COLUMBIA UNIVERSITY

Capstone Presentation

08/09/18

Xiaoran Cai, Sarah Wu,

Chinmayee Mohapatra, Yerin Jung,

Vatsal Randhar, Aditi Wadhawan,

Kylene Ye, Thibault Twahirwa

Revised by:

Kylene Ye

Agenda

COLUMBIA UNIVERSITY

Review of last presentation

COLUMBIA UNIVERSITY

Business Dilemma

- Resources underutilized
- Drop in viewership
- Increased competition
- Opportunity to attract broader and younger audience

+

Data Exploration

- Data type: unstructured script data
- Competitor comparison
  - Word association
  - Sentiment analysis

→

Solution and Plan

- Inference through analysis
- Evaluation Criteria:
  - Information Entropy
  - Penalized Difference Count

Executive Summary

Business Impact

- Generate more revenue by appealing to wider and younger audiences
- Optimization of studio and other resources through better Content
- Augmented decision making
- Increased ad Revenue from increased Viewership

Technical Challenge

- Lack of comparative or benchmark data (scripts from competitors) for market analysis
- Lack of structured data

Description of Solution

- Facilitate topic selection for content generation and management

Business Goals

Columbia University

Wider audience capture by providing better content

Augment topic selection for decision-making

Increased revenue from higher viewership

Better utilization of resources by generating more content

Project Overview

COLUMBIA UNIVERSITY

Our Overall Goal

Increase the **effectiveness of topic selection** at XXX for content generation

What topics are/aren't being discussed:
Repetition frequency

The main new element in our strategy: ***Volatility***

Differences in frequency of associated words -> Potential Uncaptured Trend

Logic Flow Diagram

COLUMBIA UNIVERSITY

Scrape data from external sources

Find the popular topics in Stocktwits (ST)

Find the frequency of these topics in XXX and XXX transcripts

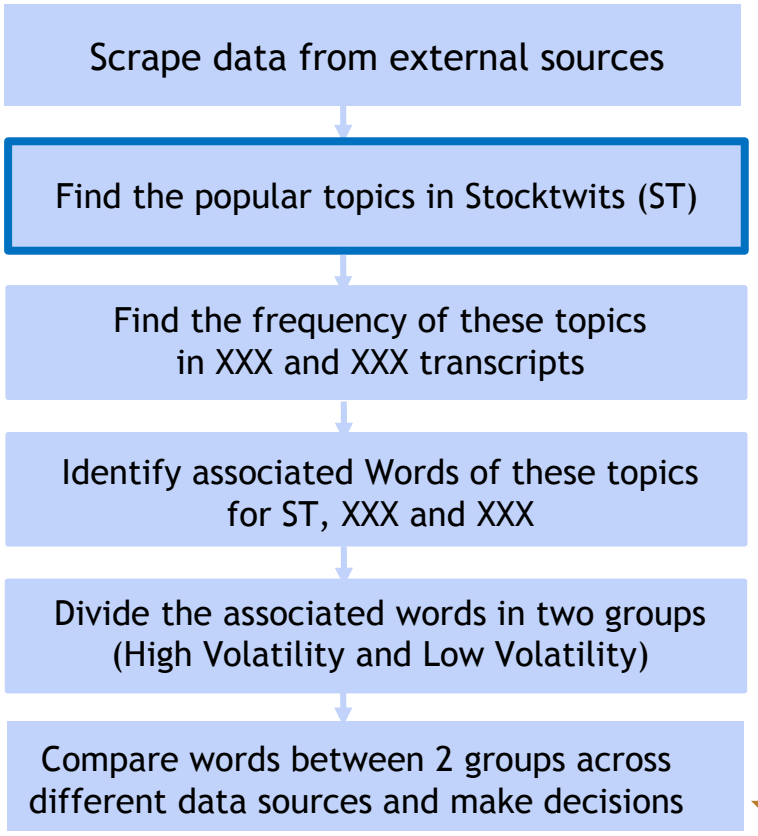
Identify associated Words of these topics for ST, XXX and XXX

Divide the associated words in two groups (High Volatility and Low Volatility)

Compare words between 2 groups across different data sources and make decisions

- ***Purpose***
  Obtain external data for analysis
- **Data from XXX**
  Unstructured data: full year news transcript
- **Extra data**
  StockTwits and transcript data of XXX for the year 2017

Logic Flow Diagram



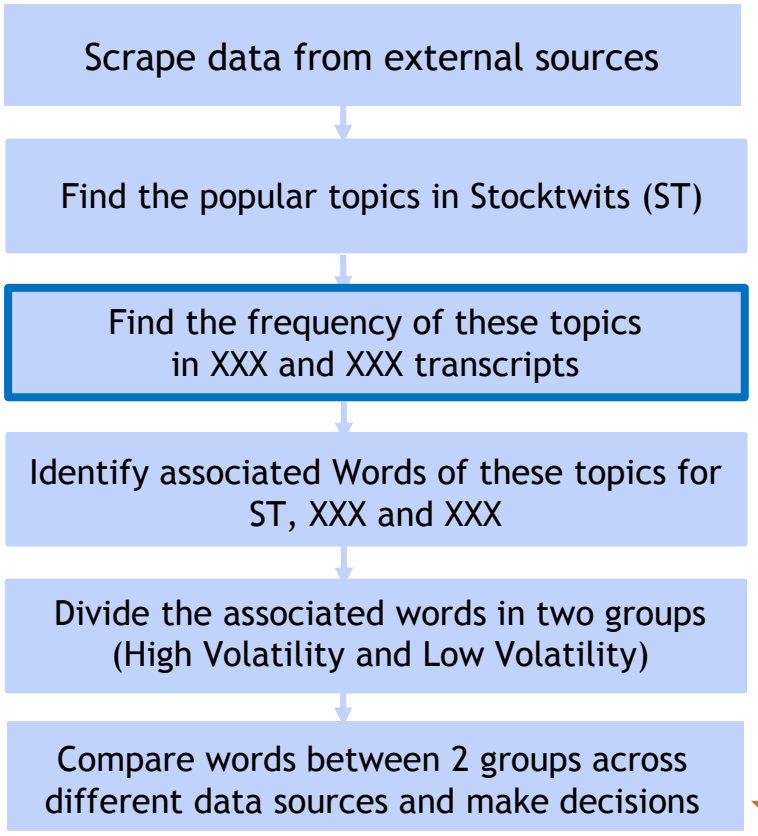
- ***Purpose***
  Find out the monthly topics (benchmark topics) that specific demographics are interested in .

- ***Procedures:***
  (1) Use ST frequent words as benchmark for profit-generating topics

  (2) Divide the data into monthly data

  (3) Use Word2Vec to generate the topics from ST transcript

Logic Flow Diagram

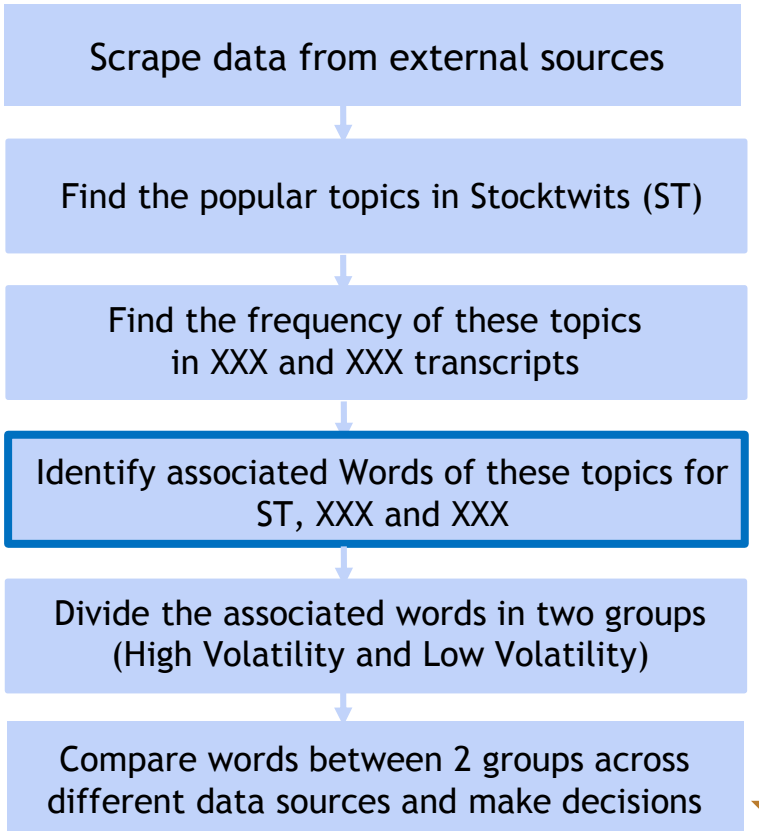


- ***Purpose***
  Find out the frequency of the benchmark topics in the XXX and XXX data

- ***Procedures:***
  Use Word2Vec to generate the frequency for each month

Logic Flow Diagram

COLUMBIA UNIVERSITY



- ***Purpose***
  Find content that can be delivered together with the topics that meet the interests of the target audience.

- ***Techniques***
  Gensim: word association

Logic Flow Diagram

COLUMBIA UNIVERSITY

Scrape data from external sources

Find the popular topics in Stocktwits (ST)

Find the frequency of these topics in XXX and XXX transcripts

Identify associated Words of these topics for ST, XXX and XXX

Divide the associated words in two groups (High Volatility and Low Volatility)

Compare words between 2 groups across different data sources and make decisions

- ***Purpose***
  For each pair of topic and its associated word, calculate their volatility score and divide into two categories:
  (1) High Volatility
  (2) Low Volatility

- ***Techniques***
  Custom algorithm

Logic Flow Diagram

COLUMBIA UNIVERSITY

Scrape data from external sources

Find the popular topics in Stocktwits (ST)

Find the frequency of these topics in XXX and XXX transcripts

Identify associated Words of these topics for ST, XXX and XXX

Divide the associated words in two groups (High Volatility and Low Volatility)

Compare words between 2 groups across different data sources and make decisions

By comparing the differences of groups we recommend content by identifying associated topics:

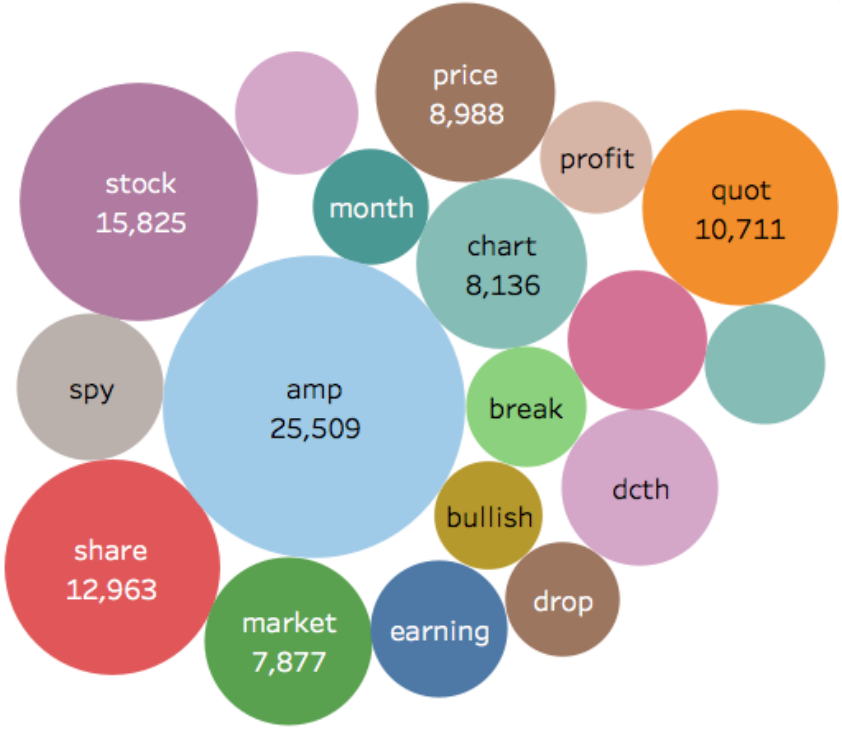
<u>Low volatility difference:</u> Appeared in Low Volatility of StockTwits or XXX, but not in Low volatility of XXX.

<u>High volatility difference:</u> Appeared in High Volatility of StockTwits or XXX, and also in low volatility of XXX.

Data Analysis

This section contains examples of topics discovered in our following analysis:

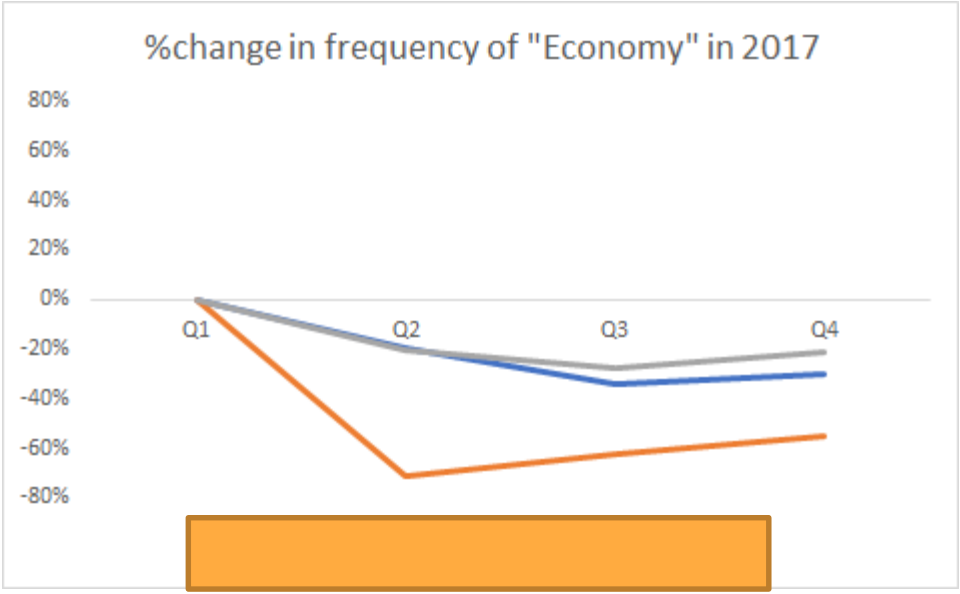
- Frequency change over time
- Word Association over time
- High/Low Volatility group division



“Economy” - Frequency

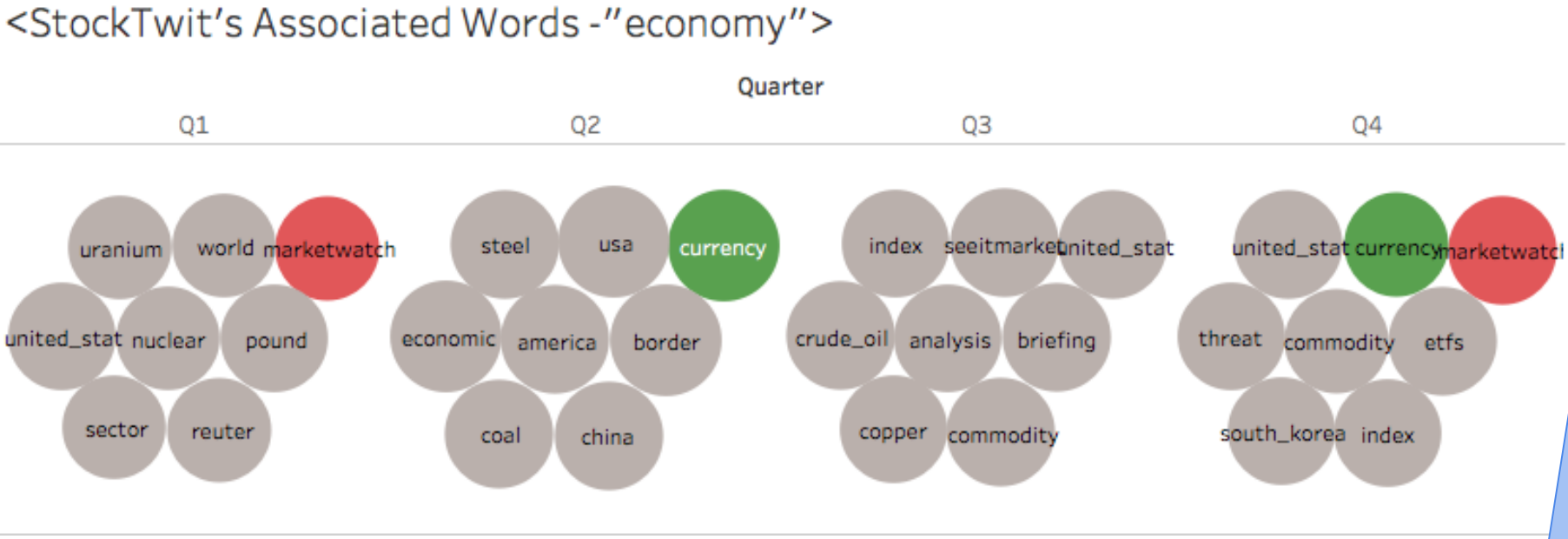
Inference:

- XXX and Stocktwits are closer to each other compared to XXX.



“Economy” - Associations

COLUMBIA UNIVERSITY

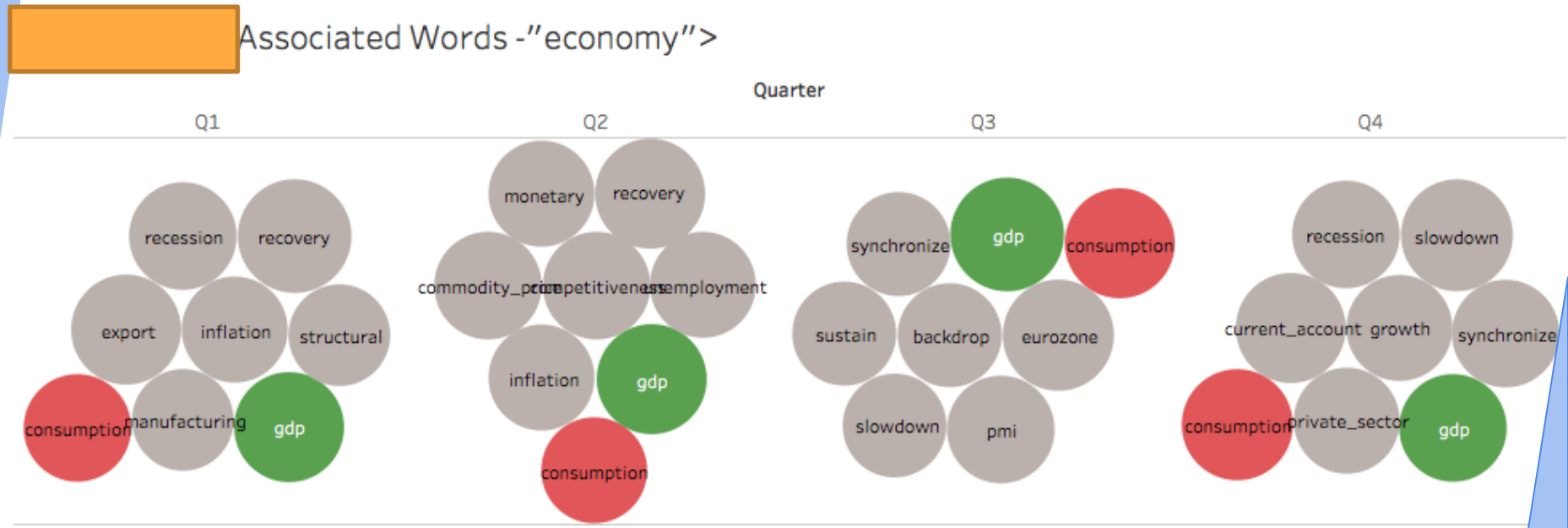


“Economy” - Associations

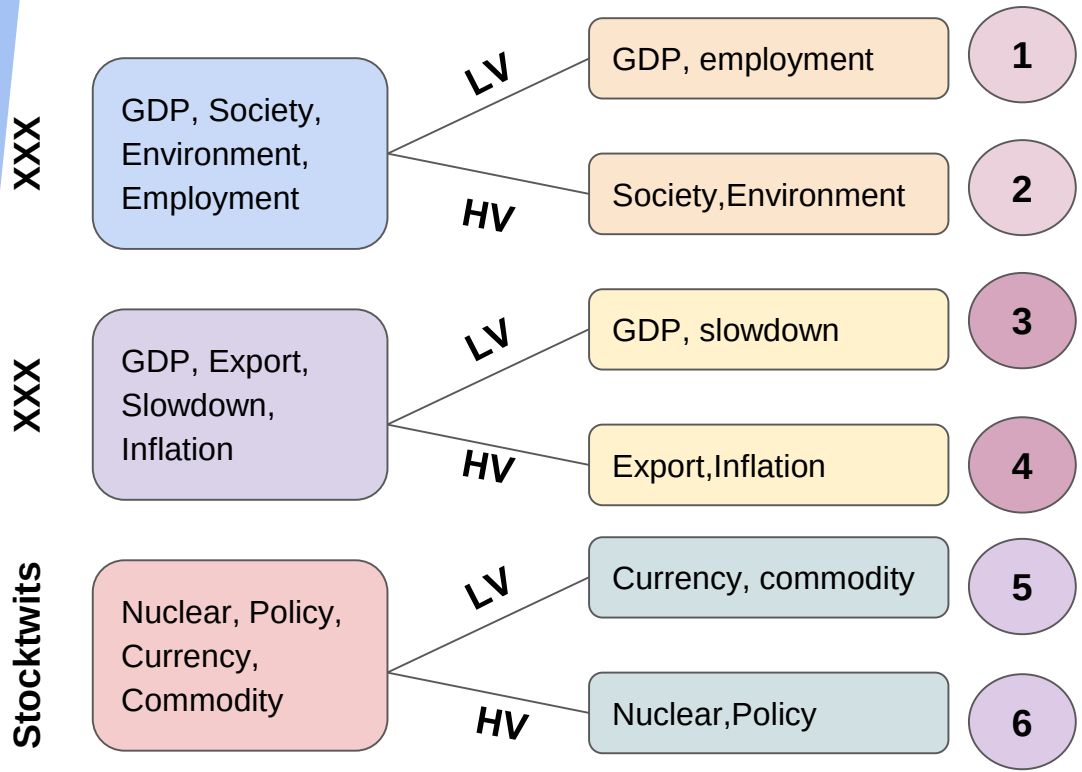
Columbia University

“Economy” - Associations

Columbia University



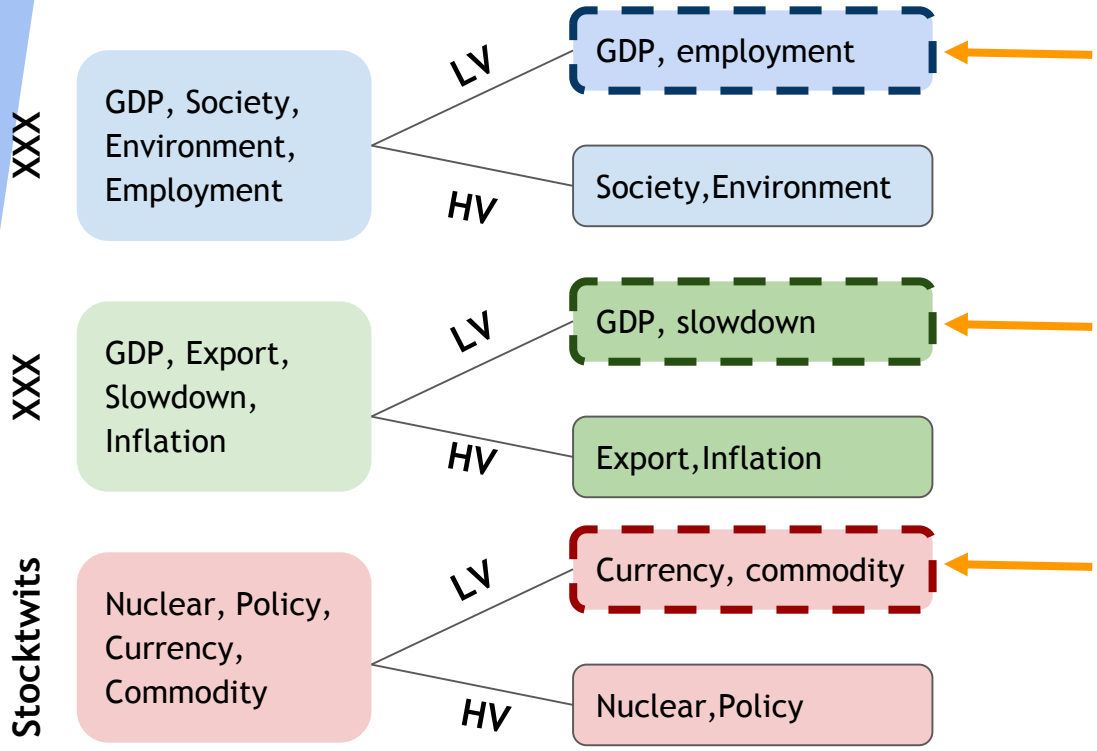
“Economy” - Volatility



Taking the union of words from **low volatility** groups of **Stocktwits** and **XXX** and removing the words from **low volatility** group of **XXX** we get:

Recession, Slowdown, Recovery, Growth

“Economy” - Volatility



Taking the union of words from **low volatility** groups of **Stocktwits** and **XXX** and removing the words from **low volatility** group of **XXX** we get:

Recession
Slowdown
Recovery
Growth

Examples of Key topics

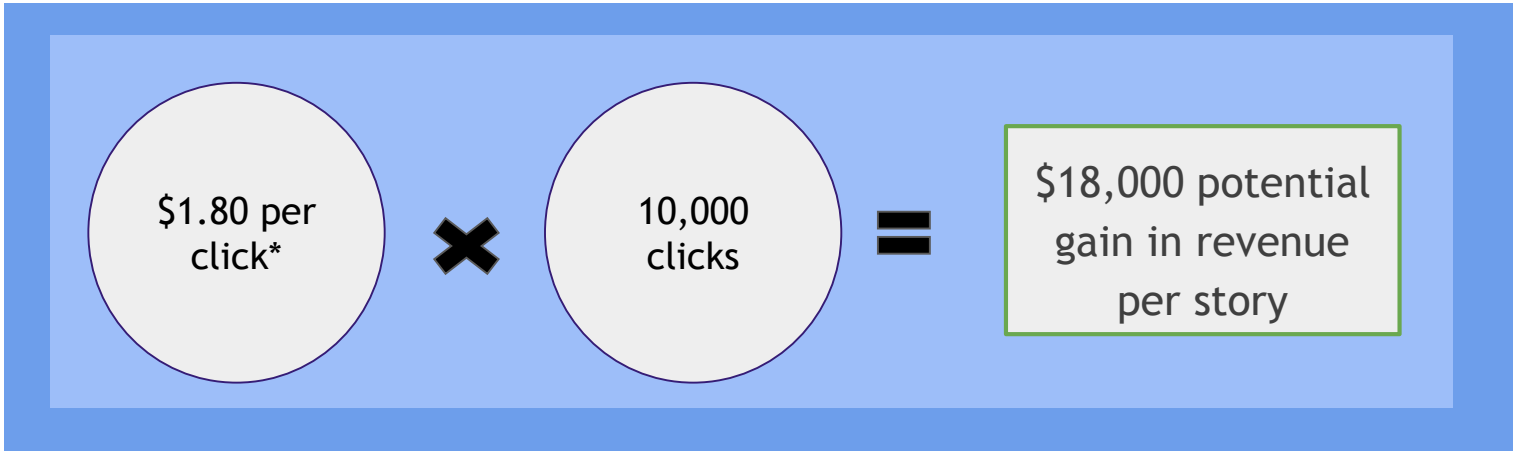
Topics	Competitors/Public are focusing on it
economy	['recovery', 'growth', 'consumption', 'slowdown', 'currency', 'commodity', 'recession']
loan	['lend', 'private', 'lender', 'asset', 'debt', 'credit', 'credit_card', 'subprime', 'borrower']
tax	['tax_reform', 'tax_code', 'repatriation', 'corporate', 'earner', 'bracket']
drug	['pipeline', 'product', 'approve', 'partner', 'partnership', 'indication', 'medication']

Topics	Competitors/Public are not focusing on it
digital	['mobile', 'advertising', 'marketing']
threat	['iran', 'nuclear', 'middle_east']

Potential Monetary Impact

COLUMBIA UNIVERSITY

- ‘Loan’ is talked about in association with ‘lend', 'private', 'lender', 'asset', 'debt', 'credit', 'credit_card', 'subprime', 'borrower'
  - Financial Services - ex. Debt refinancing services



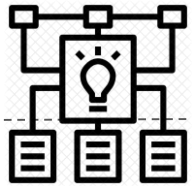
*Based on Forbes

Business Application

COLUMBIA UNIVERSITY

News Gathering

1. Stories from producers,reporters
2. Systems like Avid iNews



Our algorithm gathers input from last 24 hr news cycle

Rundown Meeting

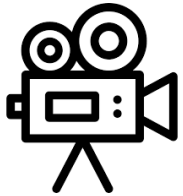
Ideas from producers + systems compared to results of our algorithm

Key outputs:

1. Producers identify viewer preferred topics
2. Higher utilization of resources as more content is generated

Production

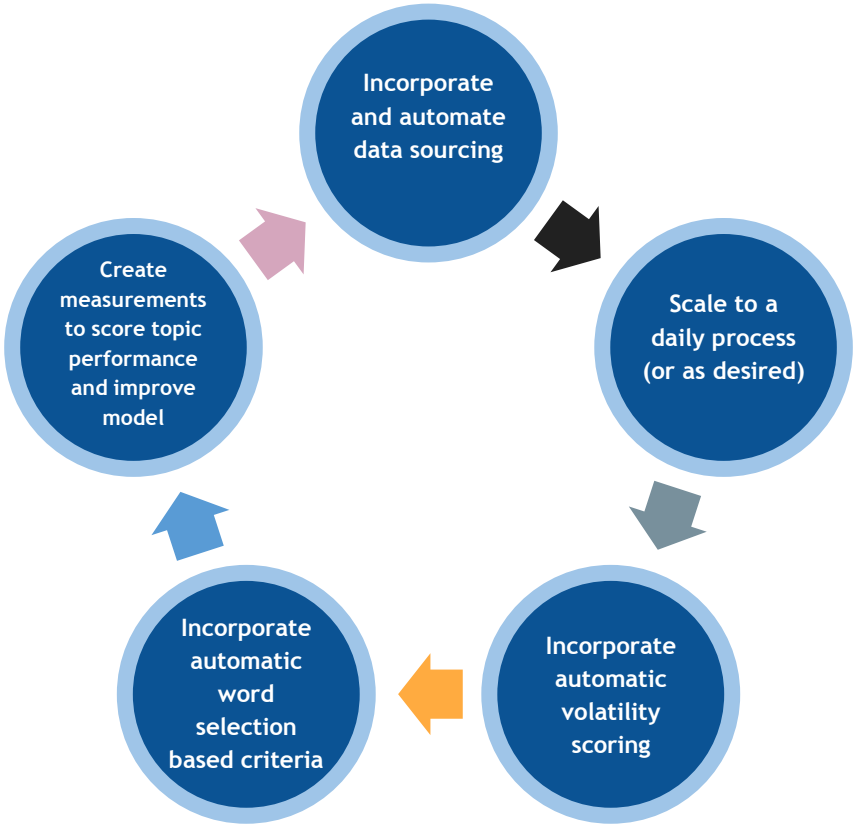
XXX able to produce larger volume of content, exploring wider aspects of key topics



Next steps:

1. Analyze and refine based on success criteria
2. Customize algorithm scale as desired (see nect slides)

Application - Next Steps(Algorithm)



Goals and Recommendations

COLUMBIA UNIVERSITY

Augment topic selection and decision-making process	▸ TS Algo will be utilized in the business day flow to aid producers in topic choice
Better utilization of resources by generating more content	▸ Volatility index can be used to identify topics that will potentially generate revenue.
Wider audience capture by providing better content	▸ High volatility index can identify topics to be considered for online platform content development to appeal to a wider audience.
Increased revenue from higher viewership	▸ Better word choice may lead to higher viewership, higher revenue per click.

COLUMBIA UNIVERSITY

Thank you!

Any Questions?