The LDA Model to Identify Job Ads

Enghin Atalay, Phai Phongthiengtham, Sebastian Sotelo, Daniel Tannenbaum[1]

Abstract

In this document, we detail the Latent Dirichlet Allocation (LDA) procedure used to identify which advertisements, in our newspaper text, are job ads.

Specifics of the LDA Model

Our next task towards producing our structured data set is to discard pages of advertisements which do not contain job ads. For each field of text, ProQuest has provided us—as a piece of metadata—a field which describes the type of newspaper content. Among these possible categories, we select text which correspond to a Display Ad or a Classified Ad.[2] Among Display Ads and Classified Ads, newspapers contain advertisements not only for job openings, but also for retail sales, real estate, or other non-job-related transactions. Unfortunately, the metadata in ProQuest's database do not distinguish between the different types of advertisements. As a result, we will need some algorithm to classify advertisements as job ads (which we want to store), as opposed to other types of advertisements (which we do not).

To do so, we construct a Latent Dirichlet Allocation (LDA) model.[3] According to the LDA model, each document (in our context, a page of advertisements) in a corpus (for us, all of the advertisements) belongs to one of K potential "topics". The topic of a given document is a hidden object. The probability that a given document belongs to topic $k \in K$ is determined according to the realization of a Dirichlet-distributed random variable. Then, for documents within a given topic, the probability that a word $w \in \{1, ...W\}$ is observed is given by a K by W matrix, β. Estimation of the LDA model involves estimation of the parameters of the Dirichlet distribution and of the elements of β. Once the model has been

[1] Atalay and Phongthiengtham: Department of Economics, University of Wisconsin-Madison. Sotelo: Department of Economics, University of Michigan-Ann Arbor. Tannenbaum: Department of Economics, University of Nebraska-Lincoln. We acknowledge financial support from the Washington Center for Equitable Growth.

[2] Other categories of newspaper text include "Articles," "Banners," Editorial Articles," "Letters to the editor," "Obituaries," and "Stock Quotes."

[3] For additional background on LDA models, see the appendix of our paper, Blei, Ng, and Jordan (2003), or Hoffman, Bach, and Blei (2010).

estimated, we can derive the probability that for any document (including those outside of the corpus used to estimate the model) the document belongs to any topic k; this likelihood is computed by multiplying the conditional probabilities (conditional on the topic) of the words in the document.

In estimating our LDA model, we separate the five subsamples of text in our database: the text from the Boston Globe, the New York Times, and the Wall Street Journal Classified Ads, and the text from the Boston Globe and the New York Times Display Ads. Quoting from the appendix of our paper, for each of these subsamples, we

> *remove* stop words *(e.g., common words like "a," "the," and "and"), numerals, and words which are not contained in the English dictionary. We then stem words; that is, we remove word affixes so that words in different forms—singular nouns, plural nouns, verbs, adjectives, adverbs—are grouped as one. (To emphasize, the removal of certain types of words and the stemming of words pertains only in constructing our LDA model.)*

Having pre-processed this text for our LDA estimation, we generate a random sample of 100 thousand pages of ads. (It would be computationally infeasible to estimate a model on the full sample of pages of ads.) Furthermore, it would be computationally infeasible to estimate a matrix β which has the number of columns equal to the number of distinct word stems in our text corpus. We follow standard practice and drop (from the β matrix) all words which appear fewer than 5 times in our text corpus, or in greater than 95 percent of the documents in our corpus. After these two deletions, we keep the top 1000 word stems in estimating our matrix, β.

Results from the LDA Model

We follow the procedure of the previous subsection. The result of this estimation is given in LDA_results.xlsx, in the in the same website as the one hosting this document. The estimation does not explicitly indicate the set of topics which relate to job ads. However, by casual inspection, it is clear that the words in one topic pertain to employment. For example, for the estimation based on Boston Globe Classified Ads, a topic has "auto," "new," and "car" as the word stems with the highest values for in their corresponding elements of the β matrix. For the fourth topic (out of five topics), the highest-β word stems are "opportun," "experi," "work," "call," and "salari." We thus identify this fourth topic as representing job ads. The number of topics, K, is chosen so that i) with K topics there is a single job-related topic, and with ii) K+1 topics, there are multiple job-related topics. This rule yields 5 topics for the Boston Globe Classified Ad subsample, 5 topics for the Boston Globe Display Ad subsample, 3 topics for the New York Times Classified Ad subsample, 12 topics for the New York Times Display Ad subsample, and 5 topics for the Wall Street Journal Classified Ad subsample.

For each page of ads (not only the 100 thousand ads used to estimate the LDA model, but instead the entire text corpus), our estimated model generates a probability distribution, characterizing the likelihood that the page of ads belongs to the topic associated with job ads. Figure 1 plots histograms of pages' LDA-model-assessed likelihood of containing job ads (as opposed to other types of advertisements). The top-left panel, for example, indicates that roughly four thousand pages of ads, among Boston Globe Classified Ads, have a near 100 percent likelihood of containing job ads. For the New York Times Classified Ads (depicted in the left middle panel), greater than 100 thousand pages of Classified Ads have a 100 percent likelihood of containing job ads. For each newspaper, the distribution of the likelihood of containing job ads is bi-modal, with peaks near 0 and near 0. In other words, for most pages of ads, the LDA model can precisely identify whether the page contains job ads.

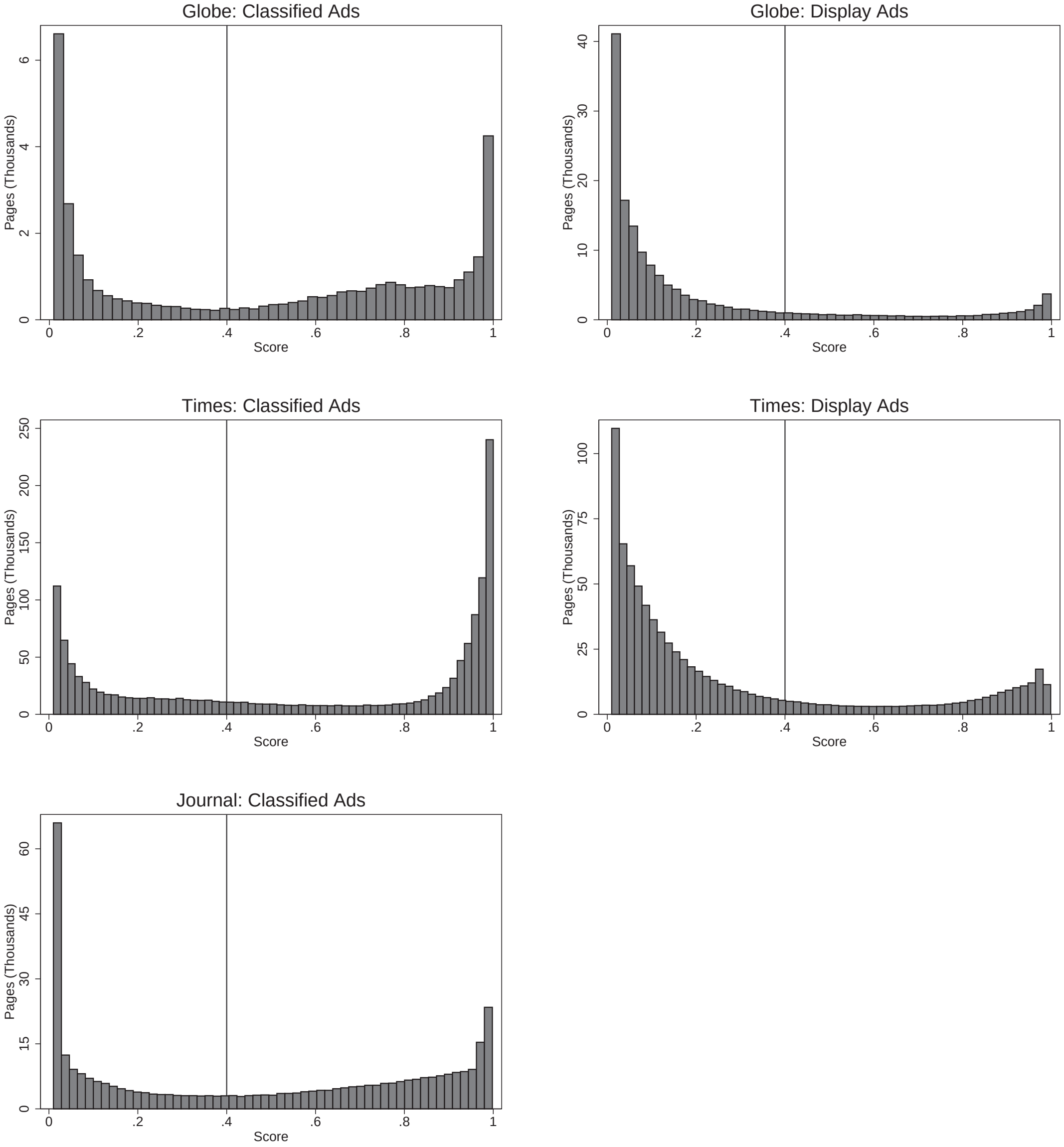
In classifying ads, we choose a 40 percent cutoff (and depict this cutoff with vertical lines in Figure (depicted using vertical lines in Figure 1): If the estimated likelihood is above 0.40, we categorize the page of ads as one which contains job ads. We discard ads for which the assessed likelihood is less than 0.40. The choice of the cutoff will determine the number of type I versus type II errors we make when classifying ads. However, given the low probability assigned to intermediate likelihood values, the the number of such (type I versus type II) errors we make is relatively insensitive to our choice of cutoff.

In the following five figures, we present several examples of chunks of display ads, all from January 1979 editions of the Boston Globe. These figures are presented in order of increasing likelihood (according to our LDA model) of representing a group of job ads.

Figure 2 comprise Display Ad #94 from the January 14, 1979 edition of the Boston Globe. For this chunk of advertisements, our LDA model assigns a likelihood of 0.36 that this chunk of text contains job advertisements. The model assigns a non-zero value for this likelihood because of the presence of the word "experience," which is recurrent in the job-ad-related topic. (Throughout Figures 2 to 6, we mark in bold the words most predictive of job-advertisements.) This assessed likelihood that this page of advertisements contain job ads, 0.36, is just below our chosen threshold of 0.4. Based on this cut-off value, we would exclude these text in our. (Even if we had chosen a lower cutoff — say 0.30 instead of 0.40 — this page of advertisements still would have been eventually excluded from our final data set for a different reason, due to the large number of garbled and misspelled words.)

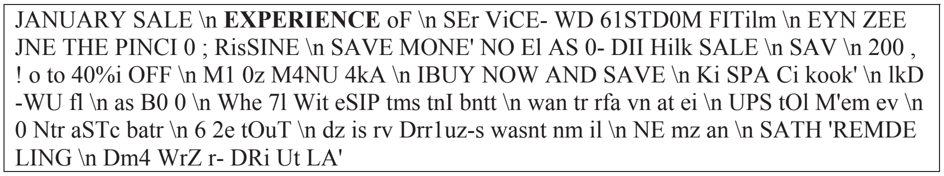
In Figures 3 and 4, we present snippets from two different chunks of display ads, the first from January 7, 1979, and the second from January 11, 1979. For each of these two pages

Figure 1: Histogram of the Likelihood of Job-Related Ads



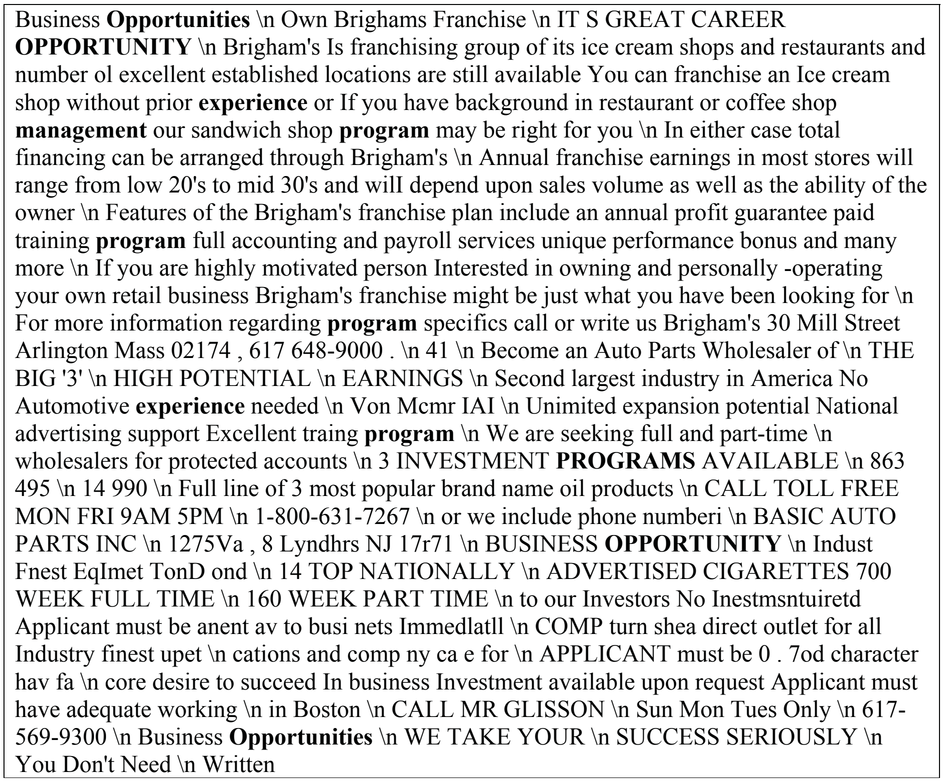
Notes: To make the graph easier to read, these histograms exclude pages of ads which were assigned exactly a zero probability of containing job ads. For the Boston Globe Classified Ads, there were 29 thousand (out of a total of 67 thousand pages) with an assigned zero probability of containing job ads. For the other subsamples, the number of pages of ads with zero assigned probability are: 177 thousand (out of a total of 332 thousand) for the Boston Globe Display Ad subsample; 444 thousand (out of a total of 1.89 million) for the New York Times Classified Ad subsample; 1.82 million (out of a total of 2.61 million) pages, for the New York Times Display Ad subsample; and 151 thousand (out of a total number of 522 thousand) pages of ads for the Wall Street Journal Classified Ads subsample.

Figure 2: January 14, 1979 Boston Globe, Display Ad #94



JANUARY SALE \n **EXPERIENCE** oF \n SEr ViCE- WD 61STD0M FITilm \n EYN ZEE JNE THE PINCI 0 ; RisSINE \n SAVE MONE' NO El AS 0- DII Hilk SALE \n SAV \n 200 , ! o to 40%i OFF \n M1 0z M4NU 4kA \n IBUY NOW AND SAVE \n Ki SPA Ci kook' \n lkD -WU fl \n as B0 0 \n Whe 7l Wit eSIP tms tnI bntt \n wan tr rfa vn at ei \n UPS tOl M'em ev \n 0 Ntr aSTc batr \n 6 2e tOuT \n dz is rv Drr1uz-s wasnt nm il \n NE mz an \n SATH 'REMDE LING \n Dm4 WrZ r- DRi Ut LA'

Figure 3: Snippet of the January 7, 1979 Boston Globe, Display Ad #258



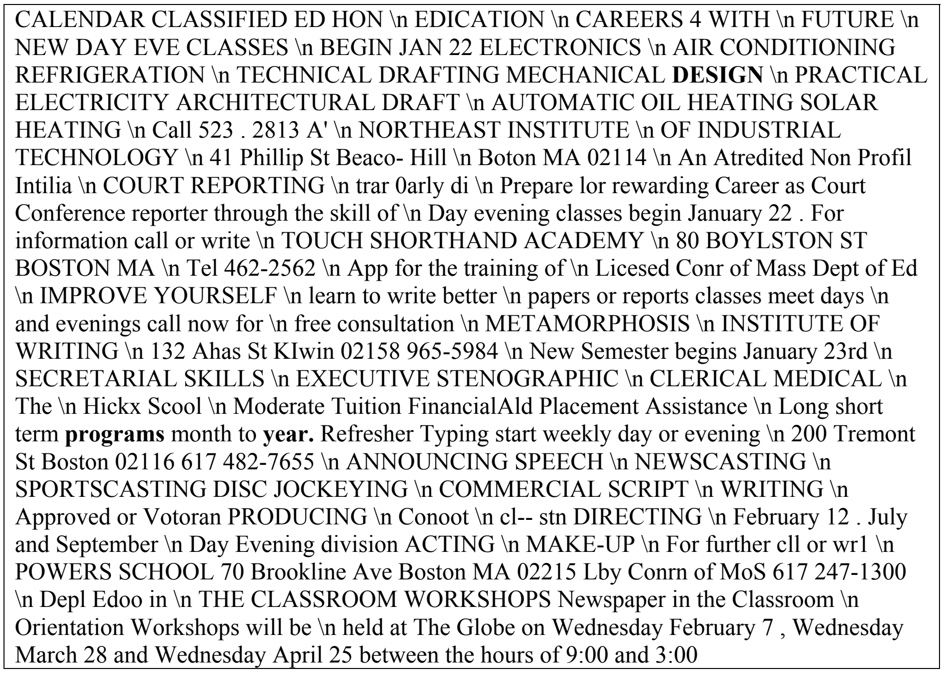
Business **Opportunities** \n Own Brighams Franchise \n IT S GREAT CAREER **OPPORTUNITY** \n Brigham's Is franchising group of its ice cream shops and restaurants and number ol excellent established locations are still available You can franchise an Ice cream shop without prior **experience** or If you have background in restaurant or coffee shop **management** our sandwich shop **program** may be right for you \n In either case total financing can be arranged through Brigham's \n Annual franchise earnings in most stores will range from low 20's to mid 30's and wilI depend upon sales volume as well as the ability of the owner \n Features of the Brigham's franchise plan include an annual profit guarantee paid training **program** full accounting and payroll services unique performance bonus and many more \n If you are highly motivated person Interested in owning and personally -operating your own retail business Brigham's franchise might be just what you have been looking for \n For more information regarding **program** specifics call or write us Brigham's 30 Mill Street Arlington Mass 02174 , 617 648-9000 . \n 41 \n Become an Auto Parts Wholesaler of \n THE BIG '3' \n HIGH POTENTIAL \n EARNINGS \n Second largest industry in America No Automotive **experience** needed \n Von Mcmr IAI \n Unimited expansion potential National advertising support Excellent traing **program** \n We are seeking full and part-time \n wholesalers for protected accounts \n 3 INVESTMENT **PROGRAMS** AVAILABLE \n 863 495 \n 14 990 \n Full line of 3 most popular brand name oil products \n CALL TOLL FREE MON FRI 9AM 5PM \n 1-800-631-7267 \n or we include phone numberi \n BASIC AUTO PARTS INC \n 1275Va , 8 Lyndhrs NJ 17r71 \n BUSINESS **OPPORTUNITY** \n Indust Fnest EqImet TonD ond \n 14 TOP NATIONALLY \n ADVERTISED CIGARETTES 700 WEEK FULL TIME \n 160 WEEK PART TIME \n to our Investors No Inestmsntuiretd Applicant must be anent av to busi nets Immedlatll \n COMP turn shea direct outlet for all Industry finest upet \n cations and comp ny ca e for \n APPLICANT must be 0 . 7od character hav fa \n core desire to succeed In business Investment available upon request Applicant must have adequate working \n in Boston \n CALL MR GLISSON \n Sun Mon Tues Only \n 617-569-9300 \n Business **Opportunities** \n WE TAKE YOUR \n SUCCESS SERIOUSLY \n You Don't Need \n Written

of text, the model-assessed probability that the page contain job ads equals 0.43 and 0.58, respectively. It appears that the ad in Figure 3 is for a business franchise, while Figure 4 is an advertisement for a job-training course. These advertisements contain a modest number of job-related words, and as a result are assigned to have a moderate probability of being job-related.

In Figure 5, we present a set of advertisements for which the LDA model assigns a 63 percent likelihood of being job-ad related. Finally, Figure 6 depicts a snippet of a page of text for which the LDA model assigns a 92 percent likelihood comprising job ads. These latter two figures contain a substantially higher frequency of words related to employment.

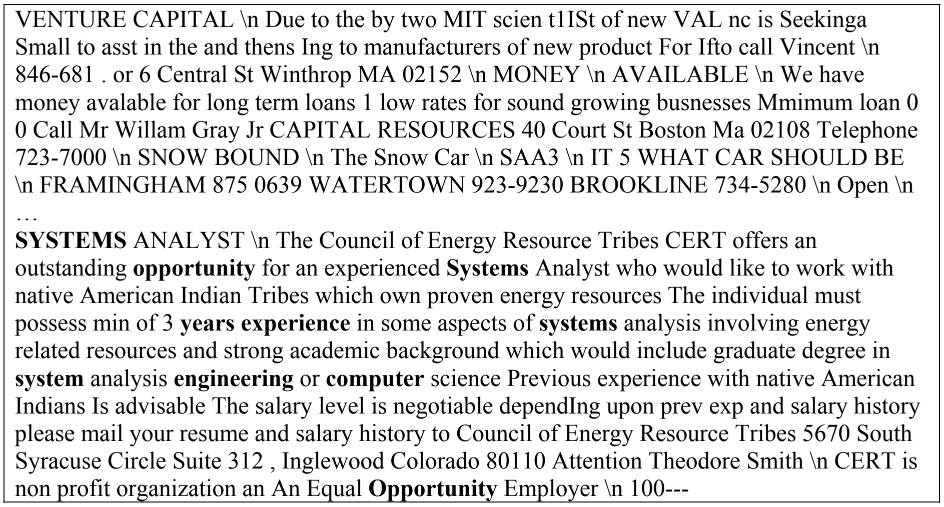
To sum up, several words and phrases appear frequently in job ads, and infrequently in other types of advertisements. Using this fact, we construct a model which identifies

Figure 4: Snippet of the January 11, 1979 Boston Globe, Display Ad #58



CALENDAR CLASSIFIED ED HON \n EDICATION \n CAREERS 4 WITH \n FUTURE \n NEW DAY EVE CLASSES \n BEGIN JAN 22 ELECTRONICS \n AIR CONDITIONING REFRIGERATION \n TECHNICAL DRAFTING MECHANICAL **DESIGN** \n PRACTICAL ELECTRICITY ARCHITECTURAL DRAFT \n AUTOMATIC OIL HEATING SOLAR HEATING \n Call 523 . 2813 A' \n NORTHEAST INSTITUTE \n OF INDUSTRIAL TECHNOLOGY \n 41 Phillip St Beaco- Hill \n Boton MA 02114 \n An Atredited Non Profil Intilia \n COURT REPORTING \n trar 0arly di \n Prepare lor rewarding Career as Court Conference reporter through the skill of \n Day evening classes begin January 22 . For information call or write \n TOUCH SHORTHAND ACADEMY \n 80 BOYLSTON ST BOSTON MA \n Tel 462-2562 \n App for the training of \n Licesed Conr of Mass Dept of Ed \n IMPROVE YOURSELF \n learn to write better \n papers or reports classes meet days \n and evenings call now for \n free consultation \n METAMORPHOSIS \n INSTITUTE OF WRITING \n 132 Ahas St KIwin 02158 965-5984 \n New Semester begins January 23rd \n SECRETARIAL SKILLS \n EXECUTIVE STENOGRAPHIC \n CLERICAL MEDICAL \n The \n Hickx Scool \n Moderate Tuition FinancialAld Placement Assistance \n Long short term **programs** month to **year.** Refresher Typing start weekly day or evening \n 200 Tremont St Boston 02116 617 482-7655 \n ANNOUNCING SPEECH \n NEWSCASTING \n SPORTSCASTING DISC JOCKEYING \n COMMERCIAL SCRIPT \n WRITING \n Approved or Votoran PRODUCING \n Conoot \n cl-- stn DIRECTING \n February 12 . July and September \n Day Evening division ACTING \n MAKE-UP \n For further cll or wr1 \n POWERS SCHOOL 70 Brookline Ave Boston MA 02215 Lby Conrn of MoS 617 247-1300 \n Depl Edoo in \n THE CLASSROOM WORKSHOPS Newspaper in the Classroom \n Orientation Workshops will be \n held at The Globe on Wednesday February 7 , Wednesday March 28 and Wednesday April 25 between the hours of 9:00 and 3:00

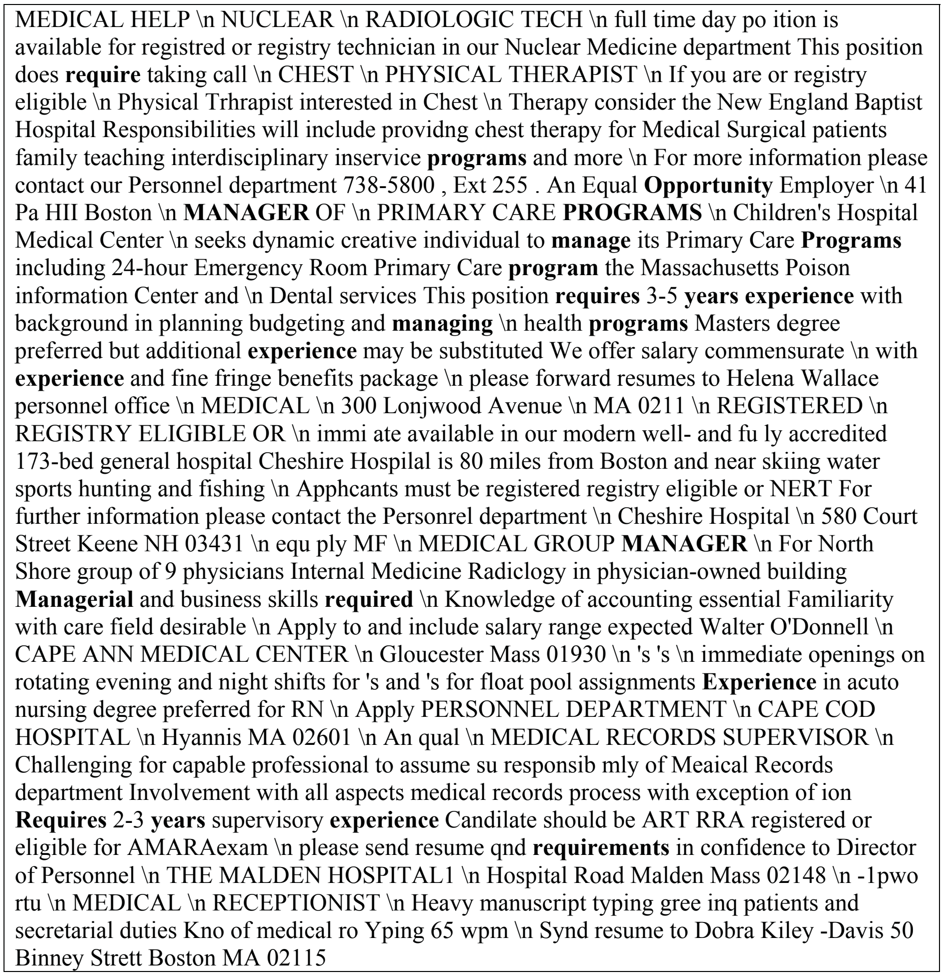
Figure 5: Snippet of the January 7, 1979 Boston Globe, Display Ad #19



VENTURE CAPITAL \n Due to the by two MIT scien t1ISt of new VAL nc is Seekinga Small to asst in the and thens Ing to manufacturers of new product For Ifto call Vincent \n 846-681 . or 6 Central St Winthrop MA 02152 \n MONEY \n AVAILABLE \n We have money avalable for long term loans 1 low rates for sound growing busnesses Mmimum loan 0 0 Call Mr Willam Gray Jr CAPITAL RESOURCES 40 Court St Boston Ma 02108 Telephone 723-7000 \n SNOW BOUND \n The Snow Car \n SAA3 \n IT 5 WHAT CAR SHOULD BE \n FRAMINGHAM 875 0639 WATERTOWN 923-9230 BROOKLINE 734-5280 \n Open \n …

SYSTEMS ANALYST \n The Council of Energy Resource Tribes CERT offers an outstanding **opportunity** for an experienced **Systems** Analyst who would like to work with native American Indian Tribes which own proven energy resources The individual must possess min of 3 **years experience** in some aspects of **systems** analysis involving energy related resources and strong academic background which would include graduate degree in **system** analysis **engineering** or **computer** science Previous experience with native American Indians Is advisable The salary level is negotiable dependIng upon prev exp and salary history please mail your resume and salary history to Council of Energy Resource Tribes 5670 South Syracuse Circle Suite 312 , Inglewood Colorado 80110 Attention Theodore Smith \n CERT is non profit organization an An Equal **Opportunity** Employer \n 100---

Figure 6: Snippet of the January 14, 1979 Boston Globe, Display Ad #226



MEDICAL HELP \n NUCLEAR \n RADIOLOGIC TECH \n full time day po ition is available for registred or registry technician in our Nuclear Medicine department This position does **require** taking call \n CHEST \n PHYSICAL THERAPIST \n If you are or registry eligible \n Physical Trhrapist interested in Chest \n Therapy consider the New England Baptist Hospital Responsibilities will include providng chest therapy for Medical Surgical patients family teaching interdisciplinary inservice **programs** and more \n For more information please contact our Personnel department 738-5800 , Ext 255 . An Equal **Opportunity** Employer \n 41 Pa HII Boston \n **MANAGER** OF \n PRIMARY CARE **PROGRAMS** \n Children's Hospital Medical Center \n seeks dynamic creative individual to **manage** its Primary Care **Programs** including 24-hour Emergency Room Primary Care **program** the Massachusetts Poison information Center and \n Dental services This position **requires** 3-5 **years experience** with background in planning budgeting and **managing** \n health **programs** Masters degree preferred but additional **experience** may be substituted We offer salary commensurate \n with **experience** and fine fringe benefits package \n please forward resumes to Helena Wallace personnel office \n MEDICAL \n 300 Lonjwood Avenue \n MA 0211 \n REGISTERED \n REGISTRY ELIGIBLE OR \n immi ate available in our modern well- and fu ly accredited 173-bed general hospital Cheshire Hospilal is 80 miles from Boston and near skiing water sports hunting and fishing \n Apphcants must be registered registry eligible or NERT For further information please contact the Personrel department \n Cheshire Hospital \n 580 Court Street Keene NH 03431 \n equ ply MF \n MEDICAL GROUP **MANAGER** \n For North Shore group of 9 physicians Internal Medicine Radiclogy in physician-owned building **Managerial** and business skills **required** \n Knowledge of accounting essential Familiarity with care field desirable \n Apply to and include salary range expected Walter O'Donnell \n CAPE ANN MEDICAL CENTER \n Gloucester Mass 01930 \n 's 's \n immediate openings on rotating evening and night shifts for 's and 's for float pool assignments **Experience** in acuto nursing degree preferred for RN \n Apply PERSONNEL DEPARTMENT \n CAPE COD HOSPITAL \n Hyannis MA 02601 \n An qual \n MEDICAL RECORDS SUPERVISOR \n Challenging for capable professional to assume su responsib mly of Meaical Records department Involvement with all aspects medical records process with exception of ion **Requires** 2-3 **years** supervisory **experience** Candilate should be ART RRA registered or eligible for AMARAexam \n please send resume qnd **requirements** in confidence to Director of Personnel \n THE MALDEN HOSPITAL1 \n Hospital Road Malden Mass 02148 \n -1pwo rtu \n MEDICAL \n RECEPTIONIST \n Heavy manuscript typing gree inq patients and secretarial duties Kno of medical ro Yping 65 wpm \n Synd resume to Dobra Kiley -Davis 50 Binney Strett Boston MA 02115 the words that are tend to appear in the same sets of documents, and tag ads as likely to represent job ads if they are rich in the words like “manager,” “experience,” “requirement,” or “opportunity.” Many of the pages of ads in our data set do not contain any of these words, and are thus assigned an exceedingly low probability of representing a set of job ads. At the same time, many pages of ads are rich in job-ad-specific words, sufficiently so that we are confident that we are accurately distinguishing job ads from other types of ads.

References

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research*, 3: 993–1022.

[2] Hoffman, M. D., Francis R. Bach, and David M. Blei. 2010. “Online Learning for Latent Dirichlet Allocation,” in *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010*: 856–864.