Event Detection from the Reuters TRC2 Corpus

Tomáš Kala

November 16, 2017

1 Dataset

The dataset used is the Reuters TRC2 corpus [1]. It comprises 1,800,370 news stories covering the period from January 1, 2008 to February 28, 2009. Each document consists of a headline, publication date, and the news story itself. On average, the document contain 8 words per headline and 203 words per body.

2 Preprocessing

As the documents consist only of plain text, some form of preprocessing is necessary before proceeding to event detection.

We first perform lemmatization and parts of speech tagging using the SpaCy package [1].

Next, we train the popular Word2Vec model [2] on the lemmatized texts. The implementation used comes from the Gensim library [3].

For the training, only punctuation symbols and words marked as unknown part of speech are discarded. The words were embedded into a 100-dimensional vector space using the skip-gram architecture, and allowing 5 passes over the collection with a 5-word window.

Having done that, we proceeded to the event detection. Here, we used only words marked as Nouns, Verbs, Adjectives and Adverbs. We also filtered out any word containing special characters. This left us with a vocabulary of size 180,114 unique words.

3 Methods and parameters

3.1 Common settings

The Dominant Power Spectrum boundary was set to 0.15, which left us with 1692 words considered for event detection. The window length for computing the moving average (Section 5.3.1 in the bachelor thesis) was set to 7.

3.2 Original method

No extra parameters.

3.3 Embedded-greedy method

No extra parameters.

3.4 Cluster-based method

Compared to the bachelor thesis where we used the DBSCAN algorithm [4], its modification, HDBSCAN [5], proved to perform better on this dataset. After experimentation, parameters were set to $\epsilon = 1.0$, $min\_samples = 2$, $min\_cluster\_size = 4$.

4 Results

Results of the same evaluations as in the bachelor thesis are shown below. Compared to the Czech texts, more events with short period (7 days or lower) were detected. Such events are not usually as important as higher-periodic ones, as they usually concern sport matches, stock reports and so on. For this reason, all evaluations were done both on the detected events and on events with period higher than 7 days.

Also, similarly to the bachelor thesis, the cluster-based method shows tendency to generate one or two clusters filled by words that simply did not fit anywhere else. It is possible that this behavior is due to ill-set parameters, but even after playing around with those, this tendency remained. However, these clusters are easy to detect and removed, as they contain over 100 words, compared to real events consisting of at most a few tens of words. For this reason, these clusters

[1] http://trec.nist.gov/data/reuters/reuters.html

(numbers 21 and 30 in the plots) were removed before evaluation. Neither of these two events remains after removing those with period lower or equal to 7 days.

4.1 Detection statistics

The results are very similar to those from the bachelor thesis. The original method generates a large number of events consisting of very few keywords. The embedded-greedy method detects fewer events with more keywords. The cluster-based method detects the lowest number of events with two outliers mentioned above.

4.1.1 Number of events

	Original	Embedded-greedy	Cluster-based
Events detected	125	77	36

4.1.2 Keywords per event

	Original	Embedded-greedy	Cluster-based
Mean	2.064	12.442	15.056
Std	0.304	9.541	23.762
Min	2.000	3.000	4.000
25%	2.000	5.000	5.000
50%	2.000	9.000	7.500
75%	2.000	16.000	16.000
Max	4.000	43.000	136.000

4.2 Detection statistics – only events with period higher than 7 days

4.2.1 Number of events

	Original	Embedded-greedy	Cluster-based
Events detected	113	18	14

4.2.2 Keywords per event

	Original	Embedded-greedy	Cluster-based
Mean	2.071	9.667	10.143
Std	0.320	7.538	10.007
Min	2.000	3.000	4.000
25%	2.000	6.000	4.250
50%	2.000	7.500	6.500
75%	2.000	10.750	8.000
Max	4.000	36.000	40.000

4.3 Precision, recall, F-measure

4.3.1 As detected

Overall, the cluster-based method reached the highest results. The high recall of the original method will be explained in Section 4.5. The original method reached high redundancy, so that a correctly detected event would be counted multiple times.

Method	Precision	Recall	F-measure
Original	13.600%	26.316%	0.179
Embedded-greedy	7.792%	18.421%	0.110
Cluster-based	20.588%	26.316%	0.231

4.3.2 Only events with period higher than 7 days

Compared to the previous table, the precision of all three methods is higher, while the recall remains the same. This confirms the intuition that there simply are not that many events of interest with a period as small as 7 days.

Method	Precision	Recall	F-measure
Original	14.159%	26.316%	0.184
Embedded-greedy	11.111%	18.421%	0.139
Cluster-based	35.714%	26.316%	0.303

4.4 Noisiness

4.4.1 As detected

The relatively high noisiness of the cluster-based method (where noisiness is expected to be low) is explained by the last 5 events detected. They all concern the same football matches, consist of words in different languages. They all have a low period though, and are filtered out in Subsection 4.4.2.

Method	Noisiness
Original	80.800%
Embedded-greedy	84.416
Cluster-based	38.235

4.4.2 Only events with period higher than 7 days

Method	Noisiness
Original	80.531%
Embedded-greedy	88.889
Cluster-based	21.429

4.5 Redundancy

Redundancy was evaluated both on all events, and on those marked as "not noisy".

After removing noisy events, the embedded-greedy method seemingly performed very well. However, when we look back to the noisiness evaluation, we see that almost *all* events were noisy. Therefore, not many events are left to evaluate redundancy after removing these.

4.5.1 As detected

Method	Redundancy (all events)	Redundancy (no noise)
Original	65.600%	33.333%
Embedded-greedy	57.143%	8.333%
Cluster-based	14.706%	19.048%

4.5.2 Only events with period higher than 7 days

Method	Redundancy (all events)	Redundancy (no noise)
Original	67.257%	36.364%
Embedded-greedy	22.222%	0.000%
Cluster-based	0.000%	0.000%

4.6 Purity

The cluster labels used for this evaluation can be found in Appendix A.

In this case, only events with period higher than 7 days were evaluated. The reason is that such short periods essentially cover the entire document stream with short bursts, and for each of those, the documents would need to be retrieved. As the Word Mover's Distance [6] is by itself a computationally expensive operation, this would take an unbearable amount of time.

Furthermore, as we have seen when comparing results before and after removing these periods, such short-periodic events do not really contribute anything meaningful.

Method	Purity
Original	33.62%
Embedded-greedy	27.51%
Cluster-based	60.59%

4.7 Computation time

The Word2Vec model can be pretrained and stored for future use. The tightest bottleneck of our two methods is therefore the Document retrieval phase.

Unit	Original	Embedded-greedy	Cluster-based
Word2Vec training	N/A	2h 10min	
Bag of words	⟵	15min	⟶
Word trajectories	⟵	7s	⟶
Event detection	5min 56s	5min 49s	2min35s
Document retrieval	1min 44s	4h 41min	4h 16min
Total	22min 47s	7h 12min	6h 44min

A Cluster labels for Purity evaluation

We have selected 60 important words from 1000 most common words among all headlines. Again, only Nouns, Verbs, Adjectives and Adverbs were considered.

The words are: *price, u.s., us, stock, india, bank, target, stocks, oil, company, financial, forecast, europe, china, euro, energy, asia, dollar, soccer, economy, nasdaq, japan, research, industries, markets, politics, eu, finance, crisis, russia, canada, morgan, money, trading, olympics, france, banking, kill, mexico, brazil, mideast, obama, school, health-care, housing, iraq, tennis, minister, iran, israel, attack, weather, president, gaza, police, protest, rebel, troop, death, afghanistan.*

B Reference events

The following table contains confirmed events that happened between January 1, 2008 and February 28, 2009. The data was taken from *https://www.onthisday.com/*.

#	Date	Headline
1	Jan 1	Malta and Cyprus officially adopt the Euro currency and become the fourteenth and fifteenth Eurozone countries.
2	Jan 14	MESSENGER spacecraft performs a Mercury flyby
3	Jan 21	Black Monday in worldwide stock markets. FTSE 100 had its biggest ever one-day points fall, European stocks closed with their worst result since 9/11, and Asian stocks drop as much as 15%.
4	Feb 5	A major tornado outbreak across the Southern United States leaves at least 58 dead, the most since the May 31, 1985 outbreak that killed 88.
5	Feb 14	Northern Illinois University shooting: a gunman opened fire in a lecture hall of the DeKalb County, Illinois university resulting in 24 casualties; 6 fatalities (including gunman) and 18 injured.
6	Feb 17	Kosovo declares independence from Serbia.
7	Feb 23	B-2 Spirit of the USAF crashes at Guam. Crew survives but aircraft written off, the most expensive air crash in human history (aircraft alone cost $1.2Bn). B-2 had a perfect safety record before the crash; not one B-2 ever crashed.
8	Feb 28	Former Prime Minister of Thailand Thaksin Shinawatra is arrested on corruption charges upon returning to Thailand after months of exile.
9	Mar 2	Riots in Yerevan, Armenia concerning the Armenian presidential election, 2008 come to a fatal end, with police forces clashing with civilians in their peaceful protest, resulting in 8 deaths.
10	Mar 24	Bhutan officially becomes a democracy, with its first ever general election.
11	May 2	Cyclone Nargis makes landfall in Myanmar killing over 130,000 people and leaving millions of people homeless
12	May 6	Chaiten Volcano erupts in Chile, forcing the evacuation of more than 4,500 people.
13	May 7	Dmitry Medvedev is sworn in as the 3rd President of the Russian Federation
14	May 12	Wenchuan earthquake, measuring 7.8 in magnitude occurs in Sichuan, China, killing over 87,000, injuring 374,643 and leaving homeless between 4.8 million and 11 million people
15	Jun 8	The Akihabara massacre took place on the Sunday-pedestrian-zoned Chuodori street. A man killed seven in an attack on a crowd using a truck and a dagger.
16	Jun 10	The Gora Prai airstrike by the United States reportedly kills 11 Pakistani paramilitary troops.
17	Jun 25	Atlantis Plastics shooting, An employee shot and killed five people after an argument, which ended in the gunman's suicide in Henderson, Kentucky.
18	Aug 8	Georgian invasion into South Ossetia. Beginning of five-day war between Georgia and Russia.
19	Aug 8	29th Olympic Games opens at Beijing, China
20	Aug 10	90th PGA Championship: Pádraig Harrington shoots a 277 at Oakland Hills Country Club
21	Aug 20	Spanair Flight 5022, from Madrid to Gran Canaria, skids off the runway and crashes at Barajas Airport. 146 people are killed in the crash, 8 more die afterwards. Only 18 people survive.
22	Sep 10	The Large Hadron Collider at CERN, described as the biggest scientific experiment in the history of mankind is powered up in Geneva, Switzerland
23	Sep 13	Hurricane Ike makes landfall on the Texas Gulf Coast of the United States, causing heavy damage to Galveston Island, Houston and surrounding areas.
24	Sep 21	Goldman Sachs and Morgan Stanley, the two last remaining independent investment banks on Wall Street, become bank holding companies as a result of the subprime mortgage crisis.
25	Oct 6	MESSENGER spacecraft performs a second Mercury flyby
26	Nov 4	Barack Obama becomes the first African-American to be elected President of the United States
27	Nov 25	A car bomb in St. Petersburg, Russia, kills three people and injures one
28	Nov 26	Terrorist attacks in Mumbai, India: Ten coordinated attacks by Pakistan-based terrorists kill 164 and injure more than 250 people in Mumbai, India.
29	Dec 1	The US economy has been in recession since December 2007, the National Bureau of Economic Research announces today
30	Jan 1	Slovakia officially adopts the Euro currency and becomes the sixteenth Eurozone country.
31	Jan 1	61 die in nightclub fire in Bangkok, Thailand.
32	Jan 3	Israeli ground forces invade Gaza.
33	Jan 8	A 6.2 magnitude earthquake hit Costa Rica's region of Volcan Poás, with an epicenter near Cinchona. It was caused by Varablanca-Angel fault.
34	Jan 15	Chesley Sullenberger lands US Airways Flight 1549 on the Hudson River shortly after takeoff from LaGuardia Airport in NYC. All passengers and crew members survive in what becomes known as the "Miracle on the Hudson"
35	Jan 20	Barack Obama, inaugurated as the 44th President of the United States of America, becomes the United States' first African-American president
36	Jan 31	In Kenya, at least 113 people are killed and over 200 injured following an oil spillage ignition in Molo, days after a massive fire at a Nakumatt supermarket in Nairobi killed at least 25 people.
37	Feb 7	Bushfires in Victoria left 173 dead in the worst natural disaster in Australia's history.
38	Feb 25	BDR massacre in Pilkhana, Dhaka, Bangladesh. 74 People are being killed, including more than 50 Army officials, by Bangladeshi Boarder Guards inside its headquarter.

References

[1] Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[3] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[5] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 160–172. Springer, 2013.

[6] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966, 2015.