Electronic Journal of Statistics
Vol. 4 (2010) 566–584
ISSN: 1935-7524
DOI: 10.1214/09-EJS522

Modeling temporal text streams using the local multinomial model

Guy Lebanon

College of Computing
Georgia Institute of Technology
Atlanta GA 30332 USA
e-mail: lebanon@cc.gatech.edu

Yang Zhao

Google
1600 Amphitheatre Parkway
Mountain View, CA 94043 USA
e-mail: zhyang@google.com

Yanjun Zhao

College of Computing
Georgia Institute of Technology
Atlanta GA 30332 USA
e-mail: zhaoyj@gmail.com

Abstract: Temporal text data such as news feeds cannot be adequately modeled by standard n-grams which correspond to multinomial or Markov chain models. Instead, we examine the application of local n-grams to modeling time stamped documents. We derive the asymptotic bias and variance and consider the bandwidth selection problem. Experimental results are presented on news feeds and web search query logs.

AMS 2000 subject classifications: Primary 62G99; secondary 62P99.
Keywords and phrases: Kernel smoothing, text modeling.

Received October 2009.

Contents

1 Introduction 567
2 The local multinomial model 569
3 Estimation and analysis 571
3.1 Bias-variance analysis of $\hat{\theta}_t$ 572
3.2 MSE, MISE and their dependence on the drift parameters 575
3.3 Bandwidth selection 577
A Description of datasets 582
A.1 RCV1 dataset 583
A.2 AOL dataset 583
Acknowledgements 583
References 583

1. Introduction

By far the most popular model for documents is the n-gram model. In the case of $n = 1$ it corresponds to the multinomial model where each word is drawn independently of the remaining words. In the case $n > 1$ it corresponds to a n-order Markov chain where the word transition probabilities $p(w_t|w_{t-n+1}, \ldots, w_{t-1})$ are estimated using empirical frequencies of string sequences. Naturally, the value of n reflects the bias-variance tradeoff. Increasing it enriches the model family but also increase the number of parameters and the difficulty in estimating them due to sparse counts. Since the dictionary size (the number of possible words in the language) is on the order of 10^4 or 10^5, values of n beyond 3 quickly become intractable. In practice, the case of $n = 1$ (unigram) is the most common with the cases of $n = 2$ (bigram) or $n = 3$ (trigram) trailing somewhat behind. Often, several n-gram models of different orders are combined as a mixture. For example, an interpolated trigram model is $p(x) = \alpha_1 p_1(x) + \alpha_2 p_2(x) + \alpha_3 p_3(x)$ where p_1, p_2, p_3 correspond to n-gram models for $n = 1, 2, 3$ respectively. Such mixtures provide outstanding modeling performance while remaining computationally tractable.

Despite their simplicity, n-grams enjoy widespread popularity as their computation scales up to the massive scale of data found on the internet and other large text archives. This popularity persists even as more complex models that require iterative estimation are investigated. Examples for specific applications in which n-gram models are used are web search [1], speech recognition [2], machine translation [7], and document classification [8]. We refer the interested reader to the references above or to standard textbooks such as [5] or [3] for more information on how n-grams are used in these applications.

Traditionally, the n-gram parameters are estimated from a document or a corpus of documents. The estimated model is used to predict probabilities associated with new arbitrary documents. This approach is inadequate for temporal document sequences where the n-gram parameters are assumed to change with the time documents are authored. For example, n-gram parameters corresponding to a news feed represent probabilities of obtaining a specific word conditioned on its context. These parameters change with time as world events are captured by the news, analyzed, and eventually forgotten. Another example is query logs in web-search where the n-gram parameters change with time as search users submit queries that reflect the time of day (day vs. night), day of week (weekend vs. weekday) and transient topics of interest.

We experiment in this paper using two temporal text datasets: the Reuters RCV1 dataset [4] and the AOL dataset [6]. The Reuters RCV1 dataset contains news stories authored during a period of 365 consecutive days by Reuters journalists. The AOL dataset contains queries issued by AOL users during a period of three months. More details regarding these datasets are provided in the appendix.

The upper part of Figure 1 displays the temporal change in the relative frequency of three words (number of word appearance in a document divided by

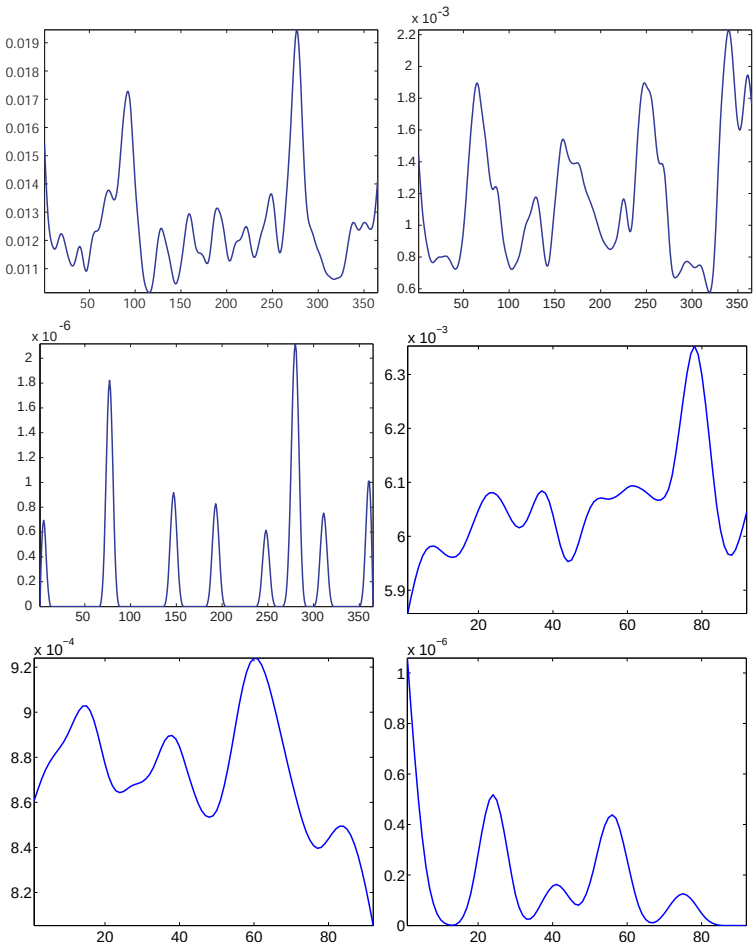

FIG 1. *Estimated relative frequency (number of appearances in a document divided by document length) of words from the most popular category in RCV1 and AOL as a function of time. The upper three panels correspond to the words* `million`, `common`, *and* `Handelsgesellschaft` *(trade unions in German) in RCV1 dataset. The lower three panels correspond to the words* `free`, `lottery`, *and* `evansville.net` *in AOL dataset.*

document length) in RCV1 dataset: `million`, `common`, and `Handelsgesellschaft` (trade unions in German) for documents in the most popular RCV1 category titled `CCAT`. It is obvious from these plots that the relative frequency of these words vary substantially in time. For example, the word `Handelsgesellschaft` appear in 8 distinct time regions, representing time points in which German trade unions were featured in the Reuters news archive. The lower part of Figure 1 displays the temporal change in the relative frequency of three words: `free`, `lottery`, and `evansville.net` in AOL dataset. Similar to RCV1 dataset, the relative frequency of these words vary substantially with time. However, due to

the short length of web queries, even the most frequent words in the AOL data such as `free` are still sparse compared to RCV1 data.

The daily relative frequencies plotted in Figure 1 correspond to the mle assuming independent draws from a multinomial distribution or n-gram with $n = 1$. We focus on this simple model as it is both very popular due to its simplicity and scalability and the fact that n-grams with $n > 1$ are straightforward generalization of the $n = 1$ case. We thus see from Figure 1 that the multinomial parameters differ substantially from day to day in both the RCV1 and AOL datasets. Attempting to use a multinomial to model the entire dataset regardless of the time documents are authored is inadequate. Similarly, attempting to estimate a single multinomial for each day using only documents from that day is suboptimal as accurate estimation is possible only for the dates in which many documents were authored. In particular, documents authored at time t are ignored when estimating the multinomial corresponding to time $t+1$. This motivates the kernel smoothing approach of the local multinomial model, which we investigate in this paper.

2. The local multinomial model

Based on the variability in Figure 1 we assume that documents authored at time t were generated by a multinomial with parameter θ_t. The main problem we are interested in is estimating the collection of multinomial parameters $\{\theta_t : t \in [a, b]\}$ where $[a, b]$ is the range of time values under consideration. It is natural to assume that the multinomial parameter generating the documents at time t is related to the multinomial parameter corresponding to time $t + \epsilon$ for small $\epsilon > 0$. In other words, we assume that the mapping $t \mapsto \theta_t$ draws a smooth curve in the simplex of multinomial parameters

$$\mathbb{P}_S \stackrel{\text{def}}{=} \left\{ q \in \mathbb{R}^{|S|} \; : \; \forall i \, q_i \geq 0, \; \sum_{i=1}^{|S|} q_i = 1 \right\}. \tag{1}$$

Above, S is the dictionary or the set of all possible words and q_i is the probability of drawing the i-word in the vocabulary.

More formally, we assume the model

$$t \sim g(t) \tag{2}$$

$$l \sim \text{Pois}(\lambda) \tag{3}$$

$$(w_1, \ldots, w_l) \sim \text{Mult}(l, \theta_t). \tag{4}$$

where $g(t)$ is the distribution of times in which documents are authored, $\text{Pois}(\lambda)$ is the distribution of the number of words in documents and $\{\theta_t, t \in [a, b]\}$ is a smooth curve in the simplex (1).

We display in Figure 2 the total number of words per day (left) and the total number of documents per day (right) for the RCV1 (top) and AOL (bottom)

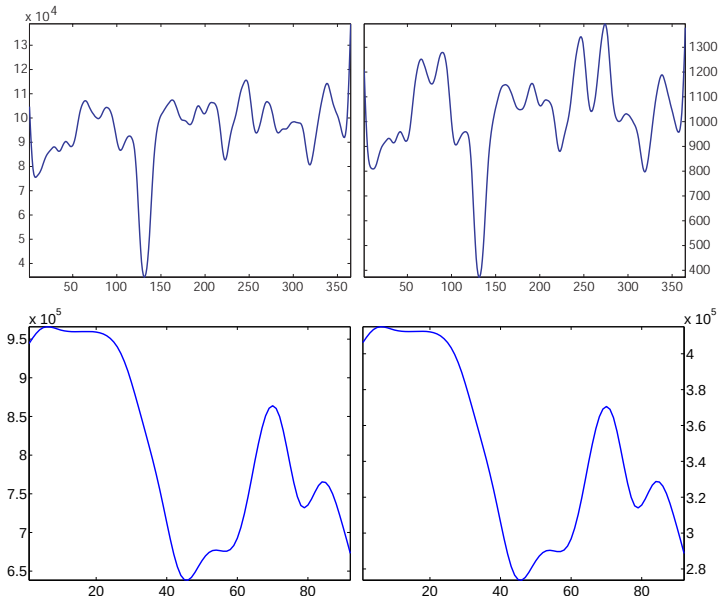

FIG 2. *The total number of words per day (left) and the total number of documents per day (right) for the most popular RCV1 class (top) and AOL (bottom) datasets. As is evident from the two right panels, $g(t)$ is a highly non-uniform density corresponding to varying amount of news content and queries in different dates. Dividing the number of words per day (left) by the number of documents per day (right) we obtain surprisingly little variation among the number of words per document. We thus conclude that documents tend to have similar lengths but the number of documents per day vary substantially.*

datasets. As is evident from the two right panels, $g(t)$ is a highly non-uniform density corresponding to varying amount of news content and queries in different dates. This high variability in $g(t)$ can be explained by the fact that some days have more news stories than other days. As we see later this variability in $g(t)$ has a direct impact on the asymptotic mse of the estimator-high variability increases the difficulty of the estimation task and consequentially increases the asymptotic mse.

Dividing the number of words per day (left) by the number of documents per day (right) we obtain surprisingly little variation among the number of words per document. We thus conclude that while the $g(t)$ vary considerably across t, the distribution of document lengths f is independent of t justifying the assumption in (3) that the document length distribution is not a function of t.

In practice, due to the discretization of time we may have multiple time points $t_1, \ldots, t_r \in [a, b]$ with N_{t_i} documents authored at time t_i. We denote the documents themselves as $x^{(t_i,1)}, \ldots, x^{(t_i,N_{t_i})}$ and use $c(x^{(t_i,j)}, w)$ to represent the number of times word w appeared in document $x^{(t_i,j)}$.

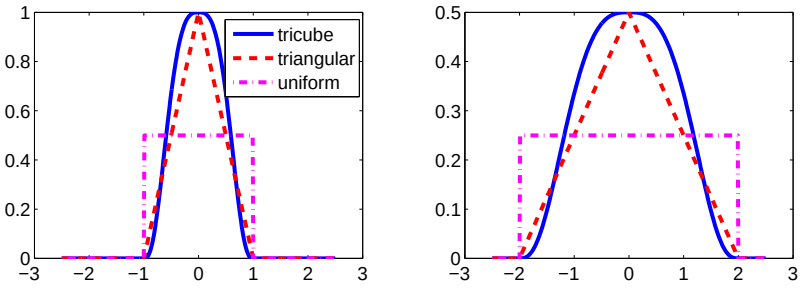

FIG 3. *Tricube, triangular, and uniform kernels with scale $h = 1$ (left) and $h = 2$ (right).*

3. Estimation and analysis

We estimate the multinomial parameter at time t by maximizing the local log-likelihood

$$\hat{\theta}_t = \arg\max_{\theta \in \Theta} \ell_t(\theta|D)$$

$$\ell_t(\theta|D) \stackrel{\text{def}}{=} \sum_{\tau} K_h(t-\tau) \sum_{j=1}^{N_\tau} \log p(x^{(\tau,j)}\,;\theta). \tag{5}$$

The function $K_h : \mathbb{R} \to \mathbb{R}$ is a smoothing kernel that is parameterized by a scale or bandwidth parameter $h > 0$ satisfying

$$K_h(r) = h^{-1} K_1(r/h)$$

where we denote $K = K_1$ and refer to it as the kernel base form. We also assume that it is a normalized distribution, that K has bounded support, and that $\int u^r K(u)\, du < \infty$ for $r \leq 2$.

Three popular kernel choices are the tricube, triangular and uniform kernels, defined as $K_h(r) = h^{-1} K(r/h)$ where the $K(\cdot)$ functions are respectively

$$K(r) = (1 - |r|^3)^3 \cdot 1_{\{|r|<1\}} \tag{6}$$

$$K(r) = (1 - |r|) \cdot 1_{\{|r|<1\}} \tag{7}$$

$$K(r) = 2^{-1} \cdot 1_{\{|r|<1\}}. \tag{8}$$

Figure 3 displays these kernels for $h = 1$ (left) and $h = 2$ (right). The uniform kernel is the simplest choice and leads to a local likelihood (5) equivalent to filtering the data by a sliding window i.e. $\hat{\theta}_t$ is computed based on data from adjacent time points with uniform weights. However, it is suboptimal in terms of its statistical efficiency or rate of convergence to the underlying distribution. In our experiments we used the triangular and tricube kernels.

The local likelihood model has a single global maximum whose closed form expression may be found by setting to 0 the gradient of the Lagrangian. Below, we denote the length of a document $x^{(t,j)}$ by $|x^{(t,j)}| \stackrel{\text{def}}{=} \sum_{v\in V} c(x^{(t,j)}, v)$, and the total number of words in day t by $|x^{(t)}| \stackrel{\text{def}}{=} \sum_{j=1}^{N_t} |x^{(t,j)}| = \sum_{v\in V} \sum_{j=1}^{N_t} c(v, x^{(t,j)})$ (recall that the number of words of type $w \in S$ in $x^{(t,j)}$ is denoted by $c(w, x^{(t,j)})$,). Using these notations the local likelihood for the multinomial becomes

$$\ell_t(\theta|D) = \sum_{\tau} K_h(t-\tau) \sum_{j=1}^{N_\tau} \sum_{w\in V} c(w, x^{(\tau,j)}) \log \theta_w, \quad \theta \in \mathbb{P}_S. \tag{9}$$

Setting the gradient of the Lagrangian to 0

$$0 = \frac{1}{[\hat{\theta}_t]_w} \sum_{\tau} K_h(t-\tau) \sum_{j=1}^{N_\tau} c(w, x^{(\tau,j)}) + \lambda_w$$

we obtain the local likelihood maximizer

$$[\hat{\theta}_t]_w = \frac{\sum_\tau K_h(t-\tau) \sum_{j=1}^{N_\tau} c(w, x^{(\tau,j)})}{\sum_\tau K_h(t-\tau)|x^{(\tau)}|}. \tag{10}$$

The estimator $\hat{\theta}_t$ is a normalized linear combination of word counts where the combination coefficients are determined by the kernel function and normalized by the number of words in different days. We note that $\hat{\theta}_t$ in (10) is different from a weighted averaging of the relative frequencies $c(w, x^{(\tau,j)})/\sum_{w'} c(w', x^{(\tau,j)})$.

We distinguish between two fundamental estimation scenarios.

Offline scenario: The goal is to estimate $\{\theta_t : t \in [a, b]\}$ given the entire dataset. In this case we will consider symmetric kernels $K(r) = K(-r)$ which will achieve an increased convergence rate of $\hat{\theta}_t \to \theta_t$ as indicated by Proposition 2.

Online scenario: The goal is estimate a model for θ_t where t represent the present using training data from the past i.e. a dataset whose time stamps are strictly smaller than t. This corresponds to situations where the data arrives sequentially as a temporal stream and at each time point a model for the present is estimated using the available stream at that time. We realize this restriction by constraining K to satisfy $K(r) = 0, r \leq 0$. As a result the local likelihood at time t incorporates documents written at times less than or equal to t.

3.1. Bias-variance analysis of $\hat{\theta}_t$

As with other statistical estimators, the accuracy of $\hat{\theta}_t$ may be measured in terms of its bias and variance.

Proposition 1. *The bias vector* $\mathsf{bias}(\hat{\theta}_t) \stackrel{\text{def}}{=} \mathsf{E}\hat{\theta}_t - \theta_t$ *and variance matrix of* $\hat{\theta}_t$ *in* (10) *are*

$$\mathsf{bias}(\hat{\theta}_t) = \frac{\sum_\tau K_h(t-\tau)|x^{(\tau)}|\,(\theta_\tau - \theta_t)}{\sum_\tau K_h(t-\tau)|x^{(\tau)}|} \tag{11}$$

$$\mathsf{Var}(\hat{\theta}_t) = \frac{\sum_\tau K_h^2(t-\tau)|x^{(\tau)}|\,(diag(\theta_\tau) - \theta_\tau\theta_\tau^\top)}{\left(\sum_\tau K_h(t-\tau)|x^{(\tau)}|\right)^2} \tag{12}$$

where $diag(z)$ *is the diagonal matrix* $[diag(z)]_{ij} = \delta_{ij} z_i$.

Proof. The random variable (RV) $c(w, x^{(\tau,j)})$ is distributed as a sum of multivariate Bernoulli RVs, or single draws from multinomial distribution. The expectation and variance of the estimator are that of a linear combination of iid multinomial RVs. To conclude the proof we note that for $Y \sim \text{Mult}(1,\theta)$, $\mathsf{E}\,Y = \theta$, $\mathsf{Var}(\theta) = \text{diag}(\theta) - \theta\theta^\top$. □

Examining Equations (11)–(12) reveals the expected dependency of the bias on h and θ_t. The contribution to the bias of the terms $(\theta_\tau - \theta_t)$, for large $|\tau - t|$, will decrease as h decreases since the kernel becomes more localized and will reduce to 0 as $h \to 0$. Similarly, for more more slowly changing parameter curve $\{\theta_t : t \in [a,b]\}$, $\|\theta_\tau - \theta_t\|, t \approx \tau$ will decrease and reduce the bias.

Despite the relative simplicity of Equations (11)–(12), it is difficult to quantitatively capture the relationship between the bias and variance, the sample size, h, λ, and the smoothness of θ_t, g. Towards this goal we derive the following asymptotic expansions.

Proposition 2. *Assuming (i)* θ, g *are smooth in* t*, (ii)* $h \to 0, hn \to \infty$*, (iii)* $g > 0$ *in a neighborhood of* t*, and (iv) document lengths do not depend on* t *and have expectation* λ*, the bias vector and variance matrix are in the offline case*

$$\mathsf{bias}(\hat{\theta}_t) = h^2\mu_{21}(K)\left(\dot{\theta}_t\frac{g'(t)}{g(t)} + \frac{1}{2}\ddot{\theta}_t\right) + o_P(h^2) \tag{13}$$

$$\mathsf{Var}(\hat{\theta}_t) = \frac{\mu_{02}(K)}{(nh)g(t)\lambda}(diag(\theta_t) - \theta_t\theta_t^\top) + o_P((nh)^{-1})$$

and in the online case

$$\mathsf{bias}(\hat{\theta}_t) = h\mu_{11}(K)\dot{\theta}_t + o_P(h) \tag{14}$$

$$\mathsf{Var}(\hat{\theta}_t) = \left(\frac{\mu_{02}(K)}{nhg(t)\lambda} + \frac{\mu_{12}(K)g'(t)}{ng^2(t)\lambda}\right)(diag(\theta_t) - \theta_t\theta_t^\top) \tag{15}$$
$$+ \frac{\mu_{12}(K)}{n\lambda g(t)}(diag(\dot{\theta}_t) - \dot{\theta}_t\theta_t^\top - \theta_t\dot{\theta}^\top) + o_P((nh)^{-1})$$

where $\dot{\theta}_t$ *is the vector* $[\dot{\theta}_t]_i = \frac{d}{dt}[\theta_t]_i$ *and*

$$\mu_{kl}(K) \stackrel{\text{def}}{=} \int t^k K^l(t)\,dt < \infty \qquad 0 \le k, l \le 2.$$

Proof. The proof follows standard expansions similar to the ones used in studying local polynomial regression but modified to our setting. We start by expanding the numerator and denominator of the bias and variance in the offline case. Our main tools are the law of large numbers, changing the integration variable, and Taylor series expansion. For notational simplicity we assume below that $t = 0$. The arguments below may be modified at some notational expense for $t \neq 0$ to produce Equations (13)–(14). In the proof below we use slightly different notation with x_{τ_i} representing the i-document authored at time τ_i.

We expand the denominator and numerator of the bias (11) multiplied by $1/n$:

$$\frac{1}{n}\sum_{i=1}^{n} K_h(\tau_i)|x_{\tau_i}| \xrightarrow{\text{p}} \lambda\int g(t)K_h(t)\,dt = \lambda h^{-1}\int g(t)K(t/h)\,dt = \lambda\int g(uh)K(u)\,du$$
$$= \lambda\int K(u)(g(0)+o(1))\,du = \lambda g(0) + o(1).$$

$$\frac{1}{n}\sum_{i=1}^{n} K_h(\tau_i)|x_{\tau_i}|(\theta_{\tau_i}-\theta_0) \xrightarrow{\text{p}} \lambda h^{-1}\int g(t)(\theta_t-\theta_0)K(t/h)\,dt$$
$$= \lambda\int g(uh)(\theta_{uh}-\theta_0)K(u)\,du$$
$$= \lambda\int (g(0)+g'(0)uh+g''(0)u^2h^2/2+o(u^2h^2))$$
$$\times(\dot\theta_0 uh+\ddot\theta_0 u^2h^2/2+o(u^2h^2))K(u)du$$
$$= \lambda h^2\mu_{21}(K)\left(g'(0)\dot\theta_0+\frac{1}{2}g(0)\ddot\theta_0\right)+o(h^2).$$

Above, we used the offline assumption by exploiting the symmetry of the kernel to deduce $\int K(u)u\,du = 0$. Dividing the two expansions and replacing $o(h^2)$ with $o_P(h^2)$ due to the law of large numbers approximation establishes (13).

Similarly we expand the denominator and numerator of the variance matrix times $1/n^2$ and $1/n$ respectively

$$\left(\frac{1}{n}\sum_{i=1}^{n} K_h(\tau_i)|x_{\tau_i}|\right)^2 \xrightarrow{\text{p}} \left(\lambda\int K(u)(g(u)+o(1))\,du\right)^2 = \lambda^2 g^2(0)+o(1))^2$$

$$\frac{1}{n}\sum_{i=1}^{n} K_h^2(\tau_i)|x_{\tau_i}|\mathsf{Var}\,(\theta_{\tau_j}) \xrightarrow{\text{p}} \lambda h^{-2}\int K^2(t/h)g(t)\mathsf{Var}\,(\theta_t)\,dt$$
$$= \lambda h^{-1}\int K^2(u)g(uh)\mathsf{Var}\,(\theta_{uh})\,du$$
$$= \lambda h^{-1}\int K^2(u)(g(0)+g'(0)uh+o(uh))$$
$$\times(\mathsf{Var}\,(\theta_0)+\dot{\mathsf{Var}}\,(\theta_0)uh+o(uh))\,du$$
$$= \lambda h^{-1}g(0)\mathsf{Var}\,(\theta_0)\mu_{02}(K)+o(h)$$

where again we used the kernel symmetry to deduce $\int K^2(u)u\,du = 0$. Since $\mathsf{Var}\,(\theta_t) = (\mathrm{diag}(\theta_t) - \theta_t\theta_t^\top)$, dividing the second expansion by the first and dividing by n^{-1} provides the desired result.

In the online setting, the kernel is no longer symmetric and $\int K(u)u\,du \neq 0$ which lowers the rate of convergence. The expansions of the numerator of the bias and variance are

$$\frac{1}{n}\sum_{i=1}^{n} K_h(\tau_i)|x_{\tau_i}|(\theta_{\tau_i} - \theta_0) \xrightarrow{\mathrm{p}} \lambda \int (g(0) + g'(0)uh + o(uh))(\dot\theta_0 uh + o(uh))K(u)du$$
$$= \lambda h \mu_{11}(K)\dot\theta_0 g(0) + o(h).$$

$$\frac{1}{n}\sum_{i=1}^{n} K_h^2(\tau_i)|x_{\tau_i}|\mathsf{Var}\,(\theta_{\tau_j}) \xrightarrow{\mathrm{p}} \frac{\lambda}{h}\int K^2(u)(g(0) + g'(0)uh + o(uh))$$
$$\times (\mathsf{Var}\,(\theta_0) + \dot{\mathsf{Var}}\,(\theta_0)uh + o(uh))\,du$$
$$= \frac{\lambda}{h} g(0)\mathsf{Var}\,(\theta_0)\mu_{02}(K) + \lambda\mu_{12}(K)$$
$$\times (g(0)\dot{\mathsf{Var}}\,(\theta_0) + g'(0)\mathsf{Var}\,(\theta_0)) + o(h).$$

Noticing that $\dot{\mathsf{Var}}\,(\theta_t) = \mathrm{diag}(\dot\theta_t) - \dot\theta_t\theta_t^\top - \theta_t\dot\theta^\top$ concludes the proof. □

3.2. MSE, MISE and their dependence on the drift parameters

We have the following direct corollary of Proposition 2.

Corollary 1. *Under the assumptions in Proposition 2, the component-wise mean squared error* $\mathsf{mse}([\hat\theta_t]_i) = \mathsf{E}([\hat\theta_t]_i - [\theta_t]_i)^2$, $i = 1,\ldots,|S|$ *are for the offline case*

$$\mathsf{mse}([\hat\theta_t]_i) = h^4\mu_{21}^2(K)\left([\dot\theta_t]_i\frac{g'(t)}{g(t)} + \frac{1}{2}[\ddot\theta_t]_i\right)^2$$
$$+ \frac{\mu_{02}(K)}{nhg(t)\lambda}[\theta_t]_i(1 - [\theta_t]_i) + o_P(h^4 + (nh)^{-1}).$$

and for the online case

$$\mathsf{mse}([\hat\theta_t]_i) = h^2\mu_{11}^2(K)[\dot\theta_t]_i^2 + \left(\frac{\mu_{02}(K)}{nhg(t)\lambda} + \frac{\mu_{12}(K)g'(t)}{ng^2(t)\lambda}\right)[\theta_t]_i(1 - [\theta_t]_i)$$
$$+ \frac{\mu_{12}(K)}{n\lambda g(t)}[\dot\theta_t]_i(1 - 2[\theta_t]_i) + o_P(h^2 + (nh)^{-1}).$$

Corollary 2. *Under the assumptions in Proposition 2, and in particular* $h \to 0, nh \to \infty$, *the estimator* $\hat\theta_t$ *is consistent i.e.* $\hat\theta_t \xrightarrow{\mathrm{p}} \theta_t$ *in both the offline and online settings.*

Proof. The proof follows from the fact that under these conditions we have convergence in the second moment of the components of $\hat{\theta}_t - \theta_t$ to 0. □

Proposition 2 is important as it specifies the conditions for consistency as well as the rate of convergence. We make the following comments on it.

1. The conditions specified in Proposition 2 for consistency of the estimator (in particular $h \to 0, nh \to \infty$) are standard conditions in non-parametric kernel smoothing and are similar to those of other related estimators such as the kernel density estimator and the Nadaraya-Watson local regression estimator.
2. The rates of convergence indicated by the argument of $o_P(\cdot)$ in Proposition 2 are important as they quantify the rates at which the estimators converges to the underlying drift. In particular, it is interesting to note the fact that the bias of online kernels converges at a linear rather than the quadratic rate of the offline kernels. This is a quantification of the fact that looking at the past and future helps predict the present more than looking only at the past.
3. Using corollary 1 and (13)–(14) we can analyze the dependency of the bias, variance, and MSE on $\{\theta_t : t \in [a, b]\}$, the parameter curve speed indicated by $\dot{\theta}_i$, the rate of change of the log sampling density $d \log g(t) = g'(t)/g(t)$, the number of documents n, and the expected length of the documents λ. Intuitively, the estimation task is easier if the drift is slower ($\dot{\theta}_t$ is smaller), the time sampling variation ($d \log g(t)/dt$) is smaller, and there are more and longer documents. Using expressions (13)–(14) we confirm these intuitive observations and quantity them: the bias is reduced as the drift speed and time sampling variation are lower while the variance is reduced as we have more (denoted by n) and longer (denoted by λ) documents.
4. Corollary 1 and expressions (13)–(14) also reveal somewhat less intuitive insights. First, the variance grows linearly with $[\theta_t]_i(1 - [\theta_t]_i)$ (in the offline case; the online variance is slightly more complicated). In the case of documents, the word probabilities θ_i are typically very small and thus the larger they are the higher the $[\theta_t]_i(1 - [\theta_t]_i)$ factor is in the asymptotic variance (see Figure 4, left). Note also that as the inverse Fisher information of the binomial $[\theta_t]_i(1 - [\theta_t]_i)$ bounds the variance of the optimal estimator.
5. Another interesting observation is that in the case of slowly varying $g(t)$ the factor $g'(t)/g(t)$ tends to be very small (see Figure 4 (right)) making the first offline bias term negligible and exposing a linear trend of the bias in $\ddot{\theta}_t$ independently of $\dot{\theta}_t$. This indicates zero (or nearly zero) offline bias for linear drift as its second derivative is zero. On the other hand, when $g(t)$ is rapidly varying the bias term (as well as the variance) exhibit more complex behavior.

The above proposition and corollary are expressed in terms of the mean squared error at a particular time point t. This is suitable in cases where we are

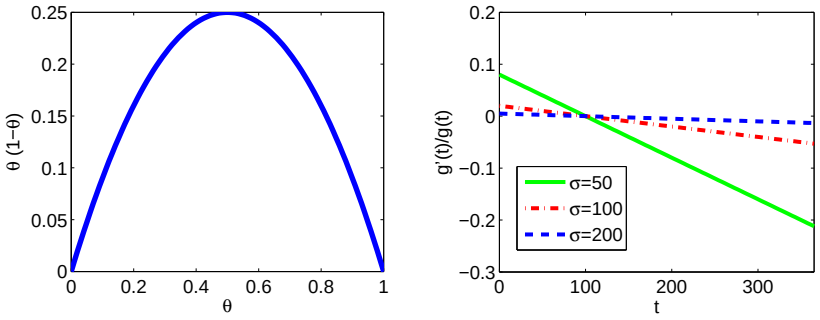

FIG 4. *Left: The variance increases linearly with* $[\theta_t]_i(1-[\theta_t])_i$ *which is monotonically increasing in* $[\theta_t]_i$ *for small values such as word probabilities. Right: The* $g'(t)/g(t)$ *factor associated with the first bias term is likely to be negligible for slow changing* $g(t)$ *compared to the second term making the bias increase linearly with* $\ddot{\theta}_t$ *and independent of* $\dot{\theta}_t$ *(indicating zero bias for linear drift). The figure plots* $g'(t)/g(t)$ *for* $t \in [1, 365]$ *and* $g(t) = N(100, \sigma)$ *with* $\sigma = 50, 100, 200$.

interested in estimation accuracy at a specific time point such as the present time. In other cases, more insightful criteria are the integrated squared error (ise) and the mean integrated square error (mise) which average the estimation error over t

$$\text{ise}([\hat{\theta}]_i) = \int ([\hat{\theta}_t]_i - [\theta_t]_i)^2 dt \tag{16}$$

$$\text{mise}([\hat{\theta}]_i) = \mathsf{E} \int ([\hat{\theta}_t]_i - [\theta_t]_i)^2 dt = \int \mathsf{mse}\,(\hat{\theta}_t) dt. \tag{17}$$

Corollary 3. *Under the assumptions in Proposition 2, we have in the offline case*

$$mise([\hat{\theta}_t]_i) = h^4 \mu_{21}^2(K)\mu_{02} \left([\dot{\theta}_t]_i \frac{g'(t)}{g(t)} + \frac{1}{2}[\ddot{\theta}_t]_i \right) + \frac{\mu_{02}(K)}{nh\lambda} \mu_{01} \left(\frac{[\theta_t]_i(1-[\theta_t]_i)}{g(t)} \right) + o_P(h^4 + (nh)^{-1}).$$

A similar expansion for the online case is straightforward. Under these assumptions and in particular $h \to 0, nh \to \infty$ *the total mise* $\sum_w mise([\hat{\theta}_t]_w)$ *converges to 0 in both the online and offline scenarios.*

3.3. Bandwidth selection

A central issue in local likelihood modeling and non-parametric estimation in general is selecting the appropriate bandwidth h. Such a selection is critical to effective modeling and is the subject of substantial research. Figure 5 displays the RCV1 test set loglikelihood for the online and offline scenarios as a function

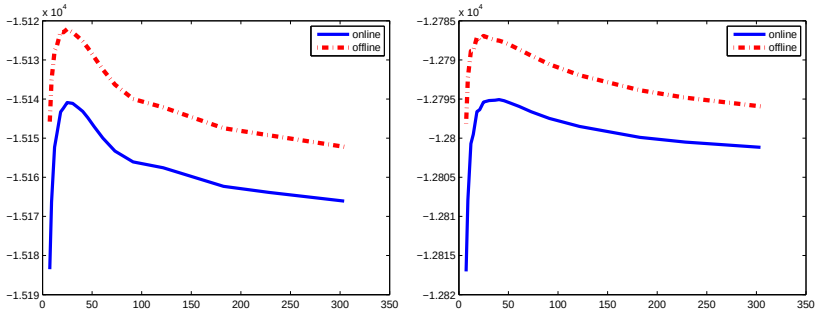

FIG 5. *Log-likelihood of held out test set as a function of the triangular kernel's bandwidth for the two largest RCV1 categories (CCAT (left) and GCAT (right)) and the most frequent 500 words. Training set size was 100 documents per day and test set performance was averaged over repeated sampling to remove noise. In all four cases, the optimal bandwidth seems to be approximately 25 which indicates a support of 25 days for the online kernels and 50 days for the offline kernels.*

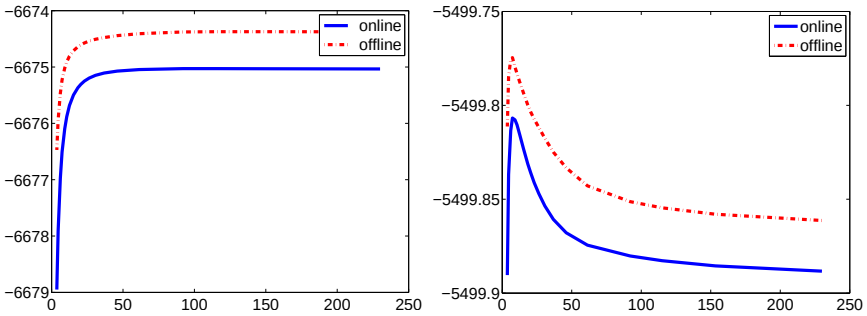

FIG 6. *Log-likelihood of held out test set as a function of the triangular kernel's bandwidth for the AOL dataset with 20000 documents per day (left) and 80000 docs per day (right)). Due to the short document length of queries, 20000 documents per day are not sufficient to motivate a non-global estimator. In the case of 80000 documents per day, the optimal bandwidth in both cases is around 7 which indicates a support of 7 days for the online kernel and 14 days for the offline kernel.*

of the (triangular) kernel's bandwidth. As expected, offline kernels performs better than online kernels with both achieving the best performance for a bandwidth approximately 25 which corresponds to a support of 25 days in the online scenario and 50 days in the offline scenario. Similar results are displayed in Figure 6 for the AOL dataset where the optimal bandwidth is 7 indicating offline support of 14 days. Note that in addition to obtaining higher accuracy than the global model corresponding to $h \to \infty$, the local model enjoys computational efficiency as it ignores a large portion of the training data.

Perhaps the most obvious technique for selecting h is maximum likelihood cross validation (MLCV). In this technique, the dataset D is randomly parti-

tioned to two subsets $D = D_A \cup D_B$ - one D_A used to construct a local likelihood estimator $\hat{\theta}_t^{(D_A)}$ and one used to evaluate the loglikelihood of the estimator

$$\ell^{(D_A)}(D_B) = \sum_{i \in D_B} \log p_{\hat{\theta}_{t_i}^{(D_A)}}(x_i).$$

Ten-fold cross validation averages this process ten times where each time 90% of the data is kept for constructing $\hat{\theta}^*$ and 10% of the data is used to evaluate the log-likelihood of $\hat{\theta}^*$. The MLCV estimator then proceeds to select the bandwidth h that maximizes the ten-fold cross validation function

$$h_{\text{MLCV}} = \arg\max_{h>0} \sum_{j=1}^{10} \ell^{(D_{A_j})}(D_{B_j}).$$

On RCV1 data, the performance of such cross validation schemes is extremely good and the estimated bandwidth possesses test set loglikelihood that is almost identical to the optimal bandwidth (see Figure 7, top). Allowing the kernel scale to vary over time results in a higher modeling accuracy than using fixed bandwidth for all dates (see Figure 7, bottom). A time-dependent cross validation procedure may be used to approximate the time-dependent optimal bandwidth which performs slightly better than the time independent cross validation estimator. Note that the accuracy with which the cross validation estimator approximates the optimal bandwidth is lower for the time-dependent bandwidth due the fact that much less data is available in each of the daily cross validation problems.

An alternative to MLCV is least squares cross validation (LSCV) which is based on the following decomposition of the mise (17)

$$\text{mise}([\hat{\theta}]_i) = \mathsf{E} \int [\hat{\theta}_t]_i^2 \, dt - 2\mathsf{E} \int [\hat{\theta}_t]_i [\theta_t]_i \, dt + \int [\theta_t]_i^2 \, dt \tag{18}$$

and noting that the third term does not depend on h. We can thus construct an unbiased estimator for $\text{mise}([\hat{\theta}]_i) - \int [\theta_t]_i^2 \, dt$ as follows

$$\text{LSCV}(h, i) = \int [\hat{\theta}_t]_i^2 \, dt - 2n^{-1} \sum_{j=1}^{n} [\hat{\theta}_t^{(-j)}]_i \tag{19}$$

where $\sum_{j=1}^{n} [\hat{\theta}_t^{(-j)}]_i$ is the local likelihood estimator using the dataset D but omitting the i-observation. Assuming we are interesting in minimizing the mise over all the parameters of $\hat{\theta}$ we obtain

$$\hat{h}_{\text{LSCV}} = \arg\min_{h>0} \sum_{i=1}^{V} \text{LSCV}(h, i) \tag{20}$$

which is an unbiased estimator of the minimizer of the total mise that is $\arg\min_{h>0} \sum_{i=1}^{V} \text{mise}([\hat{\theta}]_i)$.

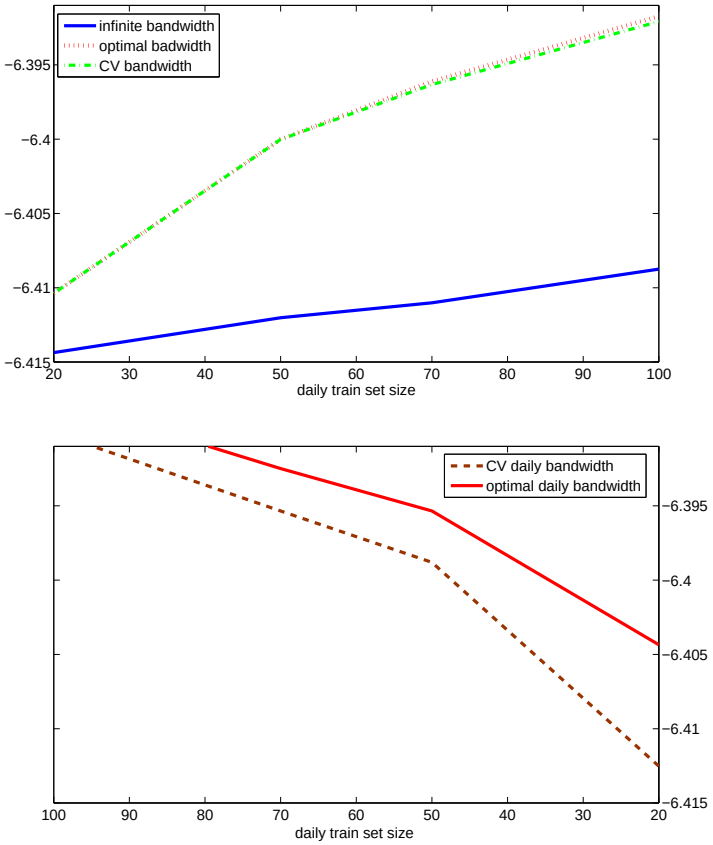

FIG 7. *Per-word log-likelihood over held-out test set for various bandwidths as a function of the daily training set size. Top: The extreme global model corresponding to $h \to \infty$ performs worst. Selecting the bandwidth by cross validation results in an accurate estimate and test-set loglikelihood almost identical to that of the optimal slope. Bottom: Allowing the kernel scale to vary over time results in a higher modeling accuracy than using fixed bandwidth for all dates.*

From a theoretical perspective, we can get additional insights by analytically minimizing the leading terms of the mse or mise as a function of h. The resulting minimizer expresses in closed form the dependency of the optimal bandwidth on the problem parameters $n, \lambda, \dot{\theta}, \ddot{\theta}, g(t), \theta_t$. For example minimizing the leading term of $\sum_{j=1}^{V} \mathsf{mse}\left([\hat{\theta}_t]_i\right)$ we obtain

$$\hat{h}_t^5 = \frac{\mu_{02}(K)\mathrm{tr}(\mathrm{diag}(\theta_t) - \theta_t\theta_t^\top)}{4n\lambda\mu_{21}^2(K)\sum_j \left([\dot{\theta}_t]_j\, g'(t)/\sqrt{g(t)} + \sqrt{g(t)}[\ddot{\theta}_t]_j\,/2\right)^2}. \tag{21}$$

Proposition 3. *Under the assumptions in Proposition 2 in the offline case we have*

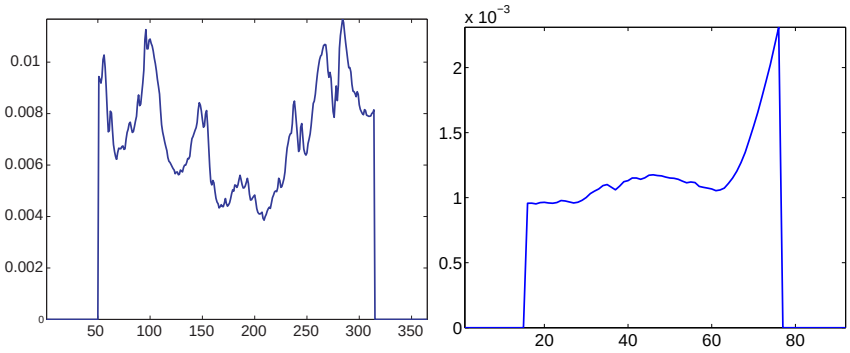

FIG 8. *Estimated gradient norm for the most popular category in RCV1 (left) and AOL (right) as a function of t. The derivatives were estimated using local smoothing. To avoid running into boundary effects we ignore the first and last 50 days in RCV1 and 15 days in AOL.*

$$\inf_{h>0} \mathit{mse}\left([\hat{\theta}_t]_i\right) =$$
$$n^{-4/5}\left\{\mu_{21}^2(K)\left([\dot{\theta}_t]_i\frac{g'(t)}{g(t)}+\frac{1}{2}[\ddot{\theta}_t]_i\right)^2+\frac{\mu_{02}(K)}{g(t)\lambda}[\theta_t]_i(1-[\theta_t]_i)+o_P(1)\right\}. \quad (22)$$

Proof. Equation (22) follows from minimizing the mse as a function of h and substituting the resulting minimizer back in the mse. □

As expected, the optimal bandwidth decreases as $n, \lambda, \|\dot{\theta}_t\|, \|\ddot{\theta}\|$ increases. Intuitively this makes sense since in these cases the variance decreases and bias either increases or stays constant. In practice, $\dot{\theta}_t, \ddot{\theta}_t$ may vary significantly with time which leads to the conclusion that a single bandwidth selection for all t may not perform adequately. These changes are illustrated in Figure 8 which demonstrates the temporal change in the gradient norm.

A more surprising result is the non-monotonic dependency of the optimal bandwidth on the time sampling distribution $g(t)$. The dependency, expressed by

$$\hat{h}_t \propto \left(\sum_{j=1}^{V}(c_{1j}/\sqrt{g(t)}+c_{2j}\sqrt{g(t)})^2\right)^{-1/5}$$

is illustrated in Figure 9 (left) where we assume for simplicity that c_{1j}, c_{2j} do not change with j resulting in

$$(\hat{h}_t)^{-1} \propto (c_1/g(t)+c_2 g(t)+c_3)^{1/5}.$$

The key to understanding this relationship is the increased asymptotic bias due to the presence of the term $g'(t)/g(t)$ in Equation (13). Indeed, plotting the inverse of the optimal bandwidth (we actually average that quantity over the word-specific optimal bandwidths for different words) for the RCV1 data as a function of the daily word count (which is proportional to $g(t)$) in Figure 9

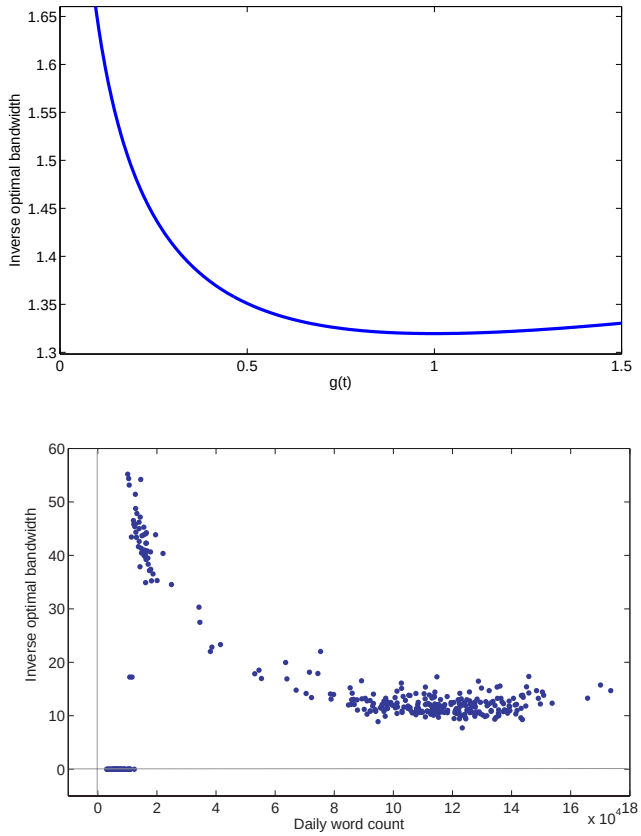

FIG 9. *Left: Inverse of the optimal bandwidth derived from Equation (21) as a function of* $g(t)$*:* $(\hat{h}_t)^{-1} \propto (c_1/\sqrt{g(t)} + c_2\sqrt{g(t)})^{2/5}$ *(we take* $c_1 = c_2 = 1$*). The graph show the non-monotonic dependency between* $\hat{h}^{opt}$ *and* $g(t)$*. Right: Inverse of the optimal bandwidth* $(\hat{h}_t)^{-1}$ *(averaged over the optimal bandwidths for the top 3000 words) as a function of the daily word count (which is proportional to* $g(t)$*). The two graphs show an interesting correspondence between theory and practice and illustrate the non-monotonic dependency between the optimal bandwidth and* $g(t)$*.*

(right) reveals a trend similar to the theoretical dependency displayed in Figure 9 (left). The mismatch in absolute numbers is due to the proportionality constant that is hard to determine in practice.

Appendix A: Description of datasets

In this paper we conducted experiments on the Reuters RCV1 dataset and the AOL query-log dataset. A description of these datasets is found below. For more information see [4] and [6].

A.1. RCV1 dataset

Reuters Corpus Volume I (RCV1) contains over 800,000 news stories which are provided by Reuters, Ltd. for research purposes. The dataset consists of all English language stories produced by Reuters journalists during the period of 365 days between August 20, 1996, and August 19, 1997. The stories have been formatted in xml and vary from a few hundred to several thousand words in length. The dataset is categorized across three dimensions: topics, industries, and regions. Special topic codes were assigned to describe the major subjects of a story.

In our experiments, the RCV1 dataset is pre-processed as follows. First the xml/html tags are removed and non-alphabetic characters (including numbers) are removed. Then all words are lowercased and stemmed while stop-words such as `the` or `of` are discarded. Lastly single character words are removed as are words appearing less than k ($k = 5$ in the experiments) times in the corpus.

A.2. AOL dataset

The AOL search query log dataset, which was provided by AOL for non-commercial research use, contains about 20 million web queries from 650,000 users. It contains all English language web queries from users during the 3 months between March 1 and May 31, 2006. Each query record contains an anonymous user ID number, the query issued by the user, and the time at which the query was submitted for search. If the user clicked on a search result, the rank of the item on which they clicked and the domain portion of the URL in the clicked result are also recorded. The dataset contains about 21 million queries averaging 3.5 words in length.

We pre-processed the AOL dataset in a similar manner to the RCV1 dataset. The web queries contain a huge mount of web addresses such as `apple.com` which were parsed as one word `apple.com` rather than two words `apple` and `com`.

Acknowledgements

The research in this paper was funded in part by NSF grant IIS-0746853.

References

[1] Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley.
[2] Jelinek, F. (1999). *Statistical methods for speech recognition*. MIT press.
[3] Jurafsky, D., Martin, J. H. and Kehler, A. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. MIT Press.

[4] Lewis, D., Yang, Y., Rose, T. and Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research* **5** 361–397.
[5] Manning, C. D. and Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press. MR1722790
[6] Pass, G., Chowdhury, A. and Torgeson, C. (2006). A picture of search. In *The First International Conference on Scalable Information Systems*.
[7] Trujillo, A. (1999). *Translation engines: techniques for machine translation*. Springer Verlag.
[8] Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval* **1** 69–90.