Report

*Please download Eluvio_DS_Challenge and GoogleNews-vectors-negative300.bin in the ./**DS_challenge** directory before running the code.

Eluvio_DS_Challenge.csv:

https://drive.google.com/file/d/15X00ZWBjla7qGOIW33j8865QdF89lyAk/view?usp=sharing\

GoogleNews-vectors-negative300.bin:

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

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

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

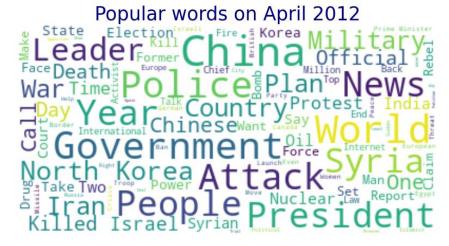


Fig.3

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]

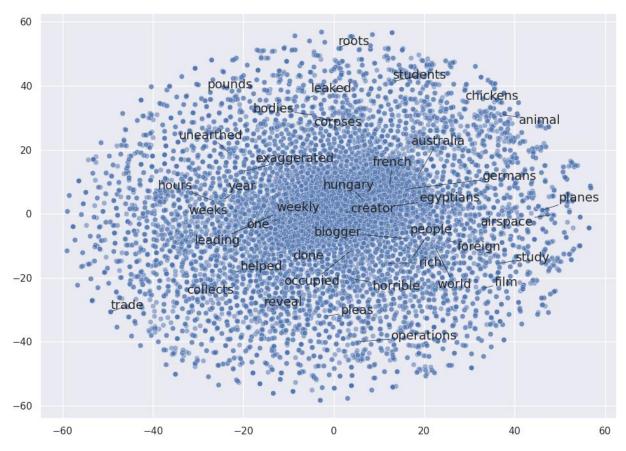


Fig.4

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 Know, we see This file stigging Vis. To Wanted To Helo Shine More Light On Our Australian Friends Potential Internet Filte
	Today I end my support of Israel
	Iceland. Sept 5/08: UN's Best Country to Live 07. Oct 5/08: Food hoarding begins, Currency rated just above Zimbabwe
	Cartoon porn kids are people, judge says in Simpsons porn case Local mafle, theComorna declares war on Raly Lower Ab exercises
	Lockerbie bomb document bid falls
	Canada: Wal-Mart tries to prevent employee unionization. Wal-Mart tries to go to Supreme Court. Wal-Mart loses.
	Local nuke war would cause world havoc - NO DOUBT! Lula suspends Brazil spy chiefs over phone taps
	In an Afghan prison, the majority of female prisoners are serving 20-year sentences for being victims of rape
	BBC uncovers lost trag billions Lawyer for Murdered Anti-Kremlin Journalist Targeted for Poisoring
	American Express gets 3.398/lifon, CapitalOnegets 3.6 Billion () Discover gets 1.28/lifon
	Italian mayor bans gatherings of three or more people as soldiers hit streets Today Israel takes down an entire apt building of civilians to kill the family of one man. In Canada, we call that terrorism. MIS burned abone of 87th building of mIS burned abone of 87th burn
	trains mayor cares garrenings or cinee or more people as socients in streets. MIS budged profess of Birtish civil servants during WW I The Saudi ask permission to build mosque in Moscope. The Rigs asin element elepipacity, an optifique church in Saudi Arabia.
	Free him! Iron rally for Bush shee attacker. Control and the state of
Ame	trican intelligence confirms that the latest military actions in South Osseba 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
	Kenya declares holiday for Obama MASJID NABAWI MADINAH The Other Man On The Podium
	Ridiculous: U.S. Olympic Cyclists had to apologize for wearing smog masks
	Luv To Surfi-Cyrthia Sommier is Set Free No Longer Culty of Killing Husband with Arsenic
	Lynchings in Congo as penis theft panic hits capital 6,000 Muslim cierics from around India approved a fatwa against terrorism.
	Luv To Surfi: Ell Manning and Abby McGrew Married in Mexico. Lurical terrorist: writes poem about beheading a hostage, wins appeal in British court Lurical terrorist writes poem about beheading a hostage, wins appeal in British court
	France calls for a boycott of the 0 ympics ceremony over Tibet Denmark considers outlawing male circumcision on minors.
	Liux Th. Suff. Mary Key Lotomona Causes a Non-Rill in Washington State Reddit upvoted the original story, why not this one? Mostlina ments those involved in acid attack equants chooligits in arghanistan
	MASIO PIEATH ERRETTS Local millitias try to stop retreating Congolese soldiers and force them back to the front to fight Munda's rebels MASID QUBA
	Lowbirth are Justices and locate memors, or the transport of the power for a day or two Lowbirth are Just turn off the power for a day or two Lowbirth are Just turn off the power for a day or two Lowbirth are Just turn off the power for a day or two
	London gallery showing a pig painting suffers wave of threats and vandalism from Muslims
id out he had l	perfriended black kids who carry guns at school. He's dead now and Luc Perre-Charles isn't getting the Bragging Nights? WTF?
	MARKETING EMAIL fo people with DOMAIN Luv To Surf: Earthquiske 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!
	Low impact, woodland nomes - Not, just not Hobbits any more! "Why won't America and Nato help us? If they won't helo us now, why did we help them in Iraq?"
	Luy 'S Surf; Paltent Uses Meat Cleaver to Hack NYC Pychologist Kathryn Faughey
	MAJOR European Union fuel crisis decision expected

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'.