What do a Million News Articles Look like?

David Corney, Dyaa Albakour,
Miguel Martinez and Samir Moussa
Signal Media
16-24 Underwood Street, London N1 7JQ
{first.last}@signalmedia.co

Abstract

We present a detailed description and analysis of the Signal Media One-Million Articles dataset. We have released this dataset to facilitate research on emerging news-related information retrieval (IR) challenges. In particular, we have observed over the past decade emerging novel paradigms for publishing and consuming news, where users can get updated on the go with news from multiple sources, and at the same time, news providers are increasingly using social media and citizen journalism as powerful news sources. As a result, a number of news-related IR tasks have emerged and attracted attention in industry and academia. These include news verification, temporal summarization of multiple news sources, and news recommendation among others. A number of news datasets were created and shared for news IR in the past. However, such datasets are often drawn from a single outlet, and heavily preprocessed and cleaned. Also, they have become outdated and are not suitable any more for the emerging news IR challenges described above. Our dataset aims to address this because it is a recent collection from a wide range of sources reflecting many real-world issues in news collection and analysis.

We present insights obtained from an analysis of certain characteristics of the dataset, such as article lengths; similarity between articles; and the temporal characteristics of news publishing. We also discuss the opportunities and the limitations of our dataset.

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

1 Introduction

We present an analysis of The Signal Media One-Million News Articles Dataset[1]. We have created and shared this collection to stimulate research into new and improved methods for large-scale text analysis related to a wide range of applications. The articles were collected from a variety of news sources in September 2015. The sources include major national and international outlets, such as Reuters, the BBC and the New York Times, along with many sources that have fewer readers and less impact, including news magazines, blogs, local outlets and specialist publications. The collection is shared under a Creative Commons licence[2] while the copyright of the articles remains with the original publishers.

There are several existing collections of news-related texts that have been widely used by the information retrieval (IR) and natural language processing (NLP) communities, such as the Twenty Newsgroups collection [twe99]; the Reuters-21578 test collection[Lew96]; several Reuters Corpora, such as RCV1 [LYRL04], RCV2 and TRC2; and more recently Yahoo's user-news feed interaction data set [Yah16]. Common Crawl[3] provides a collection of 1.8 billion pages, but this represents a snapshot sample of the entire web, rather than a focussed news collection.

While such datasets continue to be useful for evaluating and guiding research, most are limited to a single source or a website. Furthermore, such datasets

[1] Available for research purposes from `http://research.signalmedia.co/newsir16/signal-dataset.html`

[2] Attribution, non-commercial `https://creativecommons.org/licenses/by-nc/3.0/`

[3] `http://commoncrawl.org/`

are highly curated and refined, which may result in an over-estimation of performance: if an algorithm performs well on such a clean set, will it still perform well when presented with more noisy data, such as content obtained from web-scraping or other less-controlled sources? In the fast changing era with news publishing and consumption, such datasets have become outdated and are not suitable for emerging IR tasks on news. Our dataset addresses this by providing a recent large sample of news articles from multiple sources over a one month period.

Like our collection, the British National Corpus [Bur07] is monolingual (English-only), synchronic (sampled from a single time period), general (not limited to any genre or topic), and sampled from a wide range of sources. Our dataset shares many of these features, except that while the BNC is definitively from a single nation, our collection spans the globe. While focussed on news, our collection also contains imaginative works of blogs and transcriptions of broadcast speech.

The articles of the dataset were originally collected for Signal Media by Moreover Technologies from 1^{st}–30^{th} September 2015. This included repeated sampling from over 93,000 different news sources ranging from large-scale mainstream media outlets to single-author blogs. Recent decades have seen a blurring of the distinction between mainstream and citizen journalism[Gla08], hence our inclusion of multiple sources in the collection. As well as being an attractively round number, one million articles is a large enough collection to develop and evaluate a wide range of tools and models, while still being manageable enough to not require specialist or sophisticated infrastructure.

Due to the scale of the collection, and the importance of speed, the collection process is largely automated. This, along with the variation in quality of source websites, inevitably leads to imperfections in the collection. For example, the goal was to collect English-language content only, but manual analysis shows that a small proportion of articles are in other languages or a mixture of languages. Similarly, some articles contain fragments of HTML, PHP or JavaScript code, due to problems with encoding, rendering or scraping, and some are duplicated (see Section 2.2). We consider such issues as the inevitable presence of noise in any real-world collection. One of our aims in sharing this dataset is to encourage the development of tools and methods that are robust enough to cope with data of this nature.

In the remainder of the paper, we provide insights from an analysis conducted on a number of the dataset characteristics (Section 2). We share some tools for accessing the data (Section 3), and discuss the opportunities and the limitations of the dataset in Section 4.

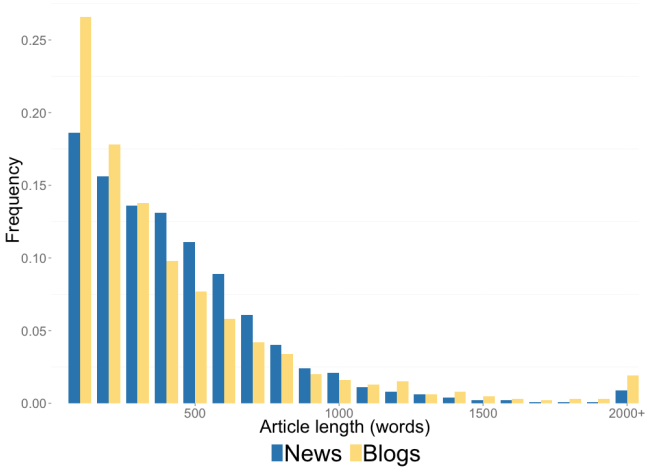

Figure 1: Distribution of article length for each media type (blogs and news)

2 Characteristics of dataset

2.1 Article length

Some news articles are little more than a single statement, especially if a news story is still breaking. Other articles may contain a lot of details and background information or discussion. To investigate this, we use a simple tokenizer[4] to count the number of words in each article. Figure 1 shows the distribution of article lengths, comparing the length of 'news' articles to the lengths of 'blog' articles. This shows that typical articles are a few hundred words long, but with a long tail reaching up past 2000 words. Note also that articles between 400 and 1000 words long are more likely to be news articles than blogs, while very short articles (200 words or less) and long articles (1200 words or more) are more likely to be blogs than news.

The tokenization also allows us to calculate that there are 407,754,159 words in the dataset; that there are 2,003,254 *distinct* words in the dataset; and that the average number of words per article is 407.75.

Which articles are unusually long? And which articles are unusually short? Just 144 articles in the collection have more than 10,000 words. The longest article has 12,450 words and is a transcript of a US college football match. Other long articles include an installment of serialized novel; an updated about a fantasy football competition; detailed personal memoirs; and a detailed list of fishing reports from Florida. While these may never attain very large readerships, they

[4]We used the Standard Tokenizer in ElasticSearch v1.7, which splits on white space and punctuation symbols, while allowing for abbreviations.

Title	First Week of ICE November 20th Options Trading
Content	Investors in Intercontinental Exchange Inc. (ICE) saw new options become available this week, for the November 20th expiration.
Media-type	News
Source	Town Hall
Published	2015-09-22T16:31:56Z

Figure 2: The shortest article in the collection, with 18 words of content.

will no doubt be of interest to certain audiences. At the other extreme, the shortest article in collection is nothing more than a brief announcement about options trading with no background discussion or detail. The article is shown in its entirety in Figure 2, including available metadata.

2.2 Duplicated articles

Identical, or near-identical articles may appear in the collection for several reasons, including:

Syndication One publisher may publish the same article through several outlets, such as regional newspapers.

Updates One source may publish multiple versions of the same article over time, especially in the case of updating breaking news stories.

Aggregation Some news aggregation sites (such as `wn.com` and `www.newslookup.com/`) display copies of articles originally published elsewhere. These articles may have already been collected from the primary sources.

Access issues Some sites give the same content when an automated tool attempts to access multiple articles on the site (e.g. a copyright or login notice), with the full text being behind a firewall.

To investigate this, we measured the cosine similarity between pairs of articles. This compares the frequency of terms found in each article, ignoring the word order. A score approaching one means that the same words appear at the same frequencies; a score close to zero means that entirely different words appear. If two articles differ by only a few words, we still wish to treat them as near-duplicates. In this way, we can ignore differences caused by different page layouts, changes in the byline or other minor edits.

In Figure 3, we show the probability distribution of cosine similarity scores between pairs of articles. To generate this, we randomly sampled 250 articles and measured the cosine similarity between each of these

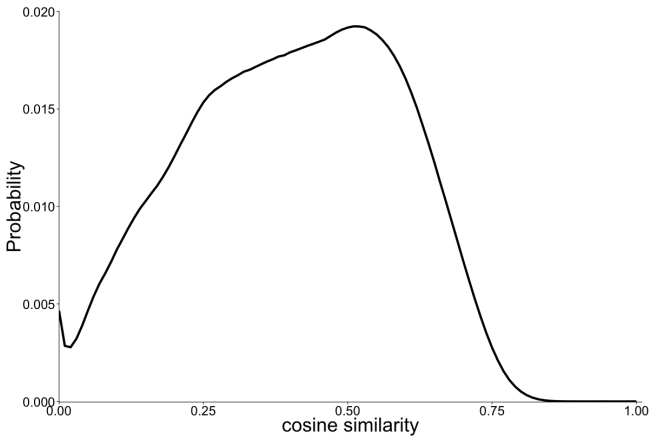

Figure 3: Distribution of pairwise cosine similarities. This shows the probability of generating different scores when comparing 250 × 1m articles.

and each of the million articles. We removed stopwords and calculated the term frequencies of the remaining words (though similar results were obtained without removing stopwords). The cosine similarity scores were grouped into 100 equal-sized bins to generate the graph. Note that we are sampling from the space of similarity scores and not the space of articles: full pairwise analysis of a smaller corpus would lead to a very different, sparser distribution.

The broad peak in central portion shows that when choosing two articles at random, their cosine similarity is likely to be around 0.5. 90% of article pairs have a cosine similarity between 0.13 and 0.70. This shows a substantial degree of overlap between articles in terms of word frequencies, despite the wide range of sources and topics covered. The slight up-tick at the extreme left shows that around 3% of articles share no terms (except stopwords) with any other article. These are typically very short articles containing rare terms (such as rare proper nouns). This includes some articles that have been corrupted during collection, and contain only fragments of code. At the other extreme, 0.0561% of pairs have a similarity score of greater than 0.95. This suggests that articles have on average around 2.2 duplicates or near-duplicates. This result is consistent with the duplicate recognition component of Signal Media's news monitoring platform[5].

2.3 Typical articles

Beyond considering duplicates, we investigated which articles are 'typical' and 'atypical' with regard to the rest of the collection. To do this, we remove stopwords and generate term frequency vectors for each article, as before. We then combined these into a single term

[5] `signalmedia.co`

Term	TF	DF
2015	930524	352629
time	642021	331982
people	531947	235954
1	518949	208164
september	471723	204583
market	423744	129407
company	416838	161796
2	414198	185599
day	387289	206976
world	365728	188220
news	364983	216282
information	351809	177431
business	327894	144896
3	313890	154149
10	312377	176963
home	293098	156997
million	278202	120327
including	275302	183904
week	271468	153210
team	263352	128957

Figure 4: The 20 most common terms, excluding stopwords, showing the total number of term occurrences (TF) and the document frequency (DF).

frequency vector representing the centroid of the entire collection. We then measured the cosine similarity between each article and this centroid. Articles with a high score can be seen as typical of the collection, in that the distribution of terms is similar. We plan to repeat this analysis using TF·IDF, but we expect to find a similar pattern of results overall.

Figure 4 lists the twenty most common terms in the collection (excluding stopwords). These include the month and year of the collection, along with words such as 'news', 'market', 'business' and company, indicating the typical focus of stories. .

One of the most typical stories by this definition, with a centroid cosine similarity of 0.9620, is about a new portable charge storage device[6]. It contains words such as 'power', 'time', 'heat' and other terms commonly found in news articles as shown in this extract: "*Harnessing the thermal energy from the included heating pot, it generates 5-watt power that can charge a device via the attached USB cable. You provide the heat source and liquid, and the PowerPot charges any compatible electronic with a USB connection. Charging time is actually comparable to a standard outlet, at about 1 to 2 hours to fully charge a phone.*"

One of the least typical stories is an article about

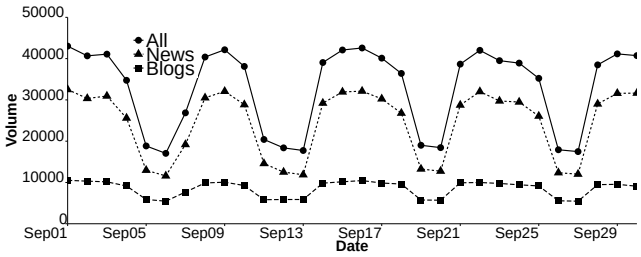

Figure 5: Daily volume over time for the different media types (Blogs and News).

tomato soup, in a small-town American newspaper[7] : "*That slight chill in the evening air? That's fall. And what better way to chase it than with a bowl of steamy, homemade soup. Chef Shereen Pavlides' recipe for Tomato Basil Soup is just what the season ordered, made with chicken stock and hearty San Marzano tomatoes.*" This has a cosine similarity with the centroid vector of just 0.00160.

While this centroid approach gives some indication of typical and atypical articles, it ignores the fact that the collection is far from homogeneous. A more sophisticated analysis could define a number of representative centroids, each representing a typical article belonging to a particular category. Such a segmentation of the collection could be achieved using any of a wide variety of document clustering algorithms.

2.4 Volume over time

In Figure 5, we plot the daily volume of articles published across the full duration of the dataset (September 2015). We can see the expected weekly pattern with more activity (published articles) during weekdays and less activity during the weekends. This is true for both media types (blogs and news). However, a couple of exceptions can be observed in these weekly cycles. First, there are fewer articles published on Monday, September 7^{th} compared to other Mondays and other weekdays in the month. This was a public holiday in the United States (Labor Day), which resulted in less media activity. Note that the dataset is sampled only from English-language sources and therefore the majority of articles originate from English speaking countries (with the USA the largest of those). The second exception is Friday, Sept. 11^{th} where the volume has dropped due to a downtime in the collection process that occurred on that day.

We now consider the hourly distribution of article publication in Figure 6. We calculate the average volume of articles in each hour of the day using the GMT timezone over the 30-day period. We observe a sharp

[6]Article ID 05bab596-55a3-40aa-adde-1a2cb48ebc41

[7]Article ID 497a3712-dd6d-4b0a-9365-5a7ade79d905

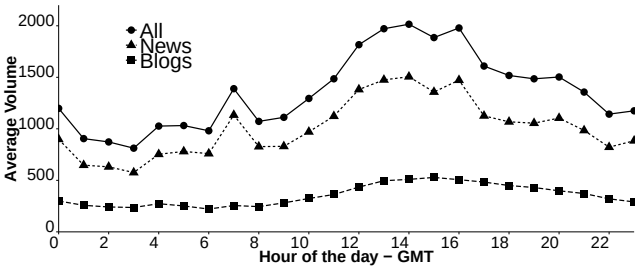

Figure 6: Average hourly volume for the different media types (Blogs and News).

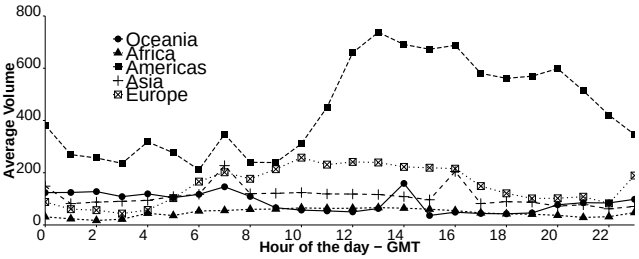

Figure 7: Average hourly volume for News sources across different geographical regions.

rise in the "News" volume at 7:00 GMT. The volume of "News" increases throughout the day and it reaches a peak between 12:00 GMT and 16:00 GMT. Afterwards the volume starts to decrease through the late hours of the day. To explain this behaviour, we further break down the "News" sources based on the continent of their sources' origin in Figure 7. We can observe that the peak at 7:00 GMT is mainly due to increased activity in Europe with the early hours in the working day there. Hours later, when it is morning in America (12:00 to 16:00 GMT), we observe an increase of activity from American sources which in fact dominate the stream.

For "Blogs" (Figure 6), we observe a different picture where the volume is more stable throughout the day with slight increases during the morning and afternoon hours.

3 Open Source Tools

We have created an open source repository on GitHub [8] to host useful tools and programming scripts for processing the dataset. For example, we have added scripts to index the data with ElasticSearch and to convert it into the TREC format for easier compatibility to some other IR tools. We will continue to maintain and promote this repository encouraging the community to provide similar tools for processing the dataset.

[8] `https://github.com/SignalMedia/Signal-1M-Tools`

4 Discussion

The Signal Media One-Million News Articles Dataset provides a distinctive research asset. The articles come from multiple sources and reflect many of the realities of practical large-scale text analytics. However, there are inevitably limitations to the data, some of which we now consider.

Language The dataset is almost monolingual, being dominated by English-language articles and with the remaining articles mostly being a mixture of English and non-English text.

Date range The articles were all collected during September 2015; a small number were published in the preceding days and weeks but only detected and downloaded in that month. By limiting the sample to a single calendar month, we achieved a relatively high density of related articles, such as multiple articles written about the same events. One month is also long enough for some of the news stories to change over time, making the dataset suitable for topic detection and tracking studies. Of course, by restricting the collection to this period makes it less useful for longer-term topic tracking.

Missing links and multimedia resource The dataset is text-only, and does not contain links to the original articles. This is partly due to issues around image licensing, but also to avoid problems with link rot: the ever-changing nature of the internet means that any published URLs would not reliably point to the same text as in the dataset. A dataset containing archived copies of multimedia content related to news would be very useful, but is beyond the scope of this work.

Labelling Many tasks require labelled documents, such as to assign documents to topic categories or to disambiguate entities. Although the collection currently has no such labels, our goal is to encourage the wider community of IR researchers to use this collection for a variety of tasks and share label sets for parts (or all) of the articles.

In conclusion, we believe that this dataset will be useful for developing, evaluating and comparing a wide range of news-related information retrieval tools and algorithms. The analysis provided here provides an initial description of the collection and forms the basis for future work.

References

[Bur07] Lou Burnard. Reference guide for the British National Corpus (XML edition). http://www.natcorp.ox.ac.uk/XMLedition/URG, February 2007.

[Gla08] Mark Glaser. Distinction between bloggers, journalists blurring more than ever. http://mediashift.org/2008/02/distinction-between-bloggers-journalists-blurring-more-than-ever059, 2008.

[Lew96] David Lewis. Reuters-21578 text categorization test collection. http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html, 1996.

[LYRL04] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004.

[twe99] 20 Newsgroups dataset. http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html, 1999.

[Yah16] Yahoo News Feed dataset. http://webscope.sandbox.yahoo.com/catalog.php?datatype=r&did=75, 2016.