Statistics of online User-generated short Documents

Giacomo Inches, Mark J. Carman, and Fabio Crestani

Faculty of Informatics, University of Lugano, Lugano, Switzerland
{giacomo.inches, mark.carman, fabio.crestani}@usi.ch

Abstract. User-generated short documents assume an important role in online communication due to the established utilization of social networks and real-time text messaging on the Internet. In this paper we compare the statistics of different online user-generated datasets and traditional TREC collections, investigating their similarities and differences. Our results support the applicability of traditional techniques also to user-generated short documents albeit with proper preprocessing.

1 Introduction and motivations

User-generated short documents are those produced online by visitors of blog or social networking websites, as well as by users of chat or instant messaging programs. The increasing popularity of these online services (Twitter, Facebook, IRC, MySpace) makes such generated content of great interest. From a commercial point of view, the analysis of these documents can highlight useful trends to focus online advertisement while, from a policing perspective, it may allow us to detect misbehavior or harassment.

While short documents and user-generated documents have been treated in recent works with different purposes (author identification, language analysis, gender prediction, documents clustering or law enforcement) to the best of our knowledge this is the first work which aims at understanding the differences between these and more traditional online collections, like the TREC “Ad-hoc” datasets.

In fact, it is important to asses the nature of similarity between user-generated short documents and more consolidated collections, to be able to infer the applicability of standard measures (of distance or similarity: BM25, Kullback-Leibler divergence, cosine with TFIDF weighting, etc.) and techniques (Probabilistic, Language or Topic Models) or to develop new ones which fit better the new datasets. For this purpose, we present the first results of our ongoing work, where we studied selected properties of 4 user-generated short document datasets and 3 more traditional ones, taken from the TREC Ad-hoc collections.

2 Datasets

As representative of user-generated short documents, we used the training dataset[1] presented at the Workshop for Content Analysis in Web 2.0 [2]. This consists of 5 distinct collections of documents crawled from 5 different online sources: *Ciao* (a movie rating service), *Kongregate* (Internet Relay Chat of online gamers), *Twitter* (short messages), *Myspace* (forum discussions) and *Slashdot* (comments on news-posts). Since we were more interested in messages exchange between users, we did not consider the *Ciao* collection and left it to future study.

We compared these datasets with a subset of the standard TREC Ad-hoc collections[2], choosing 3 of the most representative ones: *Associated Press* (AP, all years), *Financial Times Limited* (FT, all years), *Wall Street Journal* (WSJ, all years). These collections contain news article (*AP* and *WSJ*: general news, *FT*: markets and finance) published in the corresponding newspapers.

We discuss in next section the properties of these collections, which are presented in Table 1. Since the statistics of AP, WSJ and FT are similar to one anothers, we report only the values for the *WSJ* dataset.

Table 1. Statistics of collections (all values before stopwords removal unless indicated).

Collection	Collection size (# doc)	Avg. Doc. length (# word)	Vocabulary	% words			Slope $\|\alpha\|$
				stopwords	out-of-dictionary	singleton	
Kongregate	144,161	4.449	35,208	44.90	58.94	56.65	1.69
Twitter	977,569	13.989	364,367	44.99	68.37	66.95	1.54
Myspace	144,161	38.077	187,050	50.67	69.61	53.30	1.92
Slashdot	141,283	98.912	123,359	54.00	57.31	44.82	2.17
WSJ	173,252	452.005	226,469	41.45	67.57	34.33	2.70

3 Comparative analysis of datasets

We started our study by checking the average length of the documents present in each collection. We found that these user-generated documents are 5 to 100 times shorter then the ones in the traditional TREC collections (Table 1) and we performed a double analysis. First we indexed the documents without removing any stopwords, then we used a standard stopwords list to clean them.

We expected less terms to be discarded as stopwords, since we assume short documents (in particular the ones used to "chat" as *Twitter* or *Kongregate*) to be written "quicker and dirtier", with no care for orthography and using a lot of abbreviations. We found a proof of this when looking at the percentage of terms

[1] Dataset and details available at `http://caw2.barcelonamedia.org/`

[2] Datasets and details available at `http://trec.nist.gov/data/test_coll.html`

which occurred only once in the collection ("singleton terms"): short documents contain more singleton terms, which we can consider as spelling mistakes or mistyped words. This is more evident when we look at out-of-dictionary terms. These words are not contained in a standard dictionary and are identified as misspelled by a spell checker. Although the percentage of out-of-dictionary terms is similar across all datasets, we notice that for short documents collections this value is closer to the number of singleton words (from 2% to 16%), while for traditional TREC collections the distance is further (33%). This fact may indicate that in the short documents collections the presence of more singleton words could be considered as an indicator of a greater number of mistyped words. This is not the case of the traditional TREC collections, where the presence of singleton words is less evident and can be explained by the usage of particular terms such as geographical locations, foreign words or first names which are orthographically correct but not present in the spell checker used.

After this initial analysis, we took inspiration from the work of Serrano et al. [3] to investigate in more details the words distribution for each collection. We concentrate our study on two measures, the slope of the Zipf-Mandelbrot distribution and the vocabulary growth, also known as Heaps' law.

The Zipf-Mandelbrot law can be written as follows [1]:

$$\log f(w) = \log C - \alpha \; \log\left(r(w) - b\right) \tag{1}$$

where $f(w)$ denotes the frequency of a word w in the collection and $r(w)$ is the ranking of the word (in terms of its frequency), while C and b are collection specific parameters. As can be seen in Fig. 1 (left), in a log-log scale and for large values of $r(w)$, the relationship between frequency and rank of a word can be approximated with a descending straight line of slope $-\alpha$. Values for the slope α are given in Table 1 and have been calculated with χ^2 metric [4]. As expected a linear graph is observed also for short documents. Moreover we noticed a dependence between the length of the documents and the slope: the collections containing longer documents tend to have a larger negative slope, which may mean that the words in them are repeated more frequently, while the collections containing shorter documents are less repetitive (as stated previously).

Fig. 1 (right) shows the vocabulary growth with respect to the size of the whole collection. The vocabulary of user-generated short documents grows much faster in comparison with that of longer documents. This means that the conversation between users (in *Kongregate* and *Twitter*) tends to vary greatly with the usage of ever more terms. This may be in part explained by the high percentage of singleton and out-of-dictionary mistyped words or abbreviation that are continuously introduced during the dialog. We also noticed a relationship between the decreasing value of the slopes of the Zipf's law and the growth of the vocabulary: *Twitter* has the minimum slope but the maximum vocabulary growth, to the contrary *WSJ* has the maximum slope and the minimum vocabulary growth. This could, again, be explained by the high frequency of mistyped terms in the vocabulary of user-generated short documents in comparison to the standard TREC documents.

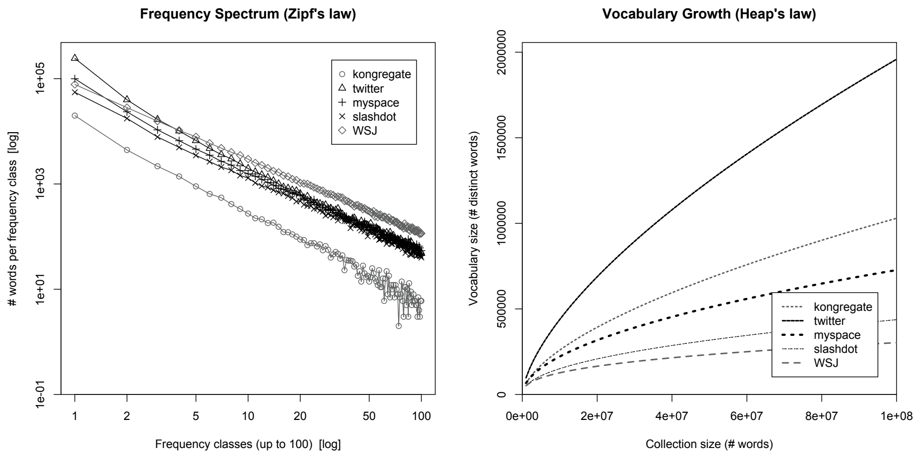

Fig. 1. Zipf's law (left) and Vocabulary Growth (right). We display only graphs after stopword removal (similar to others before).

4 Conclusions and Future work

In this work we compared user-generated short documents and standard online datasets over an initial set of properties. We were able to identify the "messy" properties of user-generated short documents, which need therefore to be preprocessed before being treated with standard techniques. These seem to be easily applicable given the Zipf nature of the short documents. In the future we would like to compare user-generated short documents with a dictionary of common online abbreviation as well as mistyped words and to enlarge the number of collections analyzed (by adding *Ciao*, Blog or Tripadvisor collections). We would also like to study other measures (such as term distribution similarity and burstiness [3]) and to investigate further the differences between *discussion-style* and *chat-style* text content [5] that we noticed but did not discuss in here.

References

1. C.D.Manning and H.Schütze. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, MA, 1999.
2. J.Codina, A.Kaltenbrunner, J.Grivolla, R. E.Banchs, and R.Baeza-Yates. Content analysis in web 2.0. In *18th International World Wide Web Conference*, 04 2009.
3. M. Serrano, A. Flammini, and F. Menczer. Modeling statistical properties of written text. *PLoS ONE*, 4(4):e5372–, 04 2009.
4. S.Evert and M.Baroni. *zipfR: Statistical models for word frequency distributions*, 2008. R package version 0.6-5.
5. D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards. Detection of harassment on web 2.0. In *CAW 2.0 '09: Proceedings of the 1st Content Analysis in Web 2.0 Workshop*, Madrid, Spain, 2009.