Master-Seminar 1: Data Analytics
Tuesday 14:00 – 16:00
Instructors: Lars Schmidt Thieme, Mofassir ul Islam Arif

Recommended Read:

A: Machine learning in automated text categorization

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*.

Abstract:
The automated categorization (or classification) of texts into predefined categories has witnessed a booming interest in the last 10 years, due to the increased availability of documents in digital form and the ensuing need to organize them. In the research community the dominant approach to this problem is based on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of preclassified documents, the characteristics of the categories. The advantages of this approach over the knowledge engineering approach (consisting in the manual definition of a classifier by domain experts) are a very good effectiveness, considerable savings in terms of expert labor power, and straightforward portability to different domains. This survey discusses the main approaches to text categorization that fall within the machine learning paradigm. We will discuss in detail issues pertaining to three different problems, namely, document representation, classifier construction, and classifier evaluation.

Category : Fundamentals

B-1: Stochastic gradient descent training for L1-regularized log-linear models with cumulative penalty

Tsuruoka, Y., Tsujii, J. I., & Ananiadou, S. (2009). Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP:*

Abstract:
Stochastic gradient descent (SGD) uses approximate gradients estimated from subsets of the training data and updates the parameters in an online fashion. This learning framework is attractive because it often requires much less training time in practice than batch training algorithms. However, L1-regularization, which is becoming popular in natural language processing because of its ability to produce compact models, cannot be efficiently applied in SGD training, due to the large dimensions of feature vectors and the fluctuations of approximate gradients. We present a simple method to solve these problems by penalizing the weights according to cumulative values for L1 penalty. We evaluate the effectiveness of our method in three applications: text chunking, named entity recognition, and part-of-speech tagging. Experimental results demonstrate that our method can produce compact and accurate models much more quickly than a state-of-the-art quasi-Newton method for L1-regularized loglinear models.

B-2: Curriculum Learning

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning* . ACM.

Abstract:
Humans and animals learn much better when the examples are not randomly presented but organized in a meaningful order which illustrates gradually more concepts, and gradually more complex ones. Here, we formalize such training strategies in the context of machine learning, and call them "curriculum learning". In the context of recent research studying the difficulty of training in the presence of non-convex training criteria (for deep deterministic and stochastic neural networks), we explore curriculum learning in various set-ups. The experiments show that significant improvements in generalization can be achieved. We hypothesize that curriculum learning has both an effect on the speed of convergence of the training process to a minimum and, in the case of non-convex criteria, on the quality of the local minima obtained: curriculum learning can be seen as a particular form of continuation method (a general strategy for global optimization of non-convex functions).

B-3: Combined Regression and Ranking

Sculley, D. (2010, July). Combined regression and ranking. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* . ACM.

Abstract:
Many real-world data mining tasks require the achievement of two distinct goals when applied to unseen data: first, to

induce an accurate preference *ranking*, and second to give good *regression* performance. In this paper, we give an efficient and effective Combined Regression and Ranking method (CRR) that optimizes regression and ranking objectives simultaneously. We demonstrate the effectiveness of CRR for both families of metrics on a range of large-scale tasks, including click prediction for online advertisements. Results show that CRR often achieves performance equivalent to the best of both ranking-only and regression-only approaches. In the case of rare events or skewed distributions, we also find that this combination can actually improve regression performance due to the addition of informative ranking constraints.

Category : Text Categorization

C-1: Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?
Leopold, E., & Kindermann, J. (2002). Text categorization with support vector machines. How to represent texts in input space?. *Machine Learning*.

Abstract:

The choice of the kernel function is crucial to most applications of support vector machines. In this paper, however, we show that in the case of text classification, term-frequency transformations have a larger impact on the performance of SVM than the kernel itself. We discuss the role of importance-weights (e.g. document frequency and redundancy), which is not yet fully understood in the light of model complexity and calculation cost, and we show that time consuming lemmatization or stemming can be avoided even when classifying a highly inflectional language like German.

C-2: Effective Use of Word Order for Text Categorization with Convolutional Neural Networks
Johnson, R., & Zhang, T. (2014). Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058*.

Abstract:
Convolutional neural network (CNN) is a neural network that can make use of the internal structure of data such as the 2D structure of image data. This paper studies CNN on text categorization to exploit the 1D structure (namely, word order) of text data for accurate prediction. Instead of using low-dimensional word vectors as input as is often done, we directly apply CNN to high-dimensional text data, which leads to directly learning embedding of small text regions for use in classification. In addition to a straightforward adaptation of CNN from image to text, a simple but new variation which employs bag-of-word conversion in the convolution layer is proposed. An extension to combine multiple convolution layers is also explored for higher accuracy. The experiments demonstrate the effectiveness of our approach in comparison with state-of-the-art methods.

C-3: Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification
Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.

Abstract:
We present a method that learns word embedding for Twitter sentiment classification in this paper. Most existing algorithms for learning continuous word representations typically only model the syntactic context of words but ignore the sentiment of text. This is problematic for sentiment analysis as they usually map words with similar syntactic context but opposite sentiment polarity, such as good and bad, to neighboring word vectors. We address this issue by learning sentimentspecific word embedding (SSWE), which encodes sentiment information in the continuous representation of words. Specifically, we develop three neural networks to effectively incorporate the supervision from sentiment polarity of text (e.g. sentences or tweets) in their loss functions. To obtain large scale training corpora, we learn the sentiment-specific word embedding from massive distant-supervised tweets collected by positive and negative emoticons. Experiments on applying SSWE to a benchmark Twitter sentiment classification dataset in SemEval 2013 show that (1) the SSWE feature performs comparably with hand-crafted features in the top-performed system; (2) the performance is further improved by concatenating SSWE with existing feature set.

D-1: An Effective Approach to Enhance Centroid Classifier for Text Categorization
Tan, S., & Cheng, X. (2007). An effective approach to enhance centroid classifier for text categorization. In *European Conference on Principles of Data Mining and Knowledge Discovery* **.** Springer, Berlin, Heidelberg.

Abstract:
Centroid Classifier has been shown to be a simple and yet effective method for text categorization. However, it is often plagued with model misfit (or inductive bias) incurred by its assumption. To address this issue, a novel Model Adjustment algorithm was proposed. The basic idea is to make use of some criteria to adjust Centroid Classifier model. In this work, the

criteria include training-set errors as well as training-set margins. The empirical assessment indicates that proposed method performs slightly better than SVM classifier in prediction accuracy, as well as beats it in running time.

D-2: Inductive learning algorithms and representations for text categorization
Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*. ACM.

Abstract:
The assignment of natural language texts to one or more predefined categories based on their content -is an important component in many information organization and management tasks. We compare the effectiveness of five different automatic learning algorithms for text categorization in terms of learning speed, real time classification speed, and classification accuracy. We also examine training set size, and alternative document representations. Very accurate text classifiers can be learned automatically from training examples. Linear Support Vector Machines (SVMs) are particularly promising because they are very accurate, quick to train, and quick to evaluate.

D-3: Character-level Convolutional Networks for Text Classification
Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In Advances in neural information processing systems .

Abstract:
This article offers an empirical exploration on the use of character-level convolutional networks (ConvNets) for text classification. We constructed several large-scale datasets to show that character-level convolutional networks could achieve state-of-the-art or competitive results. Comparisons are offered against traditional models such as bag of words, n-grams and their TFIDF variants, and deep learning models such as word-based ConvNets and recurrent neural networks.

Category: Sentiment Analysis

E-1: Thumbs up?: sentiment classification using machine learning techniques
Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*. Association for Computational Linguistics.

Abstract:
We consider the problem of classifying documents not by topic, but by overall sentiment, e.g., determining whether a review is positive or negative. Using movie reviews as data, we find that standard machine learning techniques definitively outperform human-produced baselines. However, the three machine learning methods we employed (Naive Bayes, maximum entropy classification, and support vector machines) do not perform as well on sentiment classification as on traditional topic-based categorization. We conclude by examining factors that make the sentiment classification problem more challenging.

E-2: Twitter as a Corpus for Sentiment Analysis and Opinion Mining
Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREc .

Abstract:
Microblogging today has become a very popular communication tool among Internet users. Millions of users share opinions on different aspects of life everyday. Therefore microblogging web-sites are rich sources of data for opinion mining and sentiment analysis. Because microblogging has appeared relatively recently, there are a few research works that were devoted to this topic. In our paper, we focus on using Twitter, the most popular microblogging platform, for the task of sentiment analysis. We show how to automatically collect a corpus for sentiment analysis and opinion mining purposes. We perform linguistic analysis of the collected corpus and explain discovered phenomena. Using the corpus, we build a sentiment classifier, that is able to determine positive, negative and neutral sentiments for a document. Experimental evaluations show that our proposed techniques are efficient and performs better than previously proposed methods. In our research, we worked with English, however, the proposed technique can be used with any other language.

E-3: Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts
Santos, C., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers.

Abstract:
Sentiment analysis of short texts such as single sentences and Twitter messages is challenging because of the limited contextual information that they normally contain. Effectively solving this task requires strategies that combine the small text content with prior knowledge and use more than just bag-of-words. In this work we propose a new deep convolutional neural network that exploits from character- to sentence-level information to perform sentiment analysis of short texts. We apply our approach for two corpora of two different domains: the Stanford Sentiment Treebank (SSTb), which contains sentences from movie reviews; and the Stanford Twitter Sentiment corpus (STS), which contains Twitter messages. For the SSTb corpus, our approach achieves state-of-the-art results for single sentence sentiment prediction in both binary positive/negative classification, with 85.7% accuracy, and fine-grained classification, with 48.3% accuracy. For the STS corpus, our approach achieves a sentiment prediction accuracy of 86.4%.

F-1: **Recognizing contextual polarity in phrase-level sentiment analysis**
Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* . Association for Computational Linguistics.

Abstract:
This paper presents a new approach to phrase-level sentiment analysis that first determines whether an expression is neutral or polar and then disambiguates the polarity of the polar expressions. With this approach, the system is able to automatically identify the *contextual polarity f*or a large subset of sentiment expressions, achieving results that are significantly better than baseline.

F-2: OpinionMiner: a novel machine learning system for web opinion mining and extraction
Jin, W., Ho, H. H., & Srihari, R. K. (2009). OpinionMiner: a novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* . ACM.

Abstract:
Merchants selling products on the Web often ask their customers to share their opinions and hands-on experiences on products they have purchased. Unfortunately, reading through all customer reviews is difficult, especially for popular items, the number of reviews can be up to hundreds or even thousands. This makes it difficult for a potential customer to read them to make an informed decision. The OpinionMiner system designed in this work aims to mine customer reviews of a product and extract high detailed product entities on which reviewers express their opinions. Opinion expressions are identified and opinion orientations for each recognized product entity are classified as positive or negative. Different from previous approaches that employed rule-based or statistical techniques, we propose a novel machine learning approach built under the framework of lexicalized HMMs. The approach naturally integrates multiple important linguistic features into automatic learning. In this paper, we describe the architecture and main components of the system. The evaluation of the proposed method is presented based on processing the online product reviews from Amazon and other publicly available datasets.

F-3: Coooolll: A Deep Learning System for Twitter Sentiment Classification
Tang, D., Wei, F., Qin, B., Liu, T., & Zhou, M. (2014). Coooolll: A deep learning system for twitter sentiment classification. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)* .
Abstract:
Many real-world data mining tasks require the achievement of two distinct goals when applied to unseen data: first, to induce an accurate preference *ranking*, and second to give good *regression* performance. In this paper, we give an efficient and effective Combined Regression and Ranking method (CRR) that optimizes regression and ranking objectives simultaneously. We demonstrate the effectiveness of CRR for both families of metrics on a range of large-scale tasks, including click prediction for online advertisements. Results show that CRR often achieves performance equivalent to the best of both ranking-only and regression-only approaches. In the case of rare events or skewed distributions, we also find that this combination can actually improve regression performance due to the addition of informative ranking constraints.

G-1: Twitter Sentiment Classification using Distant Supervision
Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*(12).

Abstract:
We introduce a novel approach for automatically classifying the sentiment of Twitter messages. These messages are classified as either positive or negative with respect to a query term. This is useful for consumers who want to research the sentiment of products before purchase, or companies that want to monitor the public sentiment of their brands. There is no

previous research on classifying sentiment of messages on microblogging services like Twitter. We present the results of machine learning algorithms for classifying the sentiment of Twitter messages using distant supervision. Our training data consists of Twitter messages with emoticons, which are used as noisy labels. This type of training data is abundantly available and can be obtained through automated means. We show that machine learning algorithms (Naive Bayes, Maximum Entropy, and SVM) have accuracy above 80% when trained with emoticon data. This paper also describes the preprocessing steps needed in order to achieve high accuracy. The main contribution of this paper is the idea of using tweets with emoticons for distant supervised learning.

G-2: Active learning for imbalanced sentiment classification

Li, S., Ju, S., Zhou, G., & Li, X. (2012). Active learning for imbalanced sentiment classification. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* . Association for Computational Linguistics.

Abstract:
Active learning is a promising way for sentiment classification to reduce the annotation cost. In this paper, we focus on the imbalanced class distribution scenario for sentiment classification, wherein the number of positive samples is quite different from that of negative samples. This scenario posits new challenges to active learning. To address these challenges, we propose a novel active learning approach, named co-selecting, by taking both the imbalanced class distribution issue and uncertainty into account. Specifically, our co-selecting approach employs two feature subspace classifiers to collectively select most informative minority-class samples for manual annotation by leveraging a certainty measurement and an uncertainty measurement, and in the meanwhile, automatically label most informative majority-class samples, to reduce humanannotation efforts. Extensive experiments across four domains demonstrate great potential and effectiveness of our proposed co-selecting approach to active learning for imbalanced sentiment classification.

G-3: Context-Sensitive Twitter Sentiment Classification Using Neural Network

Ren, Y., Zhang, Y., Zhang, M., & Ji, D. (2016, February). Context-Sensitive Twitter Sentiment Classification Using Neural Network. In *AAAI* .

Abstract:
Sentiment classification on Twitter has attracted increasing research in recent years. Most existing work focuses on feature engineering according to the tweet content itself. In this paper, we propose a context based neural network model for Twitter sentiment analysis, incorporating contextualized features from relevant Tweets into the model in the form of word embedding vectors. Experiments on both balanced and unbalanced datasets show that our proposed models outperform the current state-of-the-art

Category: Applications

H-1: PTE: Predictive Text Embedding through Large-scale Heterogeneous Text Networks

Tang, J., Qu, M., & Mei, Q. (2015). Pte: Predictive text embedding through large-scale heterogeneous text networks.
In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining . ACM.

Abstract:
Unsupervised text embedding methods, such as Skip-gram and Paragraph Vector, have been attracting increasing attention due to their simplicity, scalability, and effectiveness. However, comparing to sophisticated deep learning architectures such as convolutional neural networks, these methods usually yield inferior results when applied to particular machine learning tasks. One possible reason is that these text embedding methods learn the representation of text in a fully unsupervised way, without leveraging the labeled information available for the task. Although the low dimensional representations learned are applicable to many different tasks, they are not particularly tuned for any task. In this paper, we fill this gap by proposing a semi-supervised representation learning method for text data, which we call the *predictive text embedding* (PTE). Predictive text embedding utilizes both labeled and unlabeled data to learn the embedding of text. The labeled information and different levels of word co-occurrence information are first represented as a large-scale heterogeneous text network, which is then embedded into a low dimensional space through a principled and efficient algorithm. This low dimensional embedding not only preserves the semantic closeness of words and documents, but also has a strong predictive power for the particular task. Compared to recent supervised approaches based on convolutional neural networks, predictive text embedding is comparable or more effective, much more efficient, and has fewer parameters to tune.

H-2: FastXML: a fast, accurate and stable tree-classifier for extreme multi-label learning

Prabhu, Y., & Varma, M. (2014). Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM.

Abstract:
The objective in extreme multi-label classification is to learn a classifier that can automatically tag a data point with the most relevant subset of labels from a large label set. Extreme multi-label classification is an important research problem since not only does it enable the tackling of applications with many labels but it also allows the reformulation of ranking problems with certain advantages over existing formulations. Our objective, in this paper, is to develop an extreme multi-label classifier that is faster to train and more accurate at prediction than the state-of-the-art Multi-label Random Forest (MLRF) algorithm [2] and the Label Partitioning for Sub-linear Ranking (LPSR) algorithm [35]. MLRF and LPSR learn a hierarchy to deal with the large number of labels but optimize task independent measures, such as the Gini index or clustering error, in order to learn the hierarchy. Our proposed FastXML algorithm achieves significantly higher accuracies by directly optimizing an nDCG based ranking loss function. We also develop an alternating minimization algorithm for efficiently optimizing the proposed formulation. Experiments reveal that FastXML can be trained on problems with more than a million labels on a standard desktop in eight hours using a single core and in an hour using multiple cores.

H-3:Large-scale Multi-label Learning with Missing Labels
Yu, H. F., Jain, P., Kar, P., & Dhillon, I. (2014). Large-scale multi-label learning with missing labels. In *International conference on machine learning* .

Abstract:
The multi-label classification problem has generated significant interest in recent years. However, existing approaches do not adequately address two key challenges: (a) scaling up to problems with a large number (say millions) of labels, and (b) handling data with missing labels. In this paper, we directly address both these problems by studying the multi-label problem in a generic empirical risk minimization (ERM) framework. Our framework, despite being simple, is surprisingly able to encompass several recent labelcompression based methods which can be derived as special cases of our method. To optimize the ERM problem, we develop techniques that exploit the structure of specific loss functions - such as the squared loss function - to obtain effi- cient algorithms. We further show that our learning framework admits excess risk bounds even in the presence of missing labels. Our bounds are tight and demonstrate better generalization performance for low-rank promoting trace-norm regularization when compared to (rank insensitive) Frobenius norm regularization. Finally, we present extensive empirical results on a variety of benchmark datasets and show that our methods perform significantly better than existing label compression based methods and can scale up to very large datasets such as a Wikipedia dataset that has more than 200,000 labels.

I-1: A Machine Learning Approach to Twitter User Classification
Pennacchiotti, M., & Popescu, A. M. (2011). A Machine Learning Approach to Twitter User Classification. *Icwsm*, *11*(1).

Abstract:
This paper addresses the task of user classification in social media, with an application to Twitter. We automatically infer the values of user attributes such as political orientation or ethnicity by leveraging observable information such as the user behavior, network structure and the linguistic content of the user's Twitter feed. We employ a machine learning approach which relies on a comprehensive set of features derived from such user information. We report encouraging experimental results on 3 tasks with different characteristics: political affiliation detection, ethnicity identification and detecting affinity for a particular business. Finally, our analysis shows that rich linguistic features prove consistently valuable across the 3 tasks and show great promise for additional user classification needs.

I-2: Broadly Improving User Classification via Communication-Based Name and Location Clustering on Twitter
Bergsma, S., Dredze, M., Van Durme, B., Wilson, T., & Yarowsky, D. (2013). Broadly improving user classification via communication-based name and location clustering on twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Abstract:
Hidden properties of social media users, such as their ethnicity, gender, and location, are often reflected in their observed attributes, such as their first and last names. Furthermore, users who communicate with each other often have similar hidden properties. We propose an algorithm that exploits these insights to cluster the observed attributes of hundreds of millions of Twitter users. Attributes such as user names are grouped together if users with those names communicate with other similar users. We separately cluster millions of unique first names, last names, and userprovided locations. The efficacy of these clusters is then evaluated on a diverse set of classification tasks that predict hidden users properties such as ethnicity,

geographic location, gender, language, and race, using only profile names and locations when appropriate. Our readily-replicable approach and publiclyreleased clusters are shown to be remarkably effective and versatile, substantially outperforming state-of-the-art approaches and human accuracy on each of the tasks studied.

I-3: Twitter-Based User Modeling for News Recommendations
Abel, F., Gao, Q., Houben, G. J., & Tao, K. (2013). Twitter-Based User Modeling for News Recommendations. In *IJCAI* .

Abstract:
In this paper, we study user modeling on Twitter. We investigate different strategies for mining user interest profiles from microblogging activities ranging from strategies that analyze the semantic meaning of Twitter messages to strategies that adapt to temporal patterns that can be observed in the microblogging behavior. We evaluate the quality of the user modeling methods in the context of a personalized news recommendation system. Our results reveals that an understanding of the semantic meaning of microposts is key for generating high-quality user profiles.

J-1 Web-Search Ranking with Initialized Gradient Boosted Regression Trees
Mohan, A., Chen, Z., & Weinberger, K. (2011). Web-search ranking with initialized gradient boosted regression trees. In *Proceedings of the Learning to Rank Challenge* .

Abstract:
In May 2010 Yahoo! Inc. hosted the Learning to Rank Challenge. This paper summarizes the approach by the highly placed team Washington University in St. Louis. We investigate Random Forests (RF) as a low-cost alternative algorithm to Gradient Boosted Regression Trees (GBRT) (the de facto standard of web-search ranking). We demonstrate that it yields surprisingly accurate ranking results — comparable to or better than GBRT. We combine the two algorithms by first learning a ranking function with RF and using it as initialization for GBRT. We refer to this setting as iGBRT. Following a recent discussion by Li et al. (2007), we show that the results of iGBRT can be improved upon even further when the web-search ranking task is cast as classification instead of regression. We provide an upper bound of the Expected Reciprocal Rank (Chapelle et al., 2009) in terms of classification error and demonstrate that iGBRT outperforms GBRT and RF on the Microsoft Learning to Rank and Yahoo Ranking Competition data sets with surprising consistency.

J-2: **Mining text snippets for images on the web**
Kannan, A., Baker, S., Ramnath, K., Fiss, J., Lin, D., Vanderwende, L., ... & Wang, X. J. (2014). Mining text snippets for images on the web. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM.

Abstract:
Images are often used to convey many different concepts or illustrate many different stories. We propose an algorithm to mine *multiple* diverse, relevant, and interesting text snippets for images on the web. Our algorithm scales to *all* images on the web. For each image, *all* webpages that contain it are considered. The top-K text snippet selection problem is posed as combinatorial subset selection with the goal of choosing an optimal set of snippets that maximizes a combination of relevancy, interestingness, and diversity. The relevancy and interestingness are scored by machine learned models. Our algorithm is run at scale on the entire image index of a major search engine resulting in the construction of a database of images with their corresponding text snippets. We validate the quality of the database through a large-scale comparative study. We showcase the utility of the database through two web-scale applications: (a) augmentation of images on the web as webpages are browsed and (b)~an image browsing experience (similar in spirit to web browsing) that is enabled by interconnecting semantically related images (which may not be visually related) through shared concepts in their corresponding text snippets.

J-3: Smart Reply: Automated Response Suggestion for Email
Kannan, A., Kurach, K., Ravi, S., Kaufmann, T., Tomkins, A., Miklos, B., ... & Ramavajjala, V. (2016). Smart reply: Automated response suggestion for email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* . ACM.

Abstract:
In this paper we propose and investigate a novel end-to-end method for automatically generating short email responses, called Smart Reply. It generates semantically diverse suggestions that can be used as complete email responses with just one tap on mobile. The system is currently used in *Inbox by Gmail* and is responsible for assisting with 10% of all mobile

responses. It is designed to work at very high throughput and process hundreds of millions of messages daily. The system exploits state-of-the-art, large-scale deep learning. We describe the architecture of the system as well as the challenges that we faced while building it, like response diversity and scalability. We also introduce a new method for semantic clustering of user-generated content that requires only a modest amount of explicitly labeled data.

K-1: A system to grade computer programming skills using machine learning

Srikant, S., & Aggarwal, V. (2014). A system to grade computer programming skills using machine learning. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* . ACM.

Abstract:

The automatic evaluation of computer programs is a nascent area of research with a potential for large-scale impact. Extant program assessment systems score mostly based on the number of test-cases passed, providing no insight into the competency of the programmer. In this paper, we present a system to grade computer programs automatically. In addition to grading a program on its programming practices and complexity, the key kernel of the system is a machine-learning based algorithm which determines closeness of the logic of the given program to a correct program. This algorithm uses a set of highly-informative features, derived from the abstract representations of a given program, that capture the program's functionality. These features are then used to learn a model to grade the programs, which are built against evaluations done by experts. We show that the regression models provide much better grading than the ubiquitous test-case-pass based grading and rivals the grading accuracy of other open-response problems such as essay grading . We also show that our novel features add significant value over and above basic keyword/expression count features. In addition to this, we propose a novel way of posing computer-program grading as a one-class modeling problem and report encouraging preliminary results. We show the value of the system through a case study in a real-world industrial deployment. To the best of the authors' knowledge, this is the first time a system using machine learning has been developed and used for grading programs. The work is timely with regard to the recent boom in Massively Online Open Courseware (MOOCs), which promises to produce a significant amount of hand-graded digitized data.

K-2: Top-k Multiclass SVM

Lapin, M., Hein, M., & Schiele, B. (2015). Top-k multiclass SVM. In *Advances in Neural Information Processing Systems*.

Abstract:

Class ambiguity is typical in image classification problems with a large number of classes. When classes are difficult to discriminate, it makes sense to allow k guesses and evaluate classifiers based on the top-k error instead of the standard zero-one loss. We propose top-k multiclass SVM as a direct method to optimize for top-k performance. Our generalization of the well-known multiclass SVM is based on a tight convex upper bound of the top-k error. We propose a fast optimization scheme based on an efficient projection onto the top-k simplex, which is of its own interest. Experiments on five datasets show consistent improvements in top-k accuracy compared to various baselines.

K-3: Robust Top-k Multi-class SVM for Visual Category Recognition

Chang, X., Yu, Y. L., & Yang, Y. (2017). Robust Top-k Multiclass SVM for Visual Category Recognition. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

Abstract:

Classification problems with a large number of classes inevitably involve overlapping or similar classes. In such cases it seems reasonable to allow the learning algorithm to make mistakes on similar classes, as long as the true class is still among the top-k (say) predictions. Likewise, in applications such as search engine or ad display, we are allowed to present k predictions at a time and the customer would be satisfied as long as her interested prediction is included. Inspired by the recent work of [15], we propose a very generic, robust multiclass SVM formulation that directly aims at minimizing a weighted and truncated combination of the ordered prediction scores. Our method includes many previous works as special cases. Computationally, using the Jordan decomposition Lemma we show how to rewrite our objective as the difference of two convex functions, based on which we develop an efficient algorithm that allows incorporating many popular regularizers (such as the l2 and l1 norms). We conduct extensive experiments on four real large-scale visual category recognition datasets, and obtain very promising performances.