

# Mining web text for brand associations

Robert Malouf

Department of Linguistics and Oriental Langs  
San Diego State University

Bradley Davidson & Ashli Sherman

CommonHealth  
Wayne NJ

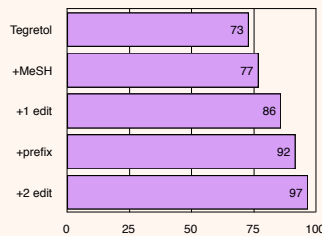
## Introduction

Web text, such as blogs, newsgroups, message boards, and email lists, can provide an easily collected and incredibly rich source of data on a nearly limitless range of topics. With this project, we are taking the first steps towards developing a methodology for mining marketing intelligence from web texts.

The corpus we are working with is a collection of posts to a number of Internet discussion groups and other websites used by epilepsy patients and their families. The corpus contains a total of 26,062,526 words in 316,373 posts from 19 different sites and 8,731 distinct users. Posts average 119 words each.

## Finding mentions

We use finite state automata to identify mentions of medication names in posts. We include the brand name (*Tegretol*), alternate names listed in MeSH (*carbamazepine*, *Amizipine*), terms with an edit distance of 1 (*tegreatol*, *tegetol*), prefixes of names (*teg*, *tegre*), and terms with an edit distance of 2 (*tegrital*).



Using an edit distance of 2 yields the best recall, but precision falls to 90.8%.

## Extracting keywords

The next step is to collect a set of candidate *keywords* which (potentially) reflect the issues surrounding the brand names which users find salient, using the *pointwise mutual information* between each brand name *b* and each term *w<sub>i</sub>*:

$$\text{score}(w_i, b) = \frac{f(w_i, b)}{N} \times \log \frac{N \times f(w_i, b)}{f(w_i) \times f(b)}$$

Out of the 20,505 terms which occurred 15 or more times, we selected 1,001 key words (the top 5% by PMI).

Next, we represent the distribution of each keyword as vector of content-bearing words that appear nearby:

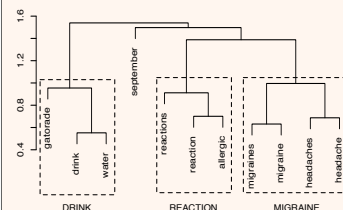
I was told not to worry about it and that it would not be causing my symptoms: burning pain, numbness, balance and coordination problems, jolts, speech problems, stiffness, etc.

The dimensionality of these vectors is reduced by SVD, yielding a representation of each term in a 100-dimensional latent semantic space.

Given this representation, we can measure the semantic distance between any two terms as:

$$\text{dist}(w_i, w_j) = 1 - \frac{w_i \cdot w_j}{\|w_i\| \|w_j\|}$$

Using this distance metric, we performed a complete-linkage hierarchical clustering of the keywords.



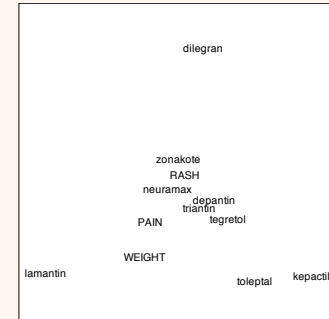
## Keyword clusters

We label each cluster with the term which is nearest to the cluster's centroid. Many (but not all) of the clusters reflected plausible brand associations:

- **MEMORY:**  
loss memory problem cognitive term short concentration speech trouble confusion recall concentrate coordination inability
- **DEFECTS**  
pregnancy pregnant risk birth defects women pregnancies risks baby childbearing dangerous trimester fetus
- **QUICKLY**  
finally eventually quickly fast awhile
- **PROV**  
fifty wbschool prov apples steven spoken silver settings brandy gold
- **SHAWN**  
shawn emily multiple tie hemiplegic

## Visualization

To visualize the terms and their associations, we represent each brand as a vector of PMI scores to produce a 160-dimension 'association space' with issue clusters as the basis vectors. We then plot brand names and clusters in two dimensions using Independent Component Analysis.



## Finding associations

We next find the issues which are most closely associated with each brand name, again using PMI:

- **Tegretol**  
CARBATROL DEFECTS DRUGS BLOOD PARTIAL SEIZURES YRS KG CARERS CP
- **Zonakote**  
EPILEPSY MG RASH DEFECTS DOSE CARBATROL SWITCHING KATHY NIGHT JME
- **Lamantin**  
WEIGHT DRINKING PROV PAIN EFFECT EAT MEMORY DOSE HEADACHE SHAWN

## References

- Church, K. W., and Hanks, P. 1989. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Meeting of the ACL*, 76–83.
- Marchini, J. L.; Heaton, C.; and Ripley, B. D. 2004. *fastICA: FastICA algorithms to perform ICA and Projection Pursuit*. R package version 1.1-6.
- Schütze, H. 1997. *Ambiguity Resolution in Language Learning*. Stanford: CSLI Publications.
- Turney, P. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 417–424.
- Widdows, D.; Cederberg, S.; and Dorow, B. 2002. Visualisation techniques for analysing meaning. In *Fifth International Conference on Text, Speech and Dialogue*, 7–115.