Toward Computer-Assisted Text Curation: Classification Is Easy (Choosing Training Data Can Be Hard…)

Robert Denroche[1], Ramana Madupu[2], Shibu Yooseph[2],
Granger Sutton[2], and Hagit Shatkay[1]

[1] Computational Biology and Machine Learning Lab,
School of Computing, Queen's University, Kingston, Ontario, Canada
{Denroche,Shatkay}@cs.queensu.ca
[2] Informatics Department, J. Craig Venter Institute, Rockville, Maryland, United States
{RMadupu,SYooseph,GSutton}@jcvi.org

Abstract. We aim to design a system for classifying scientific articles based on the presence of protein characterization experiments, intending to aid the curators populating JCVI's Characterized Protein (CHAR) Database of experimentally characterized proteins. We trained two classifiers using small datasets labeled by CHAR curators, and another classifier based on a much larger dataset using annotations from public databases. Performance varied greatly, in ways we did not anticipate. We describe the datasets, the classification method, and discuss the unexpected results.

Keywords: Classification, Biomedical Text Mining, Text Categorization, Database Curation, Imbalanced and Sparse Data.

1 Introduction

The Characterized Protein Database (CHAR) is a resource currently being developed by the J. Craig Venter Institute (JCVI) in support of their prokaryotic genome annotation pipeline, as well as broader annotation efforts. Records in the database include: protein name, gene symbol, organism name, GO terms, and synonymous accessions in other public databases. Moreover, each protein is linked to the scientific publications reporting the experimental results that characterize it. Taxon specific, functional information is drawn manually from the referenced publications by curators. CHAR is used within JCVI's auto-annotation pipeline to annotate novel prokaryotic gene products that are homologous to proteins already curated in CHAR. When CHAR is complete and populated, it is planned to be a high-quality publicly available resource for both prokaryotic and eukaryotic proteins.

As noted before by shared annotation tasks (e.g. KDD'02 [1], TREC Genomics 2004 [2], BioCreAtIvE [3]), the manual curation of each article is a slow process; the curator must locate articles about the species or protein of interest, and determine, by reading at least the abstract, if the article contains an experimental characterization. Our task is to automate parts of this process to reduce the amount of curator time

C. Blaschke and H. Shatkay (Eds.): ISBM/ECCB 2009, LNBI 6004, pp. 33–42, 2010.

needed to populate CHAR. The goal is to develop a system for classifying journal articles as either *relevant* or *irrelevant* to CHAR, based on their title and abstract. Abstract text is used because it is readily available from the PubMed database [4]. Notably, about 84% of the 834 articles manually curated for this task were classifiable as either *relevant* or *irrelevant* to CHAR, based solely on their titles and abstracts.

To train classifiers, we created three datasets, each containing both *relevant* and *irrelevant* articles. Two small sets were built by manual curation and a much larger third set by using references to articles in public databases (Swiss-Prot and GO [5,6]). An additional validation set, consisting of *relevant* articles only, was formed from existing CHAR references that were curated before we began the task.

We use a multi-variate Bernoulli model [7] based on stemmed terms to represent each article, and train a standard naïve Bayes classifier. We expected the classifier trained on the large dataset to perform at least as well as the classifiers trained on the smaller sets. However, the opposite occurred. The latter classifiers outperformed the former on both the training and the held-out validation sets. Moreover, the classifiers trained on the small datasets come close to meeting the 70% recall and 80% precision requirements specified as useful by CHAR curators. This is despite the fact that they were trained only on a relatively small number of manually selected abstracts.

As the use of a large number of GO-curated articles for training a classifier was expected to boost performance, we attempt to explain the reduction in performance that actually occurred when using them. Our analysis shows that term distributions within the articles obtained from GO differ significantly from those associated with the other relevant datasets. Our experiments and results are described throughout the rest of the paper.

2 Dataset Construction

We built and used three datasets of abstracts taken from journal articles for training and testing, and an additional set for validation. The text for all abstracts was retrieved from PubMed [4]. The three training and test sets contain both *positive* examples, i.e. abstracts of articles *relevant* to CHAR, and *negative* (*irrelevant*) ones. Two of these datasets are relatively small, containing at most a few hundreds of manually curated abstracts, and were originally intended for validation only. The third set contains thousands of abstracts referenced from reliable public databases, and was originally intended for the training and testing process. The fourth dataset, used for independent validation, consists only of *relevant* articles (*positive* examples) that were already in CHAR at the onset of the project.

Our first dataset, referred to as the ***Curated Journal*** dataset, contains 96 *positive* and 107 *negative* abstracts. These were collected by a CHAR curator (RM) from every article in four issues from three different journals[1]. These journals are the ones most commonly referenced in CHAR, and three of the four respective issues each contributes at least one reference to CHAR. RM read the title and abstract of each article, and labeled it as either *relevant*, *irrelevant*, or, if unable to determine the relevance from

[1] J. of Bacteriology Vol. 189, #5, 2007, J. of Bacteriology Vol. 189, #15, 2007, J. of Biological Chemistry vol. 257, #19, 1982 and Molecular Microbiology vol. 33, #2, 1999.

the abstract alone, assigned it the label *maybe*. Articles labeled *maybe* were discarded from the dataset and are not used in our experiments. Notably, 86.8% of the examined articles were classifiable as *relevant or irrelevant* based on their title and abstract alone. Table 1 shows the dataset statistics.

Our second dataset, called the ***Curated Swiss-Prot*** dataset, consists of 324 *positive* and 174 *negative* abstracts. To build it, first 300 articles were selected at random from references within Swiss-Prot entries [5], and 300 were collected from entries in CHAR originally populated by an automated process which used Swiss-Prot information considered to suggest experimental characterization. As the reliability of these references was uncertain, RM manually labeled these documents, producing a dataset of 498 abstracts labeled with confidence as *relevant* or *irrelevant*. The statistics for the dataset are shown in Table 2. Again, articles labeled as *maybe* were discarded and are not used in our experiments. Of the 600 articles, 83% (498) were classifiable based on their titles and abstracts alone.

Table 1. Curation Labels for the ***Curated Journal*** Dataset. Numbers shown in boldface indicate documents actually used as *positive/negative* examples in our dataset.

	Relevant	*Irrelevant*	*Maybe*	Total
J. Bacteriol	71	40	20	131
J. Biol Chem	13	65	6	84
Mol Microbiol	12	2	5	19
Total	**96**	**107**	31	234

Table 2. Curation labels for the Curated Swiss-Prot Set. Numbers shown in boldface indicate documents actually used as positive/negative examples in the dataset.

Swiss-Prot articles	*Relevant*	*Irrelevant*	*Maybe*	Total
At random	85	155	60	300
From CHAR	239	19	42	300
Total	**324**	**174**	102	600

As the above datasets are fairly small, we planned to use them for validation only. For robust testing and training of a classifier, we aimed to build a much larger dataset, denoted ***SP-GO***, utilizing curated labels assigned to journal articles by online public databases (Swiss-Prot [5] and GO [6]). We expected that adding many reliable, publically available, curated relevant examples would improve the classifier, supporting future automated curation in CHAR.

Swiss-Prot [5] entries hold references to PubMed articles, labeled to describe the information provided by the article. Articles labeled 'CHARACTERIZATION' satisfy the formal criterion of relevance to CHAR [8]. We thus collected all the articles in Swiss-Prot that were so labeled, providing 1,451 *positive* examples. The Gene Ontology project [6] also includes references to articles from PubMed that support ontology assignments. Evidence codes denote the type of information present in the referenced article. Articles assigned an experimental evidence code (EXP, IDA, IPI, IMP, IGI or IEP), satisfy the criterion of relevance to CHAR [9]. CHAR has even established an automated process for migrating such GO annotations into CHAR; as such, articles bearing the above GO codes suggest a reliable extensive data source. We thus collected all articles bearing the above evidence codes from GO, adding 8,403 *positive* examples.

Negative examples are harder to define: articles with non-experimental evidence codes in GO are not necessarily *irrelevant*, while articles in Swiss-Prot that are not labeled 'CHARACTERIZATION' may still carry experimental characterization [8,9]. To overcome this difficulty, and to obtain as *negative* examples articles associated with proteins that are unlikely to be experimentally characterized, we utilized flags attached to GenBank [10] sequence entries. GenBank entries typically carry an *'experimental'* flag for genes/proteins that have been experimentally characterized. CHAR curators estimate (based on experience) that in 80-90% of the cases where a flag is not present, the sequence is indeed not experimentally characterized. We thus gathered 10,012 GenBank entries not bearing the *'experimental'* flag, mapped the sequences (by identity) to their respective Swiss-Prot entries, and gathered all the articles referenced from these Swiss-Prot entries, resulting in 67,892 articles. Of these, 5.5% were found in the *positive* dataset, (which agrees with the estimate that only 80-90% of the un-flagged sequences are truly uncharacterized), and discarded. To have an equal number of positive and negative examples we selected 9,854 articles, at random, from the resulting negative pool.

The third dataset, ***SP-GO***, thus consists of 9,854 *relevant* abstracts (1,451 Swiss-Prot, 8,403 GO), and 9,854 *irrelevant* abstracts, selected at random from the large negative set, as described above. As many of the articles in the *negative* set have an earlier publication date than many of the *positive* ones (data not shown), to avoid temporal artifacts in the classification (*conceptual drift* as noted in TREC Genomics [2]), the sampling of the 9,854 negative documents was biased toward recent publications.

Finally, as an independent validation set, we used all articles (255 abstracts) that were referenced from fully curated CHAR entries, as of May 2008, when we started the project. We refer to this dataset as the ***CHAR*** dataset. We note that this dataset consists of *positive* (*relevant*) abstracts only.

3 Methods and Classification

The titles and abstracts of the articles in all the datasets were downloaded from PubMed, tokenized into unigrams and bigrams, and stemmed using Porter stemming [11], to obtain a set of statistical terms. Stop words were removed, as were frequent terms (occurring in more than 60% of the abstracts) and rare terms, (occurring in fewer than 3 abstracts).

For classification, documents are represented using the multi-variate Bernoulli model [7], and a standard naïve Bayes classifier was implemented [12]. Under the naïve Bayes model the probability of a document given a class is calculated as:

$$\Pr(doc \mid class) = \prod_{\substack{\textit{terms in}\\ \textit{document}}} \Pr(term \mid class) * \prod_{\substack{\textit{terms not in}\\ \textit{document}}} \left(1 - \Pr(term \mid class)\right) .$$

We use a naïve Bayes classifier as it is simple to implement and modify, while its performance is comparable to that of other classifiers. The potential performance gain by using a different classifier (e.g. SVM [13,14]), is negligible compared to the differences in performance observed by varying the training sets as shown below.

Classification results are evaluated using recall, precision and accuracy, which are all defined in terms of number of *true positives* (*TP*), number of *true negatives* (*TN*), number of *false positives* (*FP*) and *number of false negatives* (*FN*), as:

$$recall = \frac{TP}{TP+FN}, \quad precision = \frac{TP}{TP+FP}, \quad accuracy = \frac{TP+TN}{TP+FP+TN+FN}.$$

To examine the differences between the various datasets, we calculate the Kullback-Leibler (KL) divergence [15], among the term distributions within the datasets, where the distribution per dataset is a multinomial over all the terms used to represent the abstracts. The KL divergence measures the difference between two probability distributions. We use a symmetric version [15] defined, for two multinomial probability distributions over n events, P and Q where $P = \{p_1,\ldots,p_n\}$ and $Q = \{q_1,\ldots,q_n\}$, as:

$$KL(P,Q) = \sum_{i=1}^{n} (p_i - q_i)\log_2 \frac{p_i}{q_i}.$$

The KL divergence is non-negative; the lower it is, the more similar the distributions are.

4 Experiments and Results

According to CHAR curators, for a classifier to be useful for their task, its recall should be at least 70% and its precision at least 80%. With this in mind, we initially trained a naïve Bayes classifier using the ***SP-GO*** dataset. Given the magnitude and the design of the dataset (where abstracts were chosen based on curator input and annotations in public databases), we expected a classifier trained on it to perform well. We tested the classifier on both the ***Curated Swiss-Prot*** and ***Curated Journal*** datasets. Contrary to our expectations, the ***SP-GO*** trained classifier performs poorly, placing an incorrect label on more than half of our manually curated articles, as is reported in Table 3.

Table 3. ***SP-GO*** performance on the ***Curated Journal*** and ***Curated Swiss-Prot*** datasets

	Curated Journal	***Curated Swiss-Prot***
Recall	30.21%	14.20%
Precision	34.52%	51.11%
Accuracy	39.90%	35.34%

To ensure that, in each dataset, the *positive* vs. *negative* abstracts are indeed distinguishable and "classifiable", we performed complete 5-fold cross-validation, repeated 10 times (using 10 different 5-fold-splits), for each of the three datasets. Table 4 shows the results, averaged over the 5-folds and over all the 10 complete cross-validation runs.

The classifiers based on each set all perform well in the cross-validation setting. All satisfy the requirement of 70% recall, and the ***Curated Swiss-Prot*** set satisfies 80% precision as well. These results clearly show that within each dataset, classification is possible.

To verify that a classifier trained on one dataset can still perform well on another, we trained a classifier using each of the small datasets (***Curated Journal*** and ***Curated Swiss-Prot***), and tested on the other. Table 5 shows the results. The performance is much better than that of the classifier based on ***SP-GO,*** despite the smaller training sets, which were each drawn from a different data source.

For a final validation, we used each of the three trained classifiers (the above two and the one trained on ***SP-GO***) to classify the ***CHAR*** dataset, which consists of references that were already in CHAR when our project began. Results are shown in Table 6. As the ***CHAR*** dataset contains only *relevant* (*positive*) articles, only accuracy can be reported.

While the classifier trained on the ***SP-GO*** dataset labels most of the ***CHAR*** abstracts as irrelevant, the other two classifiers do label most of the articles correctly, even though they were trained on relatively little data. This is despite the fact that articles in ***SP-GO*** were curated and labeled in public databases using criteria seemingly equivalent to those employed by CHAR.

Table 4. Average results from 10 times 5-fold cross-validation classification performed over each dataset. (Standard deviation in parentheses.)

	SP-GO	***Curated Journal***	***Curated Swiss-Prot***
Recall	73.3% (0.12%)	72.5% (2.92%)	89.4% (0.77%)
Precision	73.4% (0.14%)	71.0% (1.70%)	84.2% (0.53%)
Accuracy	73.4% (0.12%)	73.0% (1.85%)	82.2% (0.37%)

Table 5. Performance of classifiers trained over one curated set and tested on the other

Trained with:	***Curated Journal***	***Curated Swiss-Prot***
Tested on:	***Curated Swiss-Prot***	***Curated Journal***
Recall	80.86%	93.75%
Precision	72.98%	52.33%
Accuracy	68.07%	56.65%

Table 6. Accuracy of the three classifiers, over the ***CHAR*** dataset

	SP-GO	***Curated Journal***	***Curated Swiss-Prot***
Accuracy	26.67%	80.00%	87.84%

To further explore the difference between the ***SP-GO*** and the other datasets, we calculate the Kullback-Leibler (KL) divergence between the term distributions associated with each dataset, separately examining the *positive* articles and the *negative* ones in each dataset. Ideally, the term distributions of two *positive* sets should be similar (low divergence) while term distributions between any pair of *positive* and *negative* sets should show much difference (higher divergence).

Table 7 shows the KL divergence between pairs of the *negative/positive* classes from all the datasets as well as the ***CHAR*** set. These values show that the *positive* abstracts in the small sets (***Curated Swiss-Prot***, ***Curated Journal***) and in ***CHAR*** share relatively similar term distributions (relatively low KL divergence), while the difference between every pair of their *positive* and *negative* subsets is much higher in terms of KL divergence. In contrast, both the *negative* and the *positive* subsets of ***SP-GO*** are almost equally dissimilar to the *positive* subset of ***Curated Journal***; moreover, the *positive* subset of ***SP-GO*** is less similar to both the ***CHAR*** dataset and to the *positive* subset of ***Curated Swiss-Prot*** (higher KL), than the *negative* ***SP-GO*** subset is (see italicized numbers in Table 7). These properties of the ***SP-GO*** term distributions, are consistent with the results reported above, that the classifier trained on ***SP-GO*** labeled most of the ***CHAR*** articles as irrelevant (Table 6).

As the ***SP-GO*** *relevant* (*positive*) abstracts originated from two distinct sources, namely Swiss-Prot and GO, to better understand the phenomenon, we separated ***SP-GO*** into two: the 1,451 abstracts collected from Swiss-Prot (denoted ***SPonly***) and the 8,403 abstracts collected from GO (***GOonly***). Table 8 shows the KL divergence between the term distributions of these two sets and the other datasets.

Table 7. Kullback-Leibler Divergence between dataset pairs (***CurSP*** denotes our ***Curated Swiss-Prot*** dataset; ***CurJol*** denotes our ***Curated Journal*** dataset. *Pos* indicates the *positive* portion of the dataset, *Neg* indicates *negative.*)

	CurJol *Pos*	***CurJol*** *Neg*	***CurSP*** *Pos*	***CurSP*** *Neg*	***CHAR*** (*Pos*)
SP-GO *Pos*	1.1442	1.2714	*0.9622*	0.5624	*0.9224*
SP-GO *Neg*	1.1733	1.3784	*0.5790*	0.2200	*0.6752*
CurJol *Pos*			0.8167	1.5319	0.7302
CurJol *Neg*			1.4791	1.7227	1.3580
CurSP *Pos*					0.3175
CurSP *Neg*					0.9667

Table 8. KL divergence between each of the ***SPonly*** and ***GOonly*** sets and all other subsets. (***CurSP*** denotes our ***Curated Swiss-Prot*** dataset; ***CurJol*** denotes our ***Curated Journal*** dataset. *Pos* indicates the *positive* portion of the dataset, *Neg* indicates *negative.*)

	CurJol *Pos*	***CurJol*** *Neg*	***CurSP*** *Pos*	***CurSP*** *Neg*	***CHAR*** (*Pos*)
SPonly (*Pos*)	0.9541	1.2336	0.3825	0.5031	0.5010
GOonly (*Pos*)	1.3226	1.3920	1.2130	0.6744	1.1381

Table 8 clearly shows that abstracts labeled as bearing experimental characterization by Swiss-Prot (***SPonly***) are more similar in their term-distribution to the *positive* abstracts in our hand-curated sets than to the *negative* abstracts. In contrast, term-distributions of the abstracts that contain experimental characterization based on GO (***GOonly***) are dissimilar to the distributions of the *positive* abstracts in our curated sets, and actually even more similar to the term distributions of the *negative* abstracts in one of the sets (***Curated SwissProt***). The apparent discrepancy between GO's notion of abstracts bearing experimental characterization and CHAR's is particularly noteworthy, given that GO curated abstracts were considered as a mainpossible source in support of CHAR's curation.

A naïve Bayes classifier trained on the 1,451 ***SPonly*** *positives* and 1,451 *negative* articles selected at random from the ***SP-GO*** *negatives*, demonstrates improved performance, closer to that obtained by classifiers trained on the small curated datasets. Performance is best on the ***Curated Swiss-Prot*** dataset where both the recall and the precision surpass the requirements set out by the CHAR curators. The results are shown in Table 9.

Table 9. ***SPonly*** performance on the ***Curated Journal***, ***Curated Swiss-Prot*** and ***CHAR*** dataset

	Curated Journal	***Curated Swiss-Prot***	***CHAR*** *(Pos)*
Recall	73.96%	70.37%	
Precision	52.21%	83.82%	
Accuracy	55.67%	71.89%	70.98%

When attempting to explain the results, our initial hypothesis was that terms indicative of a specific species or of species from specific kingdoms (such as common names, scientific names or strain identifiers) may be responsible for the differences in term distributions between the datasets. We found that terms typically associated with common model organisms, such as *E.coli*, mouse and yeast, are overrepresented in *relevant* articles, as proteins from these species are more likely to be experimentally characterized; we also found that articles from Swiss-Prot and from GO report on species from different kingdoms at different rates[2]. We have conducted experiments, completely removing all terms indicative of a specific species or, alternatively, replacing such terms with the species-generic pseudo-term SPECIES_PLACEHOLDER to capture all occurrences of a species-related term in the abstracts. While these strategies improved results slightly (1-2%), and may help prevent bias for certain species in future classification, this small effect does not explain the major difference in performance across the different datasets.

Our current hypothesis is that articles from GO cover a few types of experiments that are relevant to CHAR heavily and rarely mention others. The articles we collected from GO are not evenly distributed across the six experimental evidence codes (EXP, IDA, IPI, IMP, IGI or IEP). Approximately 70% of the articles are associated with either 'Inferred from Direct Assay' (IDA) or 'Inferred from Physical Interaction' (IPI). These codes are both strongly related to binding assay experiments [9]. Related terms such as **interact** or **complex** are therefore highly overrepresented in the ***SP-GO*** *positive* set, which mostly consists of articles from GO. It is likely that the ***SP-GO*** *positives* contain many articles based only on a few forms of experimental characterizations, leading to a classifier that labels articles reporting other types of experiments as *irrelevant,* regardless of their actual relevance to CHAR This may contribute to the low recall shown by the ***SP-GO*** classifier in Table 3, and may explain why classification performance is improved by removing the GO articles.

The *negative* articles in the ***SP-GO*** dataset were collected from Swiss-Prot entries of proteins that, based on the selection process, are expected to be uncharacterized. Such articles, which are associated with protein sequences that have not been experimentally

[2] PubMed references in GO (which make up 85% of the ***SP-GO*** *positives*) were 67% prokaryote, 14% eukaryote and 20% virus when we collected them. Sequence entries in Swiss-Prot when we collected our articles (100% of the ***SP-GO*** *negatives*) were 57% prokaryote, 36% eukaryote, 4% archaea and 3% virus [16].

characterized, are likely to be initial papers pertaining to the proteins, reporting their genomic and proteomic sequence. This is corroborated by the fact that the terms **sequence**, **genome** and **cdna** are highly overrepresented in the ***SP-GO*** *negatives*. It is likely that the ***SP-GO*** *negative* set thus contain many *irrelevant* articles discussing sequences, while it lacks *irrelevant* articles of other types. Furthermore, since the *negative* articles in the ***Curated Journal*** dataset are less likely to discuss sequence data than the *negative* articles in the ***SPonly*** dataset, classifiers trained on the ***SPonly*** dataset may erroneously label the *irrelevant*, non-sequence articles in the ***Curated Journal*** dataset as *positive*, which helps explain the low precision and low overall accuracy of the ***SPonly*** classifier over the ***Curated Journal*** dataset shown in Table 9.

5 Conclusion

We trained naïve Bayes classifiers to identify abstracts that are likely to describe experimental characterization of proteins (*relevant*), as opposed to abstracts unlikely to contain such information (*irrelevant*). The classifiers are intended to support the curation of abstracts for JCVI's CHAR (Characterized Protein) database. To train and test the classifiers we constructed small hand-curated datasets, as well as a large set based on previously curated abstracts from Swiss-Prot and GO. We expected the latter set to support training a well-performing classifier, given the dataset size and the careful consideration invested in its construction. Specifically, given that, by definition, articles containing experimental characterization are relevant to CHAR.

While the abstracts bearing the 'CHARACTERIZATION' label in Swiss-Prot proved to be effective *positive* examples, articles bearing experimental characterization flags in GO did not. The latter was particularly surprising, as GO was considered up to this point as the most likely source of information for CHAR.

Most notably, and on the positive side, we have shown that classifiers trained on relatively small hand-curated datasets perform at a high level – very close to the level required by CHAR curators to be useful – both when classifying the other sets' documents, and when classifying the left-out validation set (CHAR).

Another interesting finding of this study was that in more than 80% of the manually examined articles, the abstract and title alone contained sufficient information for determining the relevance of the article for the CHAR curation. While the actual evidence is most likely to be found in the full text, it is important to note that the coarser task of just determining relevance can be performed (in most cases) using the title and the abstract alone.

References

1. Yeh, A., Hirschman, L., Morgan, A.: Background and Overview for KDD Cup 2002 Task 1: Information Extraction from Biomedical Articles. In: ACM SIGKDD Explorations Newsletter (2002)
2. Cohen, A., Bhupatiraju, R.T., Hersh, W.R.: Feature Generation, Feature Selection, Classifiers, and Conceptual Drift for Biomedical Document Triage. In: 13th Text Retrieval Conference -TREC 2004, Gaithersburg, MD (2004)

3. Blaschke, C., Leon, E.A., Krallinger, M., Valencia, A.: Evaluation of BioCreAtIvE assessment of task 2. BMC Bioinformatics 6 (Suppl. 1), S16 (2005)
4. PubMed, `http://www.ncbi.nlm.nih.gov/pubmed`
5. Swiss-Prot Protein Knowledgebase, `http://ca.expasy.org/sprot/`
6. The Gene Ontology Project, `http://www.geneontology.org/`
7. McCallum, A., Nigam, K.: A Comparison of Event Models for Naive Bayes Text Classification. In: Learning for Text Categorization Workshop, AAAI 1998 (Tech. Report WS-98-05) (1998)
8. Swiss-Prot Protein Knowledgebase: A Primer on UniProtKB/Swiss-Prot Annotation, `http://www.uniprot.org/docs/annbioch`
9. The Gene Ontology Project: Guide to GO Evidence Codes, `http://www.geneontology.org/GO.evidence.shtml`
10. GenBank, `http://www.ncbi.nlm.nih.gov/Genbank`
11. Porter, M.F.: An Algorithm for Suffix Stripping. Program 14, 130–137 (1980)
12. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. John Wiley & Sons, Inc., New York (2001)
13. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398. Springer, Heidelberg (1998)
14. Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G.D., Michalickova, K., Pawson, T., Hogue, C.W.V.: PreBIND and Textomy – mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics 4(11) (2003)
15. Kullback, S., Leibler, R.A.: On Information and Sufficiency. Annals of Mathematical Statistics 22, 79–86 (1951)
16. Swiss-Prot Protein Knowledgebase: Release Notes for UniProtKB Release (July 22, 2008), `http://www.expasy.ch/txt/old-rel/relnotes.56.htm`