Refinitiv Lab – Data Science Challenge

Tan Tian Huat

Outline

- Installation
- Methods
  - Translation
  - Data Preprocessing
  - Stop Word Removal
  - Background Word Removal
  - Latent Dirichlet Analysis (LDA)
- Results Analysis

Installation

- Please refer to readme.md

Methods

- We assume trending as word which is popularly discussed in the corpus and also shows change of word usage over time. If a word shows stable word usage over time.
- Following is the big picture, details will be described later in subsequent slides.
  - We first translate all the different languages into English.
  - We perform analysis on the fields to get fields that contains signals rather than noise.
  - We perform various processing of the tokens and filtered tokens based on their usage
  - We make use of LDA to group trending words, so that it provides context to understand the trending words.

Translation

- There are total of 16 languages. Although we can handle them using a language dependent way, like for Chinese we can segment刘德华的歌-> [“刘德华”,“的”,“歌”]. Due to time limitation, all the languages are converted to Google Translation to English for further analysis.
- The google api provided in the submission file has throttle upon 1000 queries, since we are not subscribed users. To get around this problem, I output the text to HTML pages to use Google page translation for individual languages.

Preprocessing – analysis of fields

- We always omit **EVENT_TYPE** that is **DELETE** – there do not have useful contents, and **PRODUCTS** that is test, as they are test messages.
- The observation is that a story has its **UNIQUE_STORY_INDEX**, while it can has multiple rows in the table with the same **UNIQUE_STORY_INDEX**, with different **EVENT_TYPE** (e.g., HEADLINE, STORY_TAKE_OVERWRITE) etc.
- For **EVENT_TYPE** except DELETE, it always has **ALERT** and **STORY_TAKE_OVERWRITE**.
- For the same **UNIQUE_STORY_INDEX , STORY_TAKE_OVERWRITE** is normally a more complete writing then **HEADING** or **ALERT**, and **STORY_TAKE_APPEND** is normally a repeat of **STORY_TAKE_OVERWRITE** by moving the **TAKE_TEXT** into **ACCUMULATED_STORY_TEXT**
- Therefore in the preprocessing steps, we combine the text using the following rules:

I group by **UNIQUE_STORY_INDEX** (with **EVENT_TYPE** that is not **DELETE** and **PRODUCTS** that is not **TEST**). In each group, if **STORY_TAKE_APPEND** is available, we use the combination **HEADLINE_ALERT_TEXT** and **TAKE_TEXT**, otherwise we will use **ALERT** entry only **HEADLINE_ALERT_TEXT.**

Preprocessing

- The following precessing methods are make use of
  - Tokenization (Using TextBlob)
  - Normalization (lower case all words)
  - Lemmatization (Using TextBlob)

Stop Word Removal

- We further remove for following categories of words
  - Stop word list in nltk
  - Words that are too frequent (top 100 words)
  - Words that are least frequent (count <5)
  - Words with length <=2
  - Words with only numeric or special character
  - Words with digit and / symbol
  - Words that represents a date
  - ‘nil’
  - Words that contains ‘thompsonreuters.’ or ‘reuters’

Background words filtering

- We also filter words that has flat occurrence. This is done using Shannon wavelet describing by
  - J. Weng and B.-S. Lee, “Event detection in twitter,” in Proc. Int. Conf. Weblogs Soc. Media, 2011, pp. 401–408.
- This will help remove words like *early, last, much* etc.

Some statistics

- Total sentences after preprocessing: 2940
- Total tokens after preprocessing: 5900269
- Total unique tokens after preprocessing: 45936
- Total tokens after stop word filtering: 438497
- Total unique tokens after stop word filtering: 9753
- Total tokens after background word filtering: 254363
- Total unique tokens after background word filtering: 7753

Latent Dirichlet Analysis (LDA)

- LDA summarizes each document as a distribution of latent topics and each topic as a distribution of words. The reason for using LDA is to gives context for trending words, since many times a trending words is difficult to tell the whole story, while group of trending words can better conveys to us the theme of the story.

Results of LDA

- Topic #0: syria syrian saudi killed would force attack french sunni rebel iraq hezbollah prepared war source
- Topic #1: oklahoma tornado city storm outage killed weather moore area brkr update emergency flood houston pjm
- Topic #2: protest police erdogan protester turkey square party taksim demonstration ankara turkish tear spain city madrid
- Topic #3: cancer election say would drug house woman party study patient trial planned editing prison former
- Topic #4: india rating auction traded rupee jgb palm bbcl vessel long-term bill n.a soybean mizuho mutual
- Topic #5: breakingviews round commodity paris french story http energy beat wealth asian match third tennis emerging
- Topic #6: team win season league match cup game minute goal champion place coach club player third
- Topic #7: report please energy exchange commodity nikkei singapore metal australian tokyo future code http indonesia asian
- Topic #8: pmi manufacturing announced growth hsbc employment value purchasing quarter monetary fell forecast national would report
- Topic #9: interest share fell financial fund japan investor sale asset global exchange debt future foreign business

The results pretty much convey to us the topics.

For the first topic, it is related to war in Syria etc.

For the second topics, it is related to flood and storm in Oklahoma.

Improvement

- There are some repetitive information could be removed before the analysis, e.g,.

  LIVE PRICES & DATA
  World Stocks <0#.INDEX> Currency rates <EFX=><NFX=>
  Dow Jones/NASDAQ <.DJI><.IXIC> Nikkei <.N225>
  FTSE 100 <.FTSE> Debt <0#USBMK=><EURIBOR>

  HOW TO FIND INFORMATION YOU NEED
  |<REUTERS> | <NEWS> | <PHONE/HELP> |
  |<EQUITY> | <BONDS>| <MONEY> | <COMMODITY> | <ENERGY>

  .....
  Page Editor: Divya Sharma, Bangalore Newsroom, divya.sharma@thomsonreuters.com
  .....

- More analysis can be done to understand the topics better (e.g., sports, financial), and we can filtered the words based on topics.
- We can perform LDA with more number of rounds and topics to get better results