Identifying Low-Value Words in Text Corpora

Geoffrey P. Morgan
gmorgan@cs.cmu.edu

Kathleen M. Carley
kathleen.carley@cs.cmu.edu

Abstract

Delete Lists are lists of words that have been determined to have little useful meaning for textual analysis. One subset of words that are frequently deleted are stop-words. Stop-Words are textual tokens, such as "and", "a", or "the", that provide structural or grammatical impact to a sentence but do not themselves have significant inherent meaning.

Identifying stop-words is a routine process in most text-cleaning applications, but frequently is done via user-maintained word lists. I suggest that the corpora comparison technique I devised for word-score polarization can be used to identify low-value words while preserving the bulk of the text tokens. I will use both known and random draw corpora comparisons for this process.

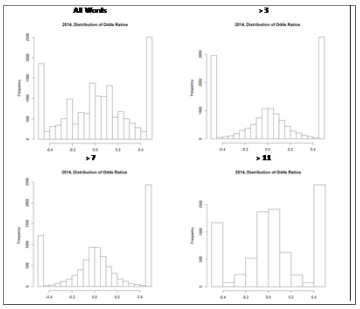
By "known" corpora, I mean corpora drawn from explicit data-sources, the emails of one company and the emails of another, for example. "Random-Draw" corpora are created by drawing document sets at random, and therefore this technique could be applied to any sufficiently large text corpus of interest. I use the ability to identify stop words as a proxy for performance in generating useful delete lists.

Comparing Corpora to Identify Terms

When we have a reason to compare two corpora on some basis, such as documents drawn from the same time-period, we can use the odds of whether a particular term (t) will be in one of the two document sets to identify key terms that distinguish the two document sets. We call these document sets A and G. The complete term set is notated as T.

Because we're using an odds-ratio, we use threshold values for a term to remain in the corpus. A term must appear at least as many times as the cut-off threshold.

Comparing two known corpus, we can evaluate performance by the algorithm's ability to identify words from a stop-word list.



Harnessing the Law of Large Numbers to Identify Low-Value Words

By repurposing the technique, we can identify valuable words in a given corpus without a comparison corpus.

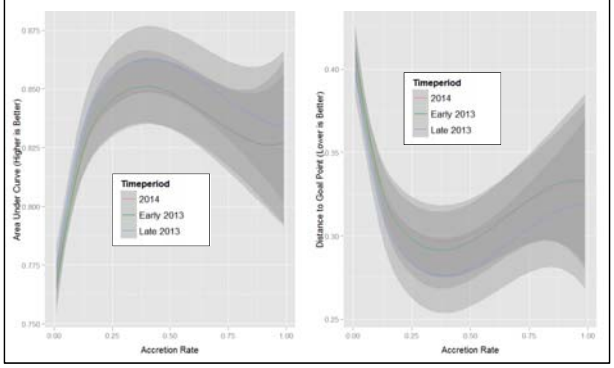
We vary Document Draw Sizes (D) and Accretion Rates (A), and use different document sets to test sensitivity. Each draw generates several sub-samples from the corpus (the number of documents is determined by D). The accretion rate is the number of terms from the draw which are marked as "low-value".

We evaluate performance by comparing the terms with the highest "low-value" count against a validated Delete List. We look at both the area under the curve (higher is better) and the distance closest to the ideal 1,0 point.

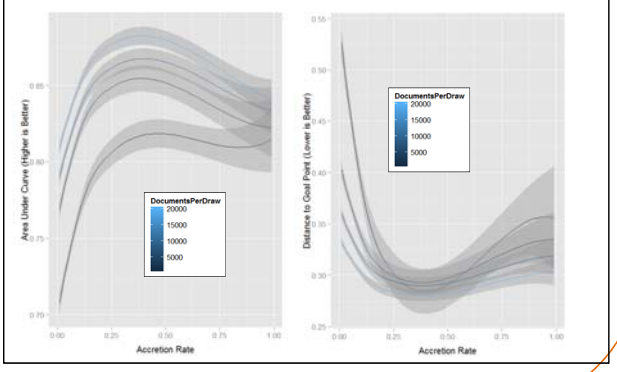
Factor	# of Values	Values
Corpus (***C***)	3	Early 2013, Late 2013, 2014
Document Draw Size (***D***)	4	1000, 5000, 10000, 20000
Accretion Rate (***A***)	8	1, 5, 10, 20, 40, 60, 80, 99
Constants	**Setting**	
Filter Value	3	
Number of Draws	1000	
Outcomes		
Best Performance	The distance of the performance point closest to 0,1	
Average Performance	Area under the curve	

Total Combinations	96
Repetitions	10
Total Runs	960

Corpus Comparisons



Document Draw Size Comparisons



This work was supported in part by the Office of Naval Research (ONR) through a Minerva N000141512797 on dynamical statistical informatics and a MURI N000140811186 on adversarial reasoning, and the Center for Computational Analysis of Social and Organization Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research, or the U.S. government.