# LIFESPAN of

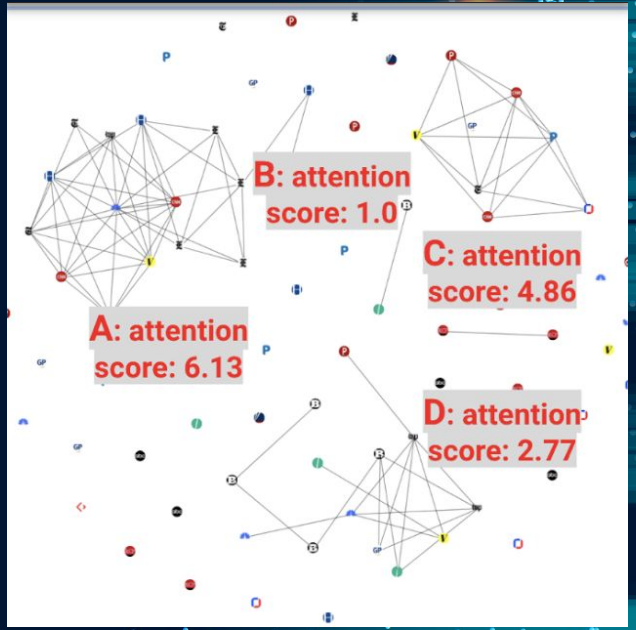
## NEWS STORIES

A NLP Approach for Extracting Trending News

Hassan Koroma

# StoryGraph: Algorithm

- Implemented by Prof. Nwala
- Measures news similarity in near-real time & quantifies attention of news stories extracted from 17 US news sources
- Clusters these articles based on set similarity
- Displays clusters in a network graph
- Refreshes every 10 minutes




storygraph.cs.odu.edu

# Research Questions

**Define a news story**: the main topic of the article, so focus on the headlines.

1. What is the duration of the shelf-life of news stories?
2. Among news stories covering political, environmental, social issues, violence and war, which ones tend to have a longer lasting public view and which ones tend to fade away quickly?

**Hypothesis**: News stories on political issues will have a longer shelf-life

# Data Sources

**3 News Agencies (from Kaggle) :**

- Reuters (financial news)
- The Guardian
- CNBC

**Data Structure:**

- About 32k news articles from 2018 - 2020
- Each article contains the headline, a short description, and publishing time

| Headlines | Time | Description |
|---|---|---|
| TikTok considers London and other locations fo... | Jul 18 2020 | TikTok has been in discussions with the UK gov... |
| Disney cuts ad spending on Facebook amid growi... | Jul 18 2020 | Walt Disney has become the latest company to ... |
| Trail of missing Wirecard executive leads to B... | Jul 18 2020 | Former Wirecard chief operating officer Jan M... |
| Twitter says attackers downloaded data from up... | Jul 18 2020 | Twitter Inc said on Saturday that hackers were... |
| U.S. Republicans seek liability protections as... | Jul 17 2020 | A battle in the U.S. Congress over a new coron... |

# Data Pipeline

| Preprocessing | Keyword Extraction | News Clustering | Visualize Trending Stories |
| --- | --- | --- | --- |
| • Remove Unwanted Text<br>• Date Normalization<br>• Lemmatize | • NER<br>• Noun Phrases<br>• Keyword Scoring<br>• Keyword Filtering<br>• Postprocessing | • Keyword Vectorization<br>• News Similarity<br>• DBSCAN Algo. | • Cluster Time Series News<br>• Visualize Top Trending Stories |

# Preprocessing

- Removed unwanted texts
  - Non-english characters, unusual headline patterns, etc.
- Date Normalization
  - M/D/Y: “Jul 18 2020”
- Lemmatization
  - Kept stopwords for noun phrase detection (“The U.K”)

```
Disney cuts ad spending on Facebook amid growing boycott: WSJ
FTC considering deposing top Facebook executives in antitrust probe: WSJ
```

# Keyword Extraction (main task)

**Spacy: Name Entity Recognition (NER)**

- Extract named entities with term frequency to reflect key points of news story (PERSON, ORG, GPE, Noun phrases, etc)

**Keyword Scoring Metric**

- Different weights are used depending on keyword type (entity or noun phrase)
- Keywords found in headlines weighs more than those in the content (entity = 4, noun chunks = 2, other = 1)

**Keyword Filtering**

- Remove stopwords, special characters , news agency in headline/content, etc.

**Postprocessing**

- Abbreviate long entities for visualization

$$Score_k = W_{type} \times T_{k,\,in\ title} + T_{k,\,in\ content}$$

where

$T_k$ : the number of times of keyword $k$ appeared in title or content

$W_{type}$ : the weight of the keyword type

```
keywords_linking_table = {
    "United States": "US",
    "United Nations": "UN",
    "European Union": "EU",
    "United Kingdom": "UK",
    "European Central Bank": "ECB",
```

# Exploratory Data Analysis

- Over 200k keywords extracted in total
- Top 30 keywords with counts
- Highlighted keywords in the same color are in the same entity

| | Keyword | Count |
|---|---|---|
| 0 | US | 10903 |
| 1 | China | 4435 |
| 2 | United States | 2037 |
| 3 | Chinese | 1678 |
| 4 | Trump | 1619 |
| 5 | company | 1323 |
| 6 | German | 1119 |
| 7 | Boeing | 1093 |
| 8 | Fed | 850 |
| 9 | Britain | 774 |
| 10 | investor | 774 |
| 11 | EU | 761 |
| 12 | deal | 745 |
| 13 | British | 703 |
| 14 | Japan | 679 |
| 15 | French | 679 |
| 16 | Federal Reserve | 679 |
| 17 | share | 670 |
| 18 | MAX | 659 |
| 19 | Europe | 657 |
| 20 | UK | 653 |
| 21 | Tesla | 630 |
| 22 | Germany | 629 |
| 23 | plan | 627 |
| 24 | coronavirus | 620 |
| 25 | European | 613 |
| 26 | Huawei | 597 |
| 27 | talk | 581 |
| 28 | Facebook | 555 |
| 29 | Wall Street | 553 |
| 30 | Apple | 540 |

# News Clustering: Unsupervised

**Vectorization**

- Convert the keywords into numerical representation using sklearn's *CountVectorizer* or *HashingVectorizer* depending on number of articles to be clustered
- Score of extracted keyword is the term frequency in vectorizer

**Cosine Similarity**

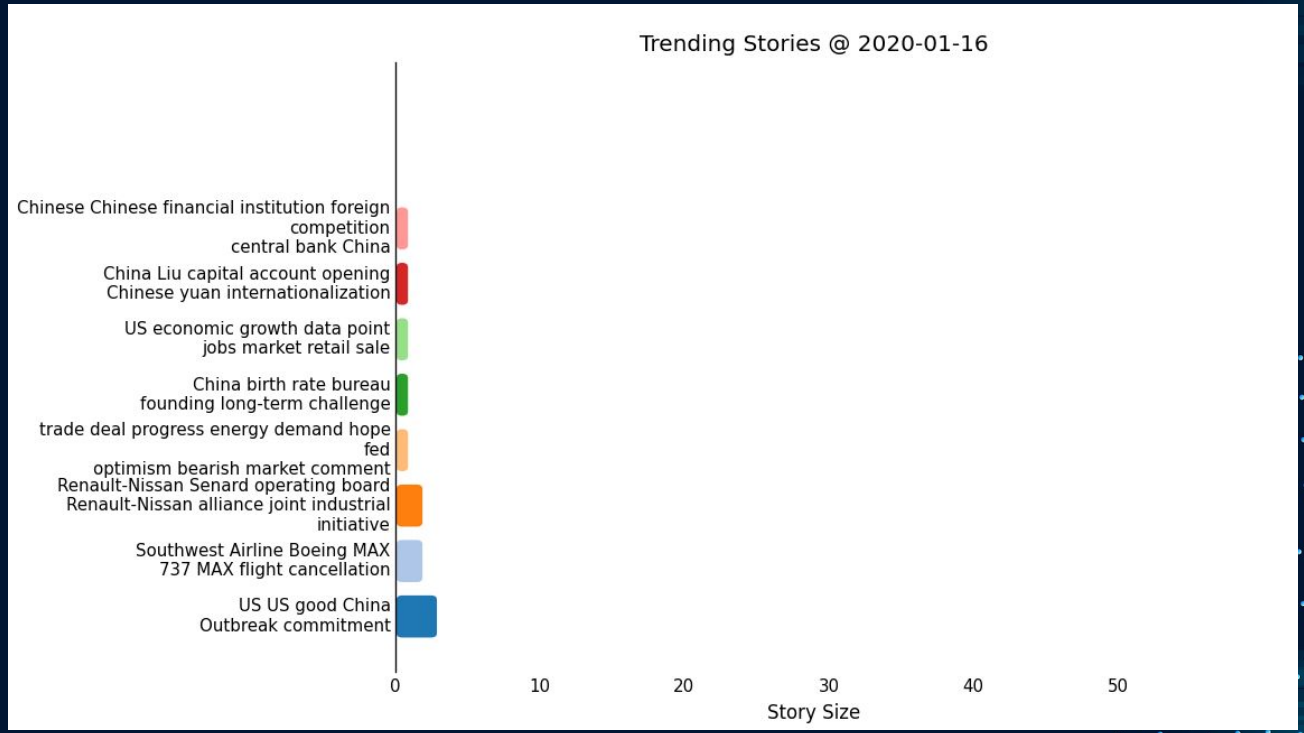
- Measure similarity between the keywords from two news articles

**DBSCAN**

- Cluster news stories based on similarity score
- No predetermined number of stories in a cluster

# DEMO

# Time Series


Trending Stories @ 2020-01-16
Chinese Chinese financial institution foreign competition central bank China
China Liu capital account opening Chinese yuan internationalization
US economic growth data point jobs market retail sale
China birth rate bureau founding long-term challenge
trade deal progress energy demand hope fed optimism bearish market comment
Renault-Nissan Senard operating board Renault-Nissan alliance joint industrial initiative
Southwest Airline Boeing MAX 737 MAX flight cancellation
US US good China Outbreak commitment
0
10
20
30
40
50
Story Size

# Findings

- Mnuchin says U.S. won't start a trade war: Brazil's Meirelles
- G20 talks on trade 'constructive,' no concern of trade war: Argentina
- U.S., Mexico resume talks to avert tariffs as deadline approaches
- Trade war, tariffs pose risks to U.S. and global growth: IMF, Fed officials
- EU readies new trade retaliation list before Trump visit
- USTR proposes $4 billion in potential additional tariffs over EU aircraft subsidies
- China announces new tariff waivers for some U.S. imports
- Trump prepares for 'productive' talks with Xi on trade war
- Factbox: Winners and losers in Trump's trade war with China
- China to slap additional tariffs on $16 billion of U.S. goods
- China rare earth prices soar on their potential role in trade war
- Oil prices slide as U.S.-China trade war escalates

## Q1: Analyzing trending news story duration

1. Trade war among various countries
    - US-China trade talks lasted for 8 months
2. Boeing 737 MAX 8 aircraft crash (also lasted for 8 months)
3. Covid-19 gained traction in February 2020
    - Rapid growth since then

## Q2: Model accepts hypothesis

- Political news stories (especially on the US economy) last longer

# Challenges + Future Work

- Memory & runtime issues due to large datasets (~1hr runtime)
- Interactivity: zooming, filtering, alternative visual representations (scope of course)
- Scalability: Ensure algorithm scales with growing dataset
    - Integrate database systems (MongoDB, PostgreSQL)
    - Check for disparity in recent trending stories
- Web Application
    - Search Engine: users can search and view different timelines
    - Custom Visualization

# References


- Tom Nicholls & Jonathan Bright (2019) Understanding News Story Chains using Information Retrieval and Network Clustering Techniques, Communication Methods and Measures, 13:1, 43–59, DOI: 10.1080/19312458.2018.1536972
- Towards Data Science
  - https://towardsdatascience.com/extract-trending-stories-in-news-29f0e7d3316a
- StoryGraph by Alexander C. Nwala
- Kaggle Dataset
  - https://www.kaggle.com/notlucasp/financial-news-headlines


**Code is available in my Github repo** - ankoroma

# Thank you!

Q/A