Understanding Documents Through Their Readers
Supplemental Material

Khalid El-Arini
Carnegie Mellon University
Pittsburgh, Pennsylvania
kbe@cs.cmu.edu

Min Xu
Carnegie Mellon University
Pittsburgh, Pennsylvania
minx@cs.cmu.edu

Emily B. Fox
University of Washington
Seattle, Washington
ebfox@uw.edu

Carlos Guestrin
University of Washington
Seattle, Washington
guestrin@cs.washington.edu

February 21, 2013

1 Learning the Dictionary

The first step of our approach involves learning a badge dictionary $\mathbf{B}$ from a training set of tweeted news articles. The observed data consists of pairs $(\mathbf{y}_i, \boldsymbol{\theta}_i)$, where each $\mathbf{y}_i$ represents the (ℓ_2-normalized) tf-idf vector for the content of document i, and $\boldsymbol{\theta}_i$ is an approximate badge vector for the users who share document i.

Specifically, we approximate $\boldsymbol{\theta}_i$ by taking each of the readers of document i, and assume a uniform distribution over the badges each of them declares in his or her profile. We then estimate $\boldsymbol{\theta}_i$ by aggregating over document i's readers.

For example, we consider an article i that was shared on Twitter by two users:

- Alice's Twitter profile contains the badges "liberal" and "feminist";
- Bob's Twitter profile contains the badges "liberal," "football" and "German."

In this case, we would assume Alice is half-liberal and half-feminist, while Bob is one third each: liberal, football and German. We would thus estimate $\boldsymbol{\theta}_i$ as the point-wise average of the two vectors: $\langle \text{liberal} : 0.5,\ \text{feminist} : 0.5 \rangle$ and $\langle \text{liberal} : 1/3,\ \text{football} : 1/3,\ \text{german} : 1/3 \rangle$.

Formally:

$$\theta_{ik} = \sum_{u \text{ tweeted } i} \frac{\delta_k^{(u)}}{\sum_j \delta_j^{(u)}},$$

where $\delta_k^{(u)}$ is a $0/1$ function indicating whether user u self-identifies with badge k.

With each $\mathbf{y}_i$ given and each $\boldsymbol{\theta}_i$ approximated, the badge dictionary $\mathbf{B}$ is learned by minimizing the following loss objective:

$$\min_{\mathbf{B} \geq 0} \sum_{i=1}^{N} \|\mathbf{y}_i - \mathbf{B}\boldsymbol{\theta}_i\|_2^2 + \lambda_B \sum_{j=1}^{V} \sum_{k=1}^{K} |\mathbf{B}_{jk}|. \tag{1}$$

We optimize Eq. (1) using a simple projected stochastic gradient descent, as outlined in Algorithm 1.

Algorithm 1 Projected stochastic gradient descent for learning the badge dictionary $\mathbf{B}$

// Data:
$\mathbf{\Theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N]$
$\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N]$
// Parameters:
Let m be the minibatch size
Let ϵ be the termination tolerance
Let α_0 be an initial step size parameter
Let λ_B be the sparsity regularization parameter
// Initialization:
$\mathbf{B}^{(0)} = \infty$
Set $\mathbf{B}^{(1)}$ to an initial value
$t \leftarrow 1$
// Termination condition
while $\|\mathbf{B}^{(t)} - \mathbf{B}^{(t-1)}\|_F > \epsilon$ **do**
  // Random minibatch
  $\mathcal{M} \leftarrow \mathrm{randsample}(N, m)$
  $\hat{\mathbf{\Theta}} \leftarrow \mathbf{\Theta}_{:,\mathcal{M}}$
  $\hat{\mathbf{Y}} \leftarrow \mathbf{Y}_{:,\mathcal{M}}$
  // Compute gradient
  $\nabla \leftarrow -\frac{1}{m}\left(\hat{\mathbf{Y}} - \mathbf{B}^{(t)}\hat{\mathbf{\Theta}}\right)\hat{\mathbf{\Theta}}^{\mathsf{T}}$
  // Set step size
  $\alpha = 1/\sqrt{\max(\alpha_0, t)}$
  // Take gradient step
  $\mathbf{B}^{(t+1)} \leftarrow \mathbf{B}^{(t)} - \alpha\nabla$
  // ℓ_1-projection via soft-thresholding, while maintaining non-negativity
  $\mathbf{B}^{(t+1)}_{>0} \leftarrow \mathbf{B}^{(t+1)}_{>0} - \lambda_B$
  $\mathbf{B}^{(t+1)}_{<0} \leftarrow 0$
  $t \leftarrow t + 1$

2 Data Processing

Here, we detail the steps of our entire data processing pipeline, for reproducibility of our experiments.

Initial Processing

1. Download three months worth of tweets from the Twitter Garden Hose: September 2010, September 2011 and September 2012.

2. Filter the tweets and extract those that are: (1) a tweet of a link; and, (2) came from a user with a non-empty profile. This leaves us with over 120 million tweets across the three months (cf. Table 1).

3. Filter the tweeted links to match one of 20,000 mainstream news sources, as defined by Google News. Additionally, we make sure the url ends in one of the following top level domains, to filter away content unlikely to be in English: {.com, .uk, .au, .ca, .us}.

4. Remove articles from *The Guardian*, as they will be in our test set.

Table 1: Training data statistics

Time period	Tweeted links with user profiles	Tweeted news articles in English	Vocabulary size	Number of badges
Sep. 2010	18,872,925	596,522	51,182	4,460
Sep. 2011	38,158,817	847,077	55,688	5,029
Sep. 2012	67,346,626	1,514,670	58,235	5,247

5. Download each tweeted news article, and remove the html boilerplate using Boilerpipe.[1]

6. Process each article by:

   - Removing all non-alphanumeric characters;
   - Filtering out stopwords;
   - Filtering out words shorter than 3 characters or longer than 25 characters;
   - Filtering out words that contain any non-English characters (fast, but imperfect, way to filter out most foreign language words).
   - Stem words using the Porter stemming algorithm.

Vocabulary Selection
At this point we proceed to select a vocabulary of words for each time period:

1. Ignore documents that have fewer than ten unique words, as well as documents where all the words appear only once.

2. Define the vocabulary to be all words that appear in at least 0.01% of all remaining tweeted news articles.

We reiterate that we have a separate vocabulary per time period. Also, as visualizing words from a stemmed vocabulary is not aesthetically pleasing, we take the convention in this chapter of displaying the most common unstemmed word for each word stem.

Badge Selection
In our experiments, we use badges that are single words. However, it is possible (and more desirable) to expand our definition of badges to include noun phrases, named entities or other syntactic constructs. Here, we compute statistics that will help us select a set of badges for each time period:

1. Extract each word appearing in any Twitter user profile from our training data set. We call these *badges*.

2. Stem the badges using the Porter stemming algorithm.

[1] `http://code.google.com/p/boilerpipe/`

Table 2: Number of articles per section of *The Guardian* in our test set

Time period	World	Sport	Opinion	Business	Life & Style	Science	Technology	UK
Sep. 2010	942	980	636	614	668	184	299	134
Sep. 2011	1,198	1,146	621	627	547	201	335	217
Sep. 2012	1,162	1,044	603	466	437	143	259	203

3. For each badge b, compute its cumulative weight across the training data by summing over all user profiles u, giving credit inversely proportional to the number of badges in a profile:

$$\sum_u \frac{\lambda_b^{(u)}}{\sum_a \lambda_a^{(u)}}.$$

4. Keep badges whose cumulative weight is at least 0.002% of the total number of tweeted articles in that time period.

Again, we have a separate badge set for each time period.

Test Data
For our test set, we download eight entire sections from *The Guardian*, a leading British newspaper, over the three months considered in our training set, comprising nearly 14,000 articles. We represent each test article as a tf-idf vector over the time-specific vocabulary constructed during training. We then code each article using the dictionaries learned from the training data. Statistics of our test data set can be found in Table 2.

3 Optimization

3.1 Dictionary Learning

In our projected stochastic gradient descent algorithm for learning our badge dictionary $\mathbf{B}$, we use the following parameters:

- Minibatch size $m = 500$.
- Our termination condition is based on a tolerance of $\epsilon = 10^{-8}$, or a maximum number of iterations of 10,000.
- Our initial step size parameter $\alpha_0 = 20$.
- Our sparsity regularization parameter $\lambda_B = 10^{-5}$.

We initialize $\mathbf{B} = \mathbf{Y}\mathbf{\Theta}^\mathsf{T}$, and then sparsify it, keeping the largest 200 values in each column, before renormalizing each column to have unit ℓ_2-norm.

3.2 Coding the Documents

We run the smoothed proximal gradient algorithm of Chen et al. [1] to optimize Eq. 3 in our paper, allowing us to code our test set of articles from *The Guardian*. We use three different settings of the regularization parameters:

1. **Heavily fused:** $\lambda_{\mathcal{G}} = 0.001$ and $\lambda_{\theta} = 2 \times 10^{-4}$;
2. **Lightly fused:** $\lambda_{\mathcal{G}} = 2 \times 10^{-5}$ and $\lambda_{\theta} = 2 \times 10^{-4}$;
3. **No fusion:** $\lambda_{\mathcal{G}} = 0$ and $\lambda_{\theta} = 2 \times 10^{-4}$.

We take our final concept representation to simply be the average of the codings generated by these three models. We found this set up to be most successful when running experiments on a separate validation data set.

4 Additional Experimental Details

4.1 Top Ten Badges

Here we can examine the ten badges that we use the most (i.e., highest total weight) to code the *Guardian* articles from September 2012, in our test set:

1. Guardian
2. Olympics
3. London
4. cricket
5. soccer
6. premier (as in, English Premier League)
7. tennis
8. fashion
9. gossip
10. Labour (as in, the British political party)

As we see in Figure 1, the words representing these ten badges align quite well with what we would expect. The most prevalent badge—"Guardian"—acts as a "background" badge in this particular data set, while the next six badges describe different aspects of sports, which represents a large proportion of our data set (cf. Table 2).

4.2 Odd-one-out

Section 5.4 of our main paper reports results on an *odd-one-out* metric, showing that our badge-based documen representation outpeforms alternative tf-idf and 100-topic LDA representations. Beyond these results presented in the main paper, here, in Figure 2, we see that this significant advantage holds true not just at an aggregate level, but in about 80% of the individual section pairings. Of the 56 possible pairings, only two resulted in significant wins on this metric by one of the competing techniques.

4.3 Case Study Details

Here are the fourteen political columnists we analyzed during our case study, and their July 2012 articles that we downloaded for analysis:

David Brooks

- http://www.nytimes.com/2012/07/13/opinion/brooks-why-our-elites-stink.html
- http://www.nytimes.com/2012/07/17/opinion/brooks-more-capitalism-please.html
- http://www.nytimes.com/2012/07/20/opinion/brooks-where-obama-shines.html
- http://www.nytimes.com/2012/07/24/opinion/brooks-more-treatment-programs.html
- http://www.nytimes.com/2012/07/27/opinion/brooks-the-olympic-contradiction.html
- http://www.nytimes.com/2012/07/31/opinion/brooks-dullest-campaign-ever.html

Ann Coulter

- http://www.anncoulter.com/columns/2012-07-04.html
- http://www.anncoulter.com/columns/2012-07-11.html
- http://www.anncoulter.com/columns/2012-07-18.html
- http://www.anncoulter.com/columns/2012-07-25.html

Maureen Dowd

- http://www.nytimes.com/2012/08/01/opinion/dowd-gadding-of-a-gawky-gowk.html
- http://www.nytimes.com/2012/07/29/opinion/sunday/dowd-mitts-olympic-meddle.html
- http://www.nytimes.com/2012/07/25/opinion/dowd-hiding-in-plain-sight.html
- http://www.nytimes.com/2012/07/22/opinion/sunday/dowd-paterno-sacked-off-his-pedestal.html
- http://www.nytimes.com/2012/07/18/opinion/dowd-whos-on-americas-side.html
- http://www.nytimes.com/2012/07/15/opinion/sunday/dowd-the-boy-who-wanted-to-fly.html
- http://www.nytimes.com/2012/07/08/opinion/sunday/cowboys-and-colleens.html
- http://www.nytimes.com/2012/07/04/opinion/gaelic-guerrilla.html
- http://www.nytimes.com/2012/07/01/opinion/sunday/the-wearing-of-the-green.html

Tom Friedman

- http://www.nytimes.com/2012/08/01/opinion/friedman-why-not-in-vegas.html
- http://www.nytimes.com/2012/07/29/opinion/sunday/friedman-coming-soon-the-big-trade-off.html
- http://www.nytimes.com/2012/07/25/opinion/friedman-syria-is-iraq.html
- http://www.nytimes.com/2012/07/22/opinion/sunday/friedman-the-launching-pad.html
- http://www.nytimes.com/2012/07/04/opinion/what-does-morsi-mean-for-israel.html
- http://www.nytimes.com/2012/07/01/opinion/sunday/taking-one-for-the-country.html

Jonah Goldberg

- http://www.nationalreview.com/articles/304711/live-free-and-uninsured-jonah-goldberg
- http://www.nationalreview.com/articles/304819/politics-and-symptoms-sick-culture-jonah-goldberg
- http://www.nationalreview.com/articles/308431/blame-barclays-not-capitalism-jonah-goldberg
- http://www.nationalreview.com/articles/309299/tilting-un-windmill-jonah-goldberg

- http://www.nationalreview.com/articles/309736/romney-and-bain-outsourcing-hysteria-jonah-goldberg
- http://www.nationalreview.com/articles/310080/co-sponsoring-your-success-jonah-goldberg
- http://www.nationalreview.com/articles/311235/brian-ross-s-brain-cramp-jonah-goldberg
- http://www.nationalreview.com/articles/312417/colorado-and-case-capital-punishment-jonah-goldberg

David Ignatius

- http://www.washingtonpost.com/opinions/david-ignatius-irans-bargaining-position-hardens/2012/07/02/gJQAi5NOJW_story.html
- http://www.washingtonpost.com/opinions/david-ignatius-israels-arab-spring-problem/2012/07/05/gJQAV5JrRW_story.html
- http://www.washingtonpost.com/opinions/david-ignatius-can-diplomacy-succeed-with-iran-and-syria/2012/07/11/gJQA7LwzdW_story.html
- http://www.washingtonpost.com/opinions/david-ignatius-pakistan-us-have-a-neurotic-relationship/2012/07/13/gJQABEDoiW_story.html
- http://www.washingtonpost.com/opinions/david-ignatius-syria-approaches-the-tipping-point/2012/07/18/gJQAFoCvtW_story.html
- http://www.washingtonpost.com/opinions/david-ignatius-central-banks-face-a-giant-bill-coming-due/2012/07/20/gJQALdJsyW_story.html
- http://www.washingtonpost.com/opinions/david-ignatius-the-day-after-in-syria/2012/07/25/gJQA4Uey9W_story.html
- http://www.washingtonpost.com/opinions/david-ignatius-senates-anti-leaking-bill-doesnt-address-the-real-sources-of-information/2012/07/31/gJQAPBElNX_story.html

Joe Klein

- http://swampland.time.com/2012/07/23/gunclingers-aurora-assault-weapons-and-the-rise-of-mass-shootings/
- http://swampland.time.com/2012/07/19/latest-column-93/
- http://swampland.time.com/2012/07/18/who-built-i-80/
- http://swampland.time.com/2012/07/15/inconvenient-truths/
- http://swampland.time.com/2012/07/13/bained/
- http://swampland.time.com/2012/07/11/a-friend-remembered/
- http://swampland.time.com/2012/07/02/you-say-tomato-i-call-bullpucky/

Charles Krauthammer

- http://www.washingtonpost.com/opinions/charles-krauthammer-the-imperial-presidency-revisited/2012/07/05/gJQAR66PQW_story.html
- http://www.washingtonpost.com/opinions/charles-krauthammer-the-islamist-ascendancy/2012/07/12/gJQArj9PgW_story.html
- http://www.washingtonpost.com/opinions/charles-krauthammer-did-the-state-make-you-great/2012/07/19/gJQAbZOiwW_story.html
- http://www.washingtonpost.com/opinions/charles-krauthammer-why-hes-going-where-hes-going/2012/07/26/gJQAGkzJCX_story.html
- http://www.washingtonpost.com/opinions/charles-krauthammer-busted-mr-pfeiffer-and-the-white-house-blog/2012/07/29/gJQA8M46IX_story.html

Nicholas Kristof

- http://www.nytimes.com/2012/07/01/opinion/sunday/africa-on-the-rise.html
- http://www.nytimes.com/2012/07/05/opinion/doughnuts-defeating-poverty.html
- http://www.nytimes.com/2012/07/08/opinion/sunday/the-coffin-maker-benchmark.html
- http://www.nytimes.com/2012/07/12/opinion/kristof-obamas-fantastic-boring-idea.html
- http://www.nytimes.com/2012/07/26/opinion/kristof-safe-from-fire-but-not-gone.html
- http://www.nytimes.com/2012/07/29/opinion/sunday/kristof-blissfully-lost-in-the-woods.html

William Kristol

- http://www.weeklystandard.com/articles/only-108-days-go_648828.html
- http://www.weeklystandard.com/articles/campaign-altogether-old_648556.html
- http://www.weeklystandard.com/articles/profiles-courage_648224.html
- http://www.weeklystandard.com/articles/obama-retreat_647776.html

Paul Krugman

- http://www.nytimes.com/2012/07/06/opinion/off-and-out-with-mitt-romney.html
- http://www.nytimes.com/2012/07/09/opinion/krugman-mitts-gray-areas.html
- http://www.nytimes.com/2012/07/13/opinion/krugman-whos-very-important.html
- http://www.nytimes.com/2012/07/16/opinion/krugman-policy-and-the-personal.html
- http://www.nytimes.com/2012/07/20/opinion/krugman-pathos-of-the-plutocrat.html
- http://www.nytimes.com/2012/07/23/opinion/krugman-loading-the-climate-dice.html
- http://www.nytimes.com/2012/07/27/opinion/money-for-nothing.html
- http://www.nytimes.com/2012/07/30/opinion/krugman-crash-of-the-bumblebee.html

Kathleen Parker

- http://www.washingtonpost.com/opinions/kathleen-parker-the-ladies-of-mount-vernon-have-preserved-washingtons-home/2012/07/03/gJQART6dLW_story.html
- http://www.washingtonpost.com/opinions/kathleen-parker-south-carolina-politics-gets-insulting/2012/07/06/gJQArcAfSW_story.html
- http://www.washingtonpost.com/opinions/kathleen-parker-doug-marlette-a-friend-remembered/2012/07/10/gJQAj8JfbW_story.html
- http://www.washingtonpost.com/opinions/kathleen-parker-romneys-critics-say-the-silliest-things/2012/07/13/gJQAXMNoiW_story.html
- http://www.washingtonpost.com/opinions/kathleen-parker-how-to-get-smart-news-literacy-programs-train-readers-to-look-beyond-infotain
2012/07/17/gJQAY1m2rW_story.html
- http://www.washingtonpost.com/opinions/kathleen-parker-obama-campaign-shows-its-desperation-in-romney-attack/2012/07/20/gJQAiYCsyW_story.html
- http://www.washingtonpost.com/opinions/kathleen-parker-in-poland-romney-addresses-economic-and-religious-freedom/2012/07/31/gJQA2c7kNX_story.html

Frank Rich

- http://nymag.com/daily/intel/2012/07/frank-rich-mitt-cant-wait-out-his-tax-storm.html

- http://nymag.com/daily/intel/2012/07/frank-rich-romney-has-a-tax-and-koch-problem.html
- http://nymag.com/news/frank-rich/declining-america-2012-7/

Fareed Zakaria

- http://fareedzakaria.com/2012/07/26/failure-to-launch/
- http://fareedzakaria.com/2012/07/18/what-voters-are-really-choosing-in-november/
- http://fareedzakaria.com/2012/07/12/tax-and-spend/
- http://fareedzakaria.com/2012/07/05/curbing-the-cost-of-health-care/

4.4 User Study Details

We recruited 118 participants for our study on Amazon Mechanical Turk, offering $0.20 per study completion. As our articles are from *The Guardian*, a British newspaper, we require participants that have good English language skills and meaningful ties to the United Kingdom. As we were unsuccessful in recruiting participants directly from Great Britain, we instead recruited exclusively from India. To improve the quality of our test set for our participants, we removed articles that were shorter than 1,200 characters, as well as those that contained the words "rugby" or "cricket."

References

[1] X. Chen, Q. Lin, S. Kim, J. Carbonell, and E. P. Xing. Smoothing proximal gradient method for general structure sparse regression. *Annals of Applied Statistics*, 6(2):719–752, 2012.

(a) 1. Guardian

(b) 2. Olympics

(c) 3. London

(d) 4. cricket

(e) 5. soccer

(f) 6. premier

(g) 7. tennis

(h) 8. fashion

(i) 9. gossip

(j) 10. Labour

Figure 1: The top ten badges used to code articles from *The Guardian* in our September 2012 test data set. The size of a word is proportional to its weight in the badge.

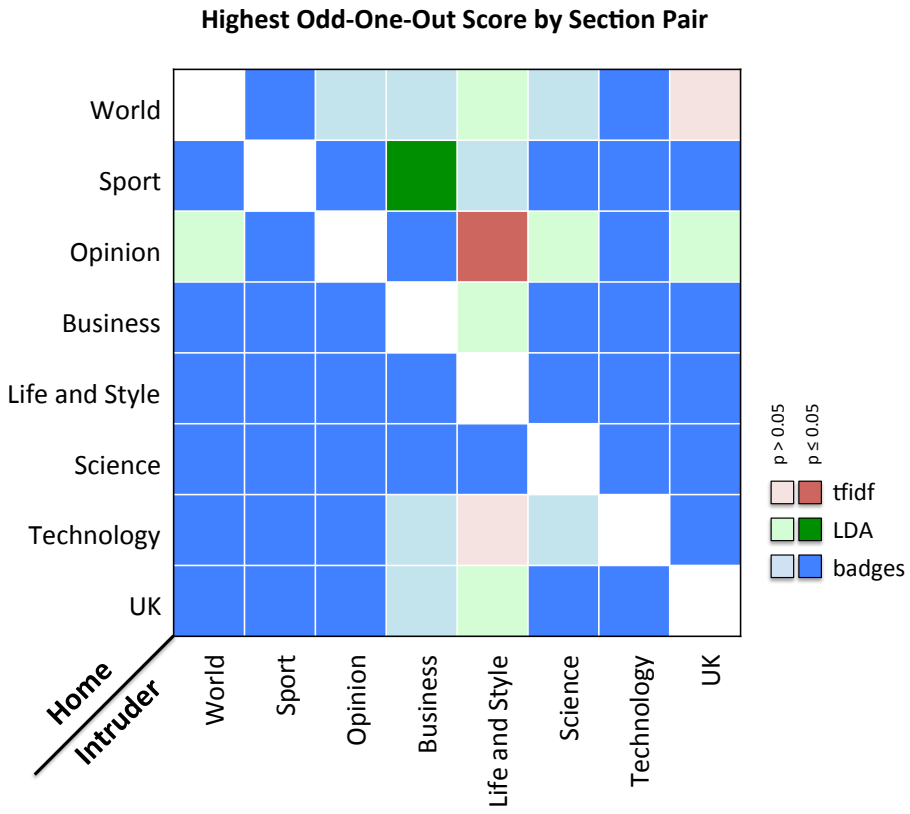

Figure 2: In about 80% of potential section pairs from *The Guardian*, the badge-based representation leads to a better odd-one-out score than the competing techniques. Darkly shaded cells are significant at the 95% confidence level, as indicated by the legend.