# Report

### The results involve 4 sections:

- 1. Most used words for each year-month segmentation
- 2. Words similarity from top voted titles per year
- 3. Title similarity from top/bottom voted records per year
- 4. Prediction on voting based on titles
- \*Please download Eluvio\_DS\_Challenge.csv and GoogleNews-vectors-negative300.bin in the main directory before running the code.

Eluvio\_DS\_Challenge.csv:

https://drive.google.com/file/d/15X00ZWBjla7qGOIW33j8865QdF89IyAk/view?usp=sharing\

GoogleNews-vectors-negative300.bin:

https://s3.amazonaws.com/dl4j-distribution/GoogleNews-vectors-negative300.bin.gz

# 1. Most used words for each year-month segmentation

#### Method:

It would be interesting to see the high frequency words for each year-month segmentation, those highly used words give us some brief ideas on the hot topics for certain months. To count the frequency of hot words in one particular month, I tokenized each topic, removed stop words and words that contain digits (such as 5k), and counted with 'Wordcloud' package. Top 100 words are selected for each month, words are capitalized for better visualization.

#### Code access:

To retrieve most popularly used words for each month in certain year, function 'Most\_used\_words' in WordingTrending.py file plots word boards for the need. Simply set 'True' in the corresponding operation request under 'if

\_\_name\_\_ == '\_\_main\_\_'' (line 254) in the same python file to save the plots under the directory 'Images/MostPopularWords'.

## Results:

12 months/plots for each year are obtained. From the results, we are clearly seeing that the word 'President' is highly used in each month of election year (2018,2012,2016) [Fig.1, Fig.2, Fig.3]. Other words like 'Russia', 'China', 'Attack' are often appeared in each plot regardless of the timing. Some hot words are very time-sensitive, for example, 'Bush' appeared in all months of 2008 and disappeared in the following years.



Fig.1



Fig.2

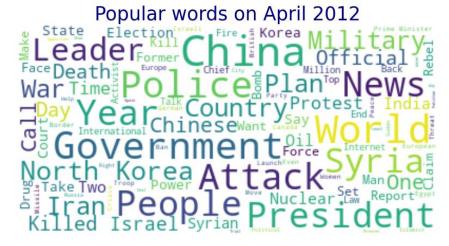


Fig.3

# 2. Words similarity from top voted titles per year

## Method:

To see if words in highly voted titles are clustered together, here we plotted the words used in top 5% up-votes titles for each year. Firstly, I selected titles whose upvotes stands in top 5% of the year, and tokenized words that are

contained in those titles. To obtain vectorized words representation, Google pre-trained word embedding (GoogleNews-vectors-negative300.bin) were used to generate word features. For better visualization, I reduced the word representation to 2 dimensions using TSNE package in Python.

#### Code access:

To retrieve top titles' words similarity for each year, function 'Google\_embedding\_analysis' in WordingTrending.py file plots for the need. Simply set 'True' in the corresponding operation request under 'if \_\_name\_\_ == '\_\_main\_\_'' (line 256) in the same python file to save the plots under the directory 'Images/TopVotesWordsSimilarity.

#### Results:

From the below result example, I observe that words that are used in highly voted titles are generally clustered together. [Fig4]

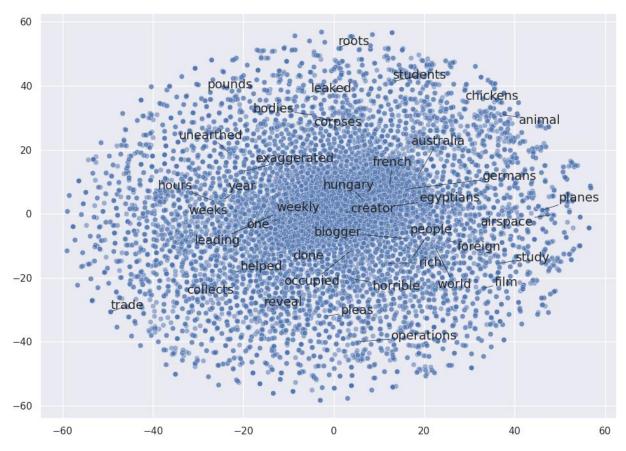


Fig.4

# 3. Titles similarity from top/down voted records per year

#### Method:

I'm very happy to see the result from section 2 that words in high voted titles are clustered together, this leads me to wonder if top voted titles are clustered together. Here I represented the whole title from the selected top 30/bottom 30 titles from each year, and visualized these titles in 2-D plots.

#### Code access:

To retrieve top/bottom titles similarity for each year, function 'google\_embedding\_titles' in WordingTrending.py file plots for the need. Simply set 'True' in the corresponding operation request under 'if \_\_name\_\_ == '\_\_main\_\_'' (line 258) in the same python file to save the plots under the directory 'Images/TopBottomVotesTitleSimilarity.

#### Results:

From the below result example, I observed that top voted titles are more likely to be at the upper part of the plot and the least voted titles tend to be at a lower position. [Fig5] However, this pattern is not very clear, and my suspicion is on the loss of information for the reduction of word feature dimensions.

	India sinks Somali pirate ship Georgian ambassador admits that they started it.
	London Police boycott tasers because of their potential to cause fear and damage public confidence
	Blue:Top 99% Up Votes ; Orange: Below 1% Up Votes
	I Kneed yet are Service Thinks so that the service of the service
	Today I end my support of Israel
	Kertan porn kids are people, judge says in Simpsons porn careful of the Country to Live 07. Oct 5/08: Food hearding begins, Currency rated just above Zimbabwe Cartoon porn kids are people, judge says in Simpsons porn care
	Local maffe, theCamorra declares war on Italy  Lockerbie bomb document bid fails  Lockerbie bomb document bid fails
	Canada: Wal-Mart tries to prevent employee unionization. Wal-Mart tries to go to Supreme Court. Wal-Mart loses.
	Local right war would cause world have: - NO DOUBT!
	in an A'ghan prison, the majority of female prisoners are serving 20-year sentences for being victims of rape
	BBC uncovers lost traq billions Lawyer for Murdered Anti-Kremlin Journal of Torgeted for Poisoning
	mentican Express gets 3.398/kinn, Capital/Oncepets3.6 Billion, Discover gets 1.28/kinn Letshavethenintestrateotratismoneyar2.993 million late tee.  Lounge Acts the Iraq activate approves the SOON by a wide margin, inumy what a U.S. president who scommitted to withdrawal can yield
	Italian mayor hans natherings of three or more people as soldiers his streets. Today Israel Lakes down an entire apt building of civilians to kill the family of one man. In Canada, we call that terrorism.
	The Saudi ask permission to build mosque in bloscope. The Riss and general ordingriphy in I have larbeit.  The Saudi ask permission to build mosque in bloscope. The Riss and general ordingriphy, an orthodox church in Saudi Arabia.
	Free him! Irag rally for Bush shoe attacker
Ame	rican intelligence confirms that the latest military actions in South Ossetia were started by Georgia and Russia's position in the conflict was correct
	Twelve Year Old Girl Shocks Fox News Says Georgia was the Aggressor, Not Russia
	RIP Brenden Foster, the 11-year-old whose last wish was to feed the homeless.  Gorbachev forms new Russian party opposing Putin  Kerya declares holiday for Obama  The Other Man On The Podium
	Karya declares holiday for Obama MaSID NABAWI NADIDAH NASID NABAWI NADIDAH Ridiculous: U.S. Olympic Cytilish had to applogize for weering smag masks
	Luv To Surf: Cyrthia Sommer is Set Free   No Longer Guilty of Killing Husband with Arsenic
	Lynchings in Congo as penis theft panic hits capital 6,000 Muslim clerics from around India approved a fatwa against terrorism.
	Luv To Surf: Eli Manning and Abby McGrew Married in Mexico
	Aprical terrorist writes poem about beheading a hostage, wins appeal in British court  France calls for a boycott of the O ympics ceremony over Tibet  Denmark considers outlawing male circumcision on minors.
	2
	Reddit upvoted the original story, why not this one? Muslims are those involved in acid attack against schoolgiris in adjantscho
	Local militias try to stop retreating Congolese soldiers and force them back to the front to fight Nkunda's rebels  MASJID QUBA
	Lowbirth rate? Just turn off the power for a day or two Loyalists Lead Vote in Belarus
	London gallery showing a pig painting suffers wave of threats and vandalism from Muslims
he had b	efriended brack kids who carry guns at school. He's dead now and Luc Pierre-Charles isn't getting the Bragging (lights? WTF?
	MARKETING EMAIL to people with DOMAIN  Luv To Surt: Earthquake Rocks Midwest   5.4 on the Richter Scale   Video
	Lying to police about secret lover okay in Italy
	Georgia downs two Russian warplanes as countries move to brink of war
	The Bush administration this month is quietly cutting off birth control supplies to some of the world's poorest women in Africa.
	Low impact, woodland homes - Not just for Hobbits any more!
	"Why won't America and Nato help us? If they won't help us now, why did we help them in Iraq?"
	Luv <u>to</u> Surft Paltent Uses Meat Cleaver to Hack NYC Pychologist Kathryn Faughey
	MAJOR European Union fuel crisis decision expected
	indical curduous ruinos, ruino, ruino

Fig.5

# 4. Prediction on voting based on titles

#### Method:

This section builds classification models to predict the votes based on titles. The voting is classified into 'high' and 'low' using the cutoff - the mean value of overall upvotes score in the dataset. Neural network model is applied using Keras package in Python with 2 benchmark models: logistic regression and Naïve Bayes. Two feature sets (word embedding) are generated for comparison,

one used Google pre-trained embedding and the other used the simple frequency matrix (counts of words in positive/negative class) obtained from training data. To avoid negative values in Naïve Bayes model, only frequency matrix feature set is used for the Naïve Bayes model. In summary, this section compared the following 5 models:

- 1. Logistic regression (frequency matrix as word embedding)
- 2. Logistic regression (Google word embedding)
- 3. Naïve Bayes (frequency matrix as word embedding)
- 4. Keras NN model (frequency matrix as word embedding)
- 5. Keras NN model (Google word embedding)

#### Code access:

To run the above model, function 'title\_vote\_prediction' in WordingTrending.py file does the tick. Simply set 'True' in the corresponding operation request under 'if \_\_name\_\_ == '\_\_main\_\_'' (line 261) in the same python file to save the result plots under the directory 'Images/ Prediction'.