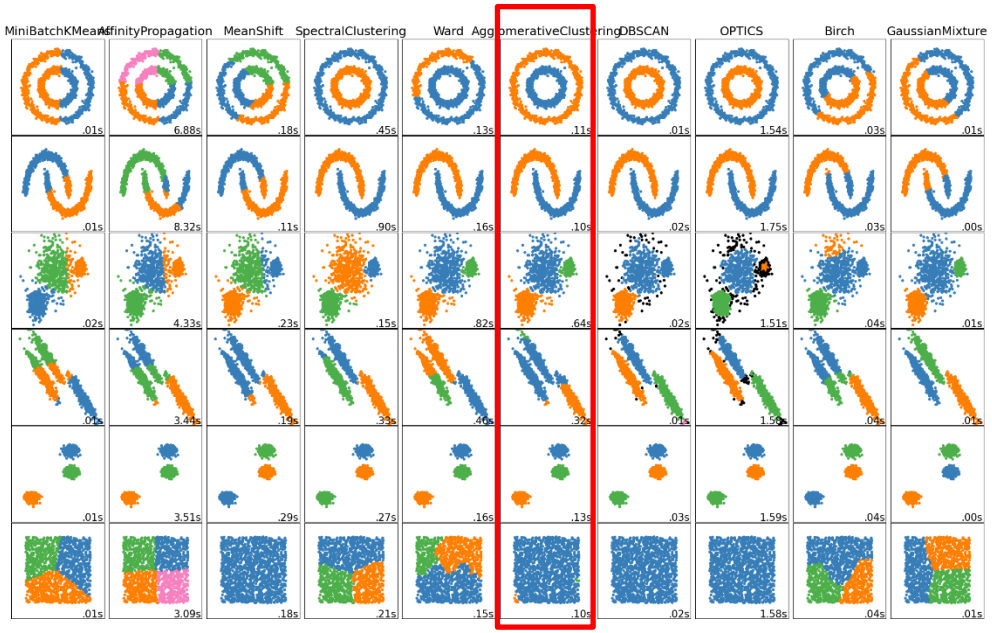
	TF-IDF	Word2Vec	Universal Sentence Encoder (USE)	BERT
WE Semantic Performance	Poor	Medium	-	High
SE Semantic Performance	Poor	Poor (Average of W2V)	High	High

Clustering

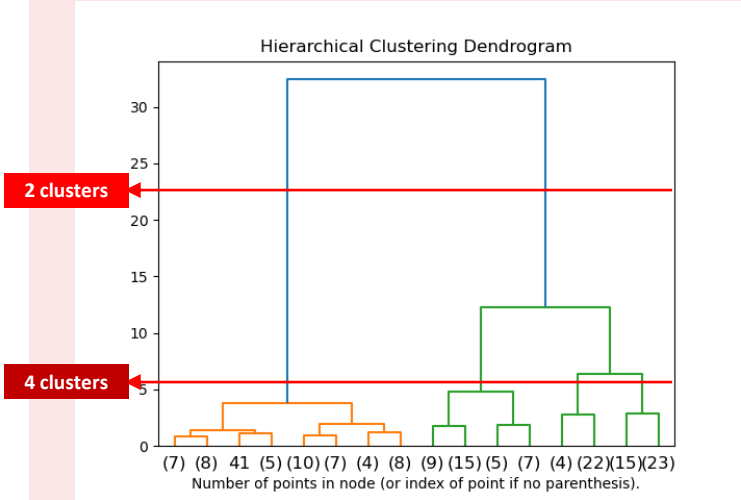

Dataset Preparation
Word Embedding
Clustering
Evaluation Metrics
Web Presentation

Recap: Hierarchical Agglomerative Clustering (HAC)

Selecting <u>Hierarchical Clustering</u> because of
(1) High Accuracy
(2) Compatibility for Cosine Similarity
(3) Ease of Use



HAC Visual Output: Dendrogram



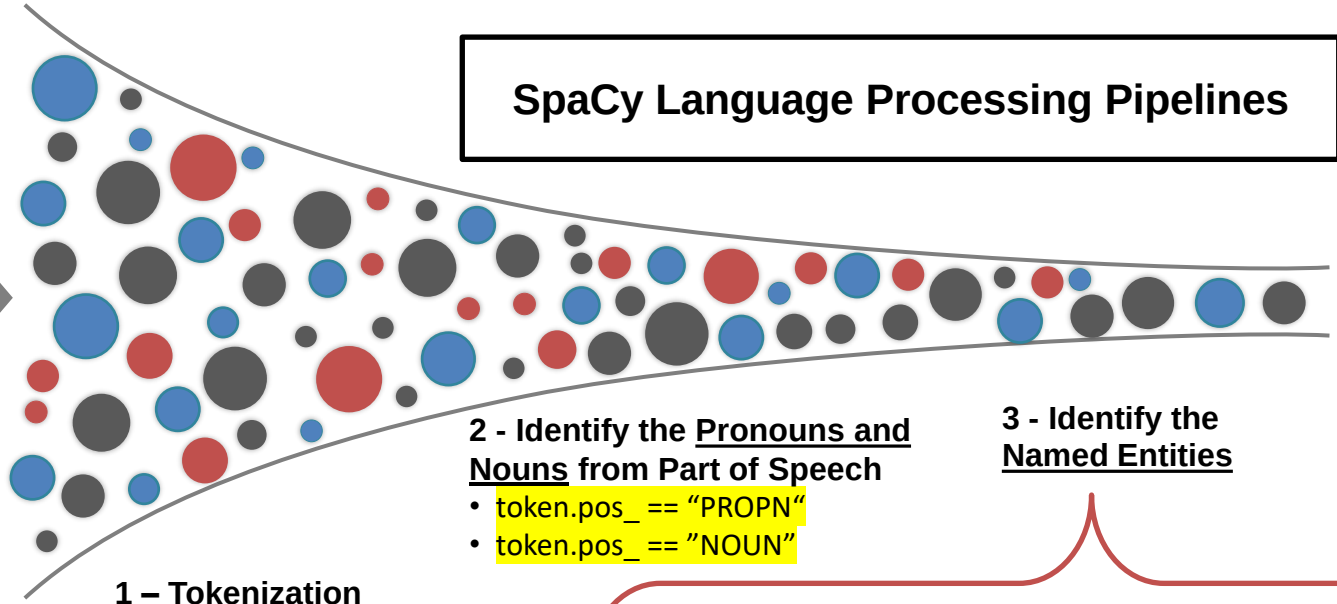
Dendrogram represents a clear overall of the structure of the clustering data.

A **linking threshold** can be varied to determine directly the number of clusters to be classified (see the horizontal lines in the above figure).

Keyword and Named Entity Recognition (NER)

Restricting the Scope of Keyword to obtain more to-the-point and meaningful insights

A News Cluster



SpaCy Language Processing Pipelines

1 – Tokenization

- spacy.load("en_core_web_lg")

2 - Identify the <u>Pronouns and Nouns</u> from Part of Speech

- token.pos_ == “PROPN“
- token.pos_ == ”NOUN”

3 - Identify the <u>Named Entities</u>

“PERSON”: People, including fictional.
“NORP”: Nationalities or religious or political groups.
“FAC”: Buildings, airports, highways, bridges, etc.
“ORG”: Companies, agencies, institutions, etc.
“GPE”: Countries, cities, states.
“LOC”: Non-GPE locations, mountain ranges, bodies of water.
“EVENT”: Named hurricanes, battles, wars, sports events, etc.

Top Key Nouns

Top Key NER

Recursive Clustering

Testing approaches of second word embedding to improve accuracy of recursive clustering

Problem

After 1st Clustering, we have

≥ 40 Fragmented Clusters

Solution

Appling HAC twice recursively to **reduce fragmentation** where similar clusters are left-unclustered.

3 Approaches for Implementation

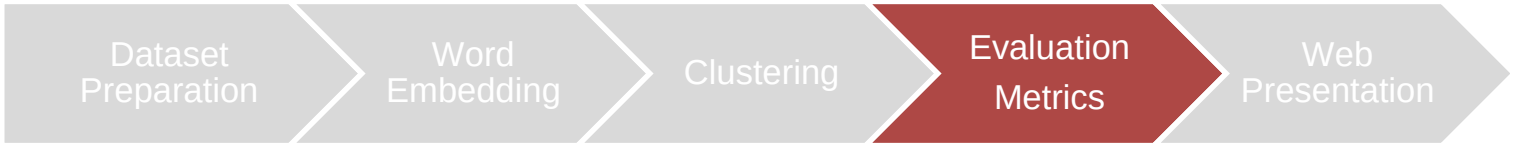
The comparison of approaches will be elaborated in our **Demonstration** later.

Trail 1	Trail 2	Trail 3
Concatenate all **raw content** in one cluster	Take the **Average** of BERT's word embedding within the cluster	Leverage the Retrieved **Nouns and NER**
Embedding of the "**Long Article Chunk**"		Embedding of the **sentence joined by Nouns and NER**

Second Hierarchical Clustering

Reference: Bambrick, J., Xu, M., Almonte, A., Malioutov, I., Perarnau, G., Selo, V., & Chan, I. C. (2020). NSTM: Real-time QUERY-DRIVEN News OVERVIEW composition at Bloomberg. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. doi:10.18653/v1/2020.acl-demos.40

Evaluation Metrics



Recap: A Basket of Evaluation Metrics

SOCIETE GENERALE

Deploying Systematic Evaluation of Clustering

	Extrinsic Metrics				Intrinsic Metrics	
	Adjusted Rand Index (RI)	**Homogeneity Score**	**Completeness Score**	**V Measure**	**Silhouette Coefficient (Unadjusted)**	**Adjusted Silhouette Coefficient (Specific to 'Cosine')**
Truth Labels Needed	✓	✓	✓	✓	-	-
Measurement	Similarity of predicted clusters and truth tables	Extent of homogeneity of members (from a single class) in each cluster	Extent of assigning members of a given class to the same cluster.	Harmonic mean of homogeneity and completeness	The ratio of intra-cluster cohesion over inter-cluster separation based on distance	**Modified** Silhouette scores specifically for measuring intra- and inter-cluster **similarity instead of distance**
Formula	$\mathrm{RI} = \frac{a+b}{C_2^{n_{samples}}}$	$h = 1 - \frac{H(C\|K)}{H(C)}$	$c = 1 - \frac{H(K\|C)}{H(K)}$	$v = 2 \cdot \frac{h \cdot c}{h + c}$	$s = \frac{b-a}{max(a,b)}$	$s = \frac{a-b}{max(a,b)}$
Outputs	**Bounded Scores within [-1,1]:** -1 for poor match; 1 for perfect match; 0 for random match	**Bounded Scores within [0,1]:** 0 for poor clustering; 1 for perfect clustering			**Bounded Scores within [-1,1]**: -1 for incorrect clustering; +1 for highly dense clustering; 0 for overlapping clusters.	
Advantages	• Proportional Interpretation • corrects for random labelling	Intuitive Interpretation			Intuitive Interpretation	Tailor-made for evaluation of similarity-based clustering
Drawbacks	Requires Truth Table	• Requires Truth Table • Not Normalized with regards to random labelling			-	-

Website Demonstration

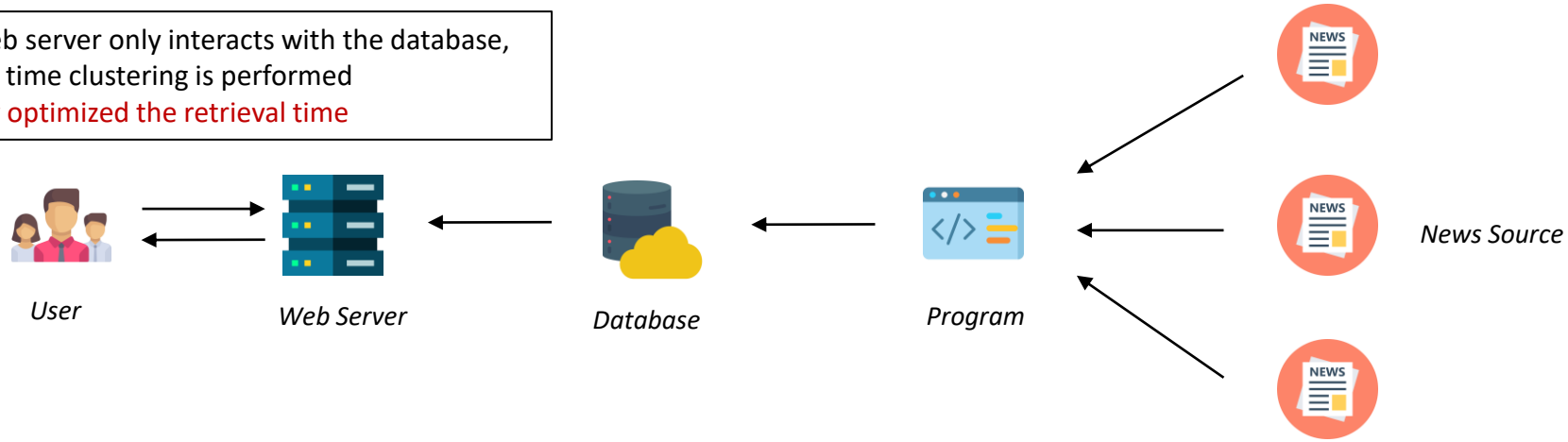
Architecture

Backend
- Clustering results are outputted as JSON format

Database
- JSON files are stored in MongoDB

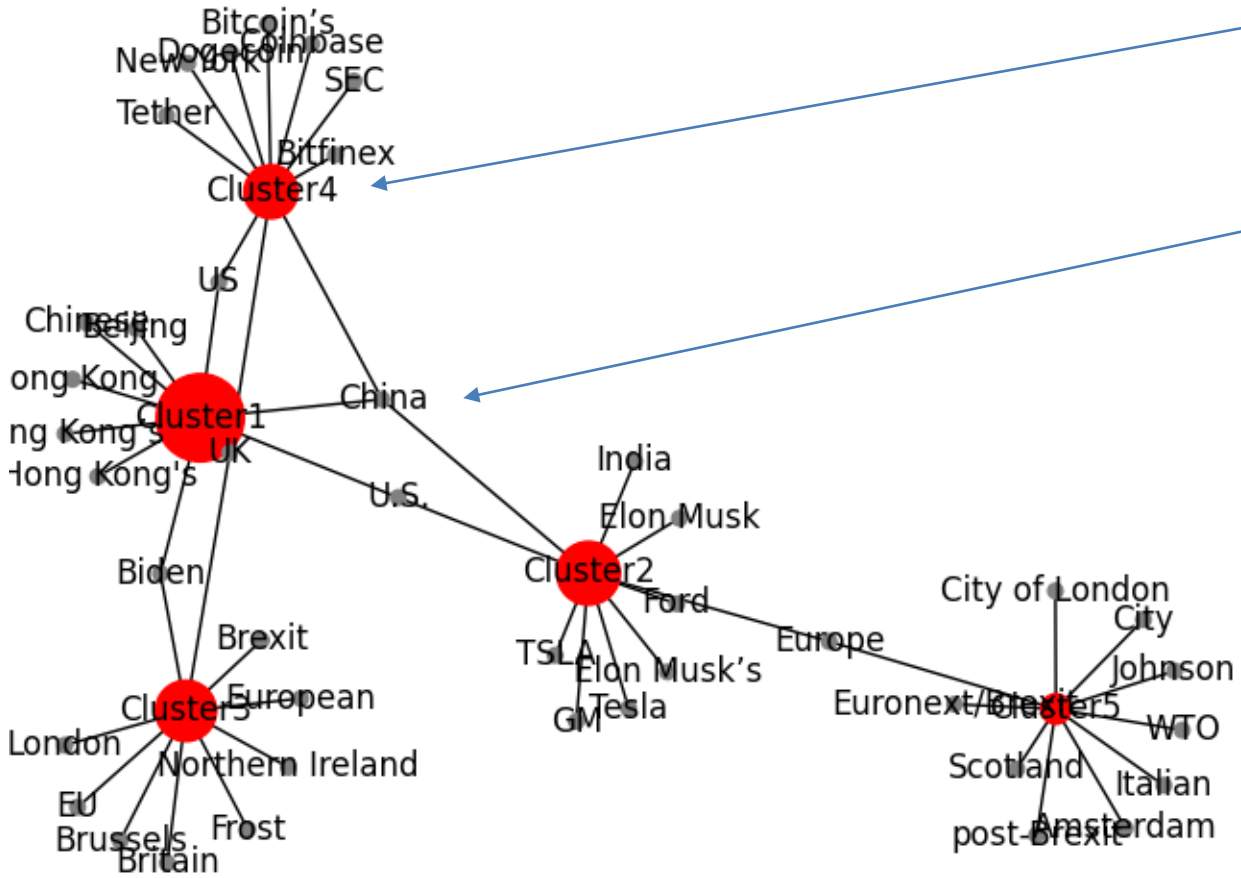
Web Server
- Gets data from database

- The web server only interacts with the database, no real time clustering is performed
- Largely optimized the retrieval time



- The program scraps news from different sources and performs clustering in a daily manner
- Saves the output at the Database

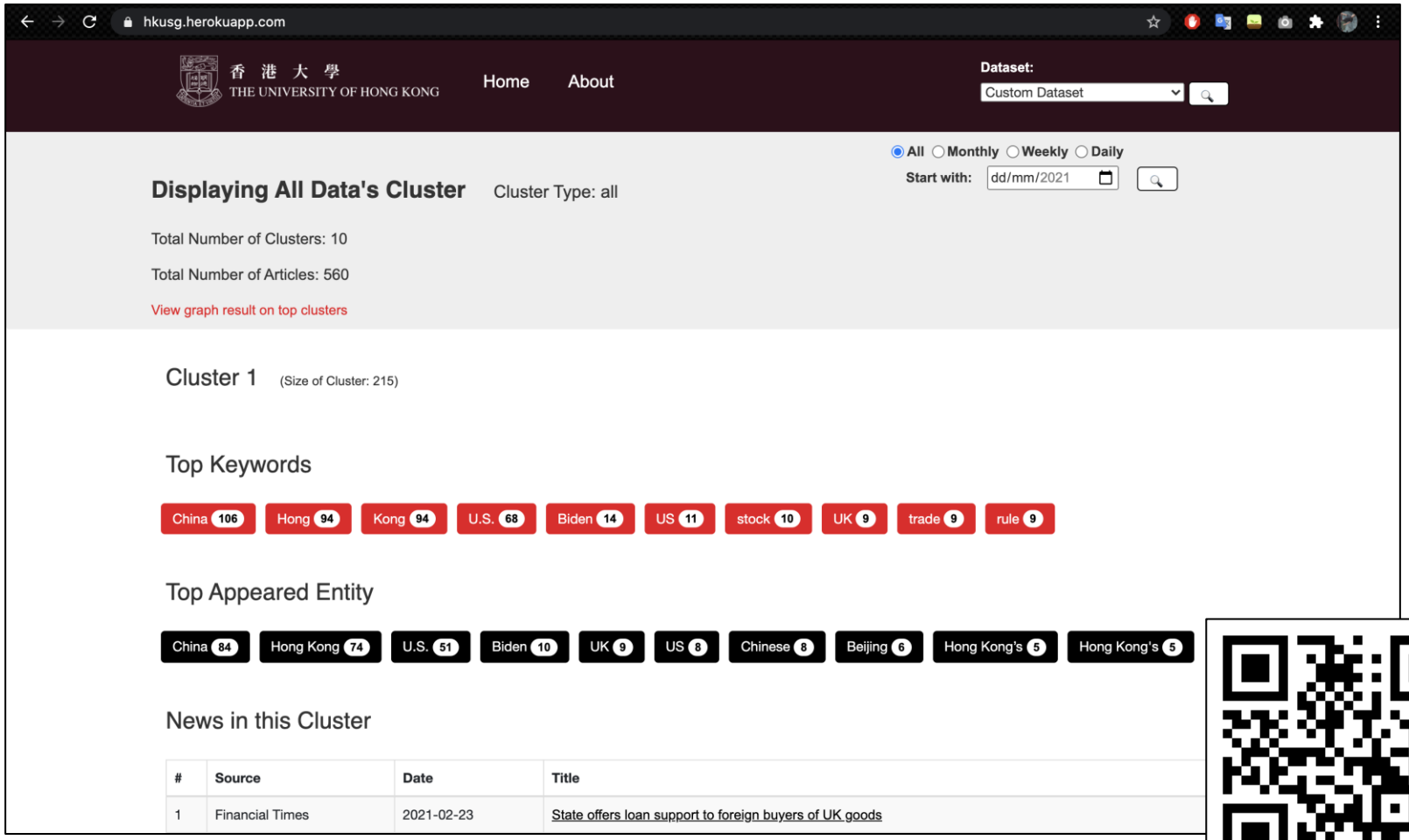
Network Visualization Graph



Node size represents the number of articles in the cluster

Keywords connect to clusters that contain it

Website Demonstration



Access our demo website

1) Enter URL in the browser http://hkusg.herokuapp.com

2) Scan the QR Code

Evaluation

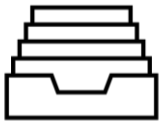
Clustering Demonstration and Evaluation

Illustrated Workflow

Data Collection

Retrieve news articles from major news providers

Word Embedding



Capture context of a word in a document

Clustering

Build a hierarchy of clusters

2nd Round Clustering

Redo the clustering

Named Entity Recognition

Website Visualization

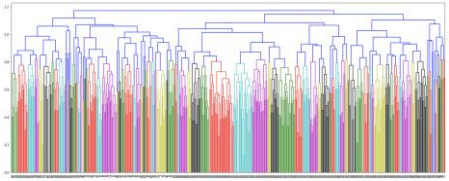
Evaluation by fixed number of clusters:
- Ignores the variance brought by the value of threshold
- Brings fairness to the evaluated models

Evaluation by a generalized threshold:
- Replicating the real-life scenario
- Evaluation on the threshold value

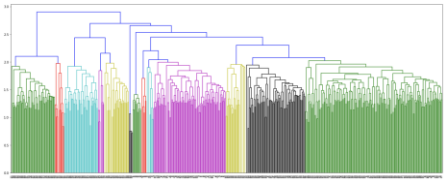
Flow of our demonstration:

1. Word Embedding Evaluation with fixed number of clusters
2. Recursive Clustering Evaluation with generalized threshold
3. Proof of the generalized threshold
4. More result evaluation with UCI News Aggregator

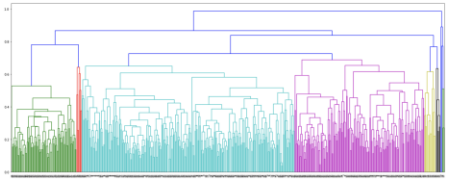
Word Embedding Evaluation with fixed number of clusters



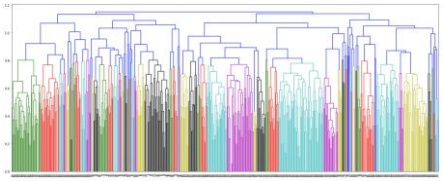
BERT



TF-IDF



Word2Vec



USE

Model	Rand	Homogeneity	Completeness	V Measure	Silhouette Score	Amended Silhouette Score	Loss
BERT	0.56	0.65	0.58	0.62	0.10	0.35	23
TF-IDF	0.13	0.42	0.24	0.30	0.01	0.14	391
W2V	0.10	0.43	0.17	0.25	0.09	0.06	465
USE	0.48	0.57	0.50	0.53	0.12	0.36	11

```
Brexit              [  5 121   3   0   6]
Cryptocurrency      [ 73   4  12  13   1]
Electric Vehicle    [  0   0   5  88   8]
Hong Kong           [  0   6  98   1   7]
US-China Relations  [  1   1 106   0   1]
```

BERT

```
Brexit              [  3   0   0  11 121]
Cryptocurrency      [  0  14   0   0  89]
Electric Vehicle    [  1  66   0   0  34]
Hong Kong           [ 60   1   0   0  51]
US-China Relations  [  5   3   5   0  96]
```

TF-IDF

```
Brexit              [  0 130   2   2   1]
Cryptocurrency      [  5  81  15   2   0]
Electric Vehicle    [  0  99   2   0   0]
Hong Kong           [ 80  32   0   0   0]
US-China Relations  [  6 101   2   0   0]
```

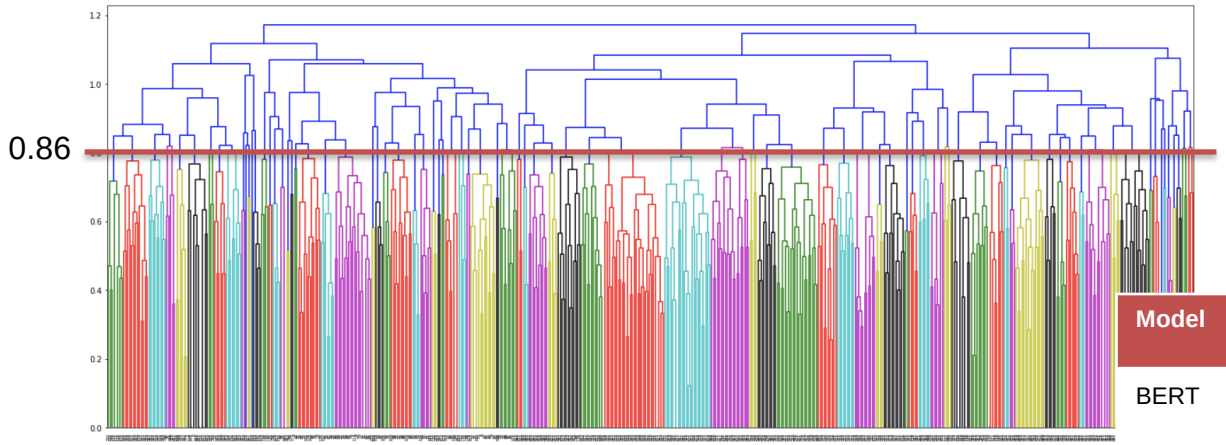
Word2Vec

```
Brexit              [  0 109  11   2  13]
Cryptocurrency      [  9   5  11   9  69]
Electric Vehicle    [ 89   2   1   0   9]
Hong Kong           [  4   2  91   0  15]
US-China Relations  [  1   0 105   0   3]
```

USE

Recursive Clustering Evaluation with generalized threshold

Realistically, we will never know the number of clusters



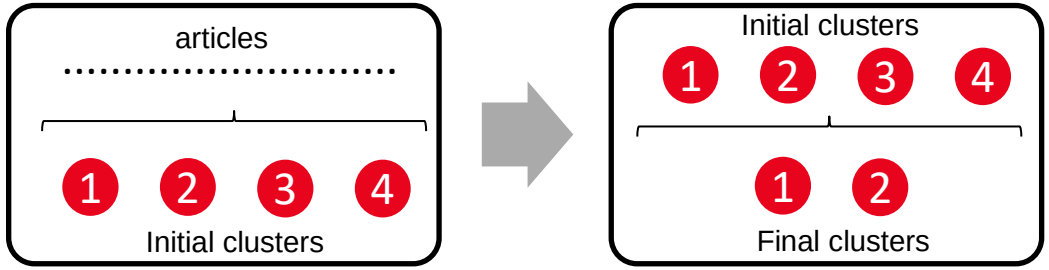
Model	Num of Clusters
BERT	61
USE	50

Challenges:

- There are too many clusters
- There is no one-size-fits-all threshold
  - Smaller threshold is needed when news are concentrated
  - Vice versa when news appear in multiple topics

Despite this, at least, each cluster is a correctly classified sub-cluster
→ Need to explore method to merge the clusters

Recursive Clustering Evaluation with generalized threshold



Sequence:
"Hong Kong" -> "Kong Hong"

Weight:
("China", 100) ("US", 93)
("Iran", 5)

-> "China US Iran"

How do we convert clusters into vectors?

	Approach 1	Approach 2	Approach 3
Description	Concatenate the articles Article1 Article2 Article3 → One Large Article	Average of the embeddings Embedding1 Embedding2 Embedding3 } Average Of Embedding	Extract the top nouns "HK announced election" "China blocks US" } "HK election China US"
Result	Number of Clusters: 14 Brexit [7 3 1 0 8 101 1 1 3 4 0 4 1 1] Cryptocurrency [0 0 0 0 2 3 1 7 84 0 0 1 5 0] Electric Vehicle [0 0 4 0 0 3 1 0 17 0 74 0 2 0] Hong Kong [0 2 1 1 0 97 3 0 0 2 1 0 5 0] US-China Relation[0 0 0 0 0 2 0 0 2 0 0 14 91 0] Adjusted Rand Score: 0.51	Number of Clusters: 10 Brexit [6 0 1 1 0 94 22 2 3 6] Cryptocurrency [0 0 0 0 15 1 3 5 76 3] Electric Vehicle [0 0 0 0 95 1 0 1 0 4] Hong Kong [2 1 3 0 1 4 0 3 0 98] US-China Relations [0 0 0 0 1 3 0 0 1 104] Adjusted Rand Score: 0.54	Number of Clusters: 10 Brexit [12 2 13 98 0 0 3 3 3 1] Cryptocurrency [18 1 2 0 0 0 73 3 6 0] Electric Vehicle [15 19 0 0 0 59 0 5 3 0] Hong Kong [6 0 101 0 1 1 0 2 1 0] US-China Relations [0 1 12 0 0 0 1 94 1 0] Adjusted Rand Score: 0.61
Evaluation	• Difficult to control the size of the article • The larger the article is, the more inaccurate the embedding will be Abandoned	• Easy to implement • Safe control • Acceptable performance Chosen	• Best results of three • Hard to ensure the sequence of words • Cannot demonstrate the weight of words Requires further study

Proof of the generalized threshold

What threshold value should we choose?

10 Combinations: ("Brexit", "Cryptocurrency"), ("Brexit", "Electric Vehicle"), ("Brexit", "Hong Kong"),
... ("Electric Vehicle", "US-China Relations"), ("Hong Kong", "US-China Relations")

Threshold: [0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1, 1.05, 1.1, 1.15, 1.2]

Threshold	Brexit v Crypto	Brexit v EV	Brexit v HK	Brexit v US-China	Crypto v EV	Crypto v HK	Crypto v US-China	EV v HK	EV v US-China	HK v US-China	Average
0.6	0.087	0.088	0.096	0.126	0.058	0.088	0.127	0.091	0.112	0.108	0.098
0.65	0.134	0.149	0.159	0.211	0.105	0.332	0.291	0.224	0.167	0.306	0.208
0.7	0.233	0.328	0.402	0.390	0.223	0.409	0.450	0.347	0.421	0.513	0.372
0.75	0.504	0.431	0.430	0.581	0.465	0.457	0.634	0.462	0.675	0.644	0.528
0.8	0.589	0.648	0.553	0.696	0.594	0.773	0.764	0.811	0.804	0.820	0.705
0.85	0.704	0.735	0.679	0.716	0.629	0.821	0.882	0.889	0.943	0.868	0.787
0.9	0.845	0.924	0.006	0.774	0	0.890	0.907	-0.001	0	0.000	0.435
0.95	0	-0.002	0	-0.002	0	-0.002	0	-0.001	0	0.000	-0.001
1	0	0	0	0	0	0	0	0	0	0	0
1.05	0	0	0	0	0	0	0	0	0	0	0
1.1	0	0	0	0	0	0	0	0	0	0	0
1.15	0	0	0	0	0	0	0	0	0	0	0
1.2	0	0	0	0	0	0	0	0	0	0	0

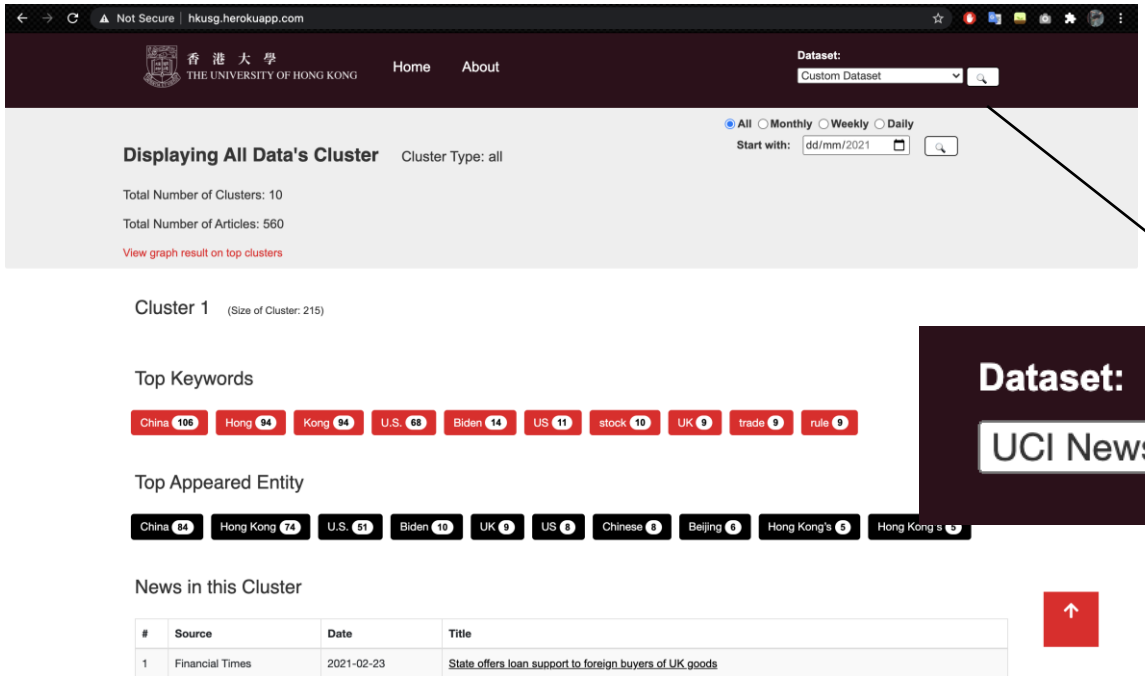
Adjusted Rand Score

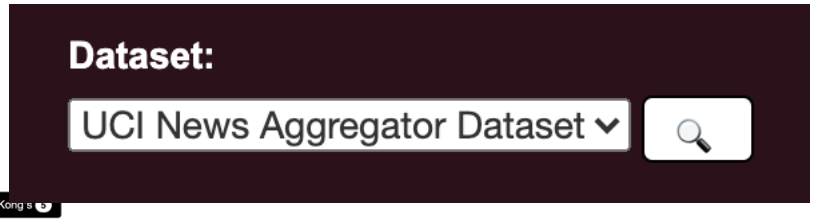
More result evaluation with UCI News Aggregator

400k news in 2014

Categories:

- Business
- Technology
- Entertainment
- Health





More result evaluation with UCI News Aggregator

2014 March 11 Number of Articles: 4291 Number of Clusters: 36

Justin Bieber 86 | Stacy Keibler 83 | Miley Cyrus 57 | Selena Gomez 44 | George Clooney's 32 | Justin Bieber's 18 | SXSW 16 | Bieber 13

Entertainment
Justin Bieber, Stacy Keibler

Titanfall 107 | Apple 68 | iOS 7.1 28 | Microsoft 18 | CarPlay 17 | Xbox One 16 | Apple TV 16

Technology
Apple, Microsoft, XBOX

Juan Pablo Galavis 62 | Juan Pablo 58 | Nikki Ferrell 25 | Chris Harrison 16 | Nikki 14 | Juan Pablo's 5 | Juan Pablo 5

Entertainment
Juan Pablo, Nikki Ferrell

Mt. Gox 55 | US 43 | EBay 18 | Icahn 16 | eBay 12 | EU 10 | Carl Icahn 9

Business
Cryptocurrency, eBay

China 90 | US 36 | McDonald 25 | Chinese 25 | Asia 15 | Asian 10 | Hong Kong 10

Business
US-China, Asia, HK, Japan

Takeaway

“We successfully cluster news titles in a general manner without training”

Takeaway

"We successfully cluster news titles in a general manner without training"

Values to Société Générale

Helps understanding the market from the myriad of news

- Serves as the first line of analysis in the incoming news
- Systematically understands the market behaviour and anomalies

Assists in the research for future NLP application

- Clustering is an important component in most NLP applications
- Word Embedding can be embedded in other NLP algorithms as well

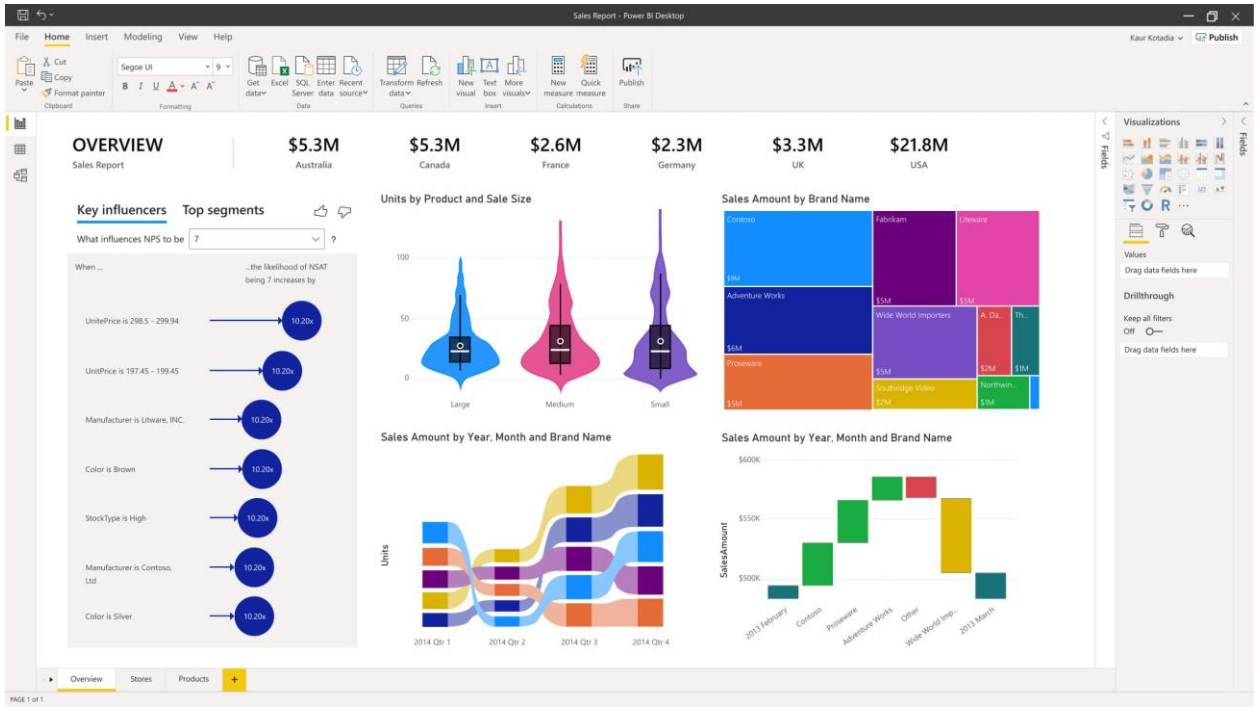
Future Plans

Future Plans

Advanced Visualization

Leverage dashboard to easily understand latest market news

- Enables users to interact with the dashboard
- Looking into the details of information with just a click
- E.g. Network visualization graph



Challenge

- Customized use case shall be investigated with SG analysts

Future Plans

Expand the scope of clustering

Enable clustering in both general / specific dataset

- Currently only focus on general news, for identifying the topics
- Ideally: Able to work on any dataset, ex. Clustering on search result
- E.g. Bloomberg use case, where Canada → Covid/ Political/ Oil Price

Challenge

- Real time clustering is desired
- Dynamic threshold is desired

Q&A