The Initial Text Cleaning

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

Abstract

In this document, we summarize the initial processing of the raw text, provided to us by ProQuest. The main components of this step are to retrieve document metadata, to remove markup from the newspaper text, and to perform an initial spell-check of the text.

The underlying text from which our database is constructed was provided to us by ProQuest. The text was produced via a OCR (Optical Character Recognition) conversion of scanned images of previously published editions of the Boston Globe, New York Times, and Wall Street Journal. The original data were published in 1960 to 1983 for the Boston Globe, 1940 to 2000 for the New York Times, and 1940 to 1998 for the Wall Street Journal.

Figure 1 presents the unprocessed text and some of the associated metadata, as provided to us by ProQuest. Figure 2 contains the same page of ads, having removed the superfluous markup, combining hyphenated words across subsequent lines, and performing a preliminary spell check. The first few fields within Figure 1 characterize the newspaper, the date of publication, and the record type and title. The record type and title indicate that Figure 1 represents the 45th page of classified ads in the September 12 edition of the Wall Street Journal. The final field contains the text of this page of classified ads. This page contains a single set of ads, posted by the Singer Corporation, for three separate positions: for an Electromechanical Engineer, for an Electronic or Mechanical Engineer, and for a Designer Specialist. We store the publication date and publisher, the record title, and the text. (Newspaper Ads could be advertising not only for open jobs, but also for other types of products or transactions. In the "LDA Model to Identify Job Ads" document, we specify our procedure to separate job ads from other types of ads.)

First, we must remove any of the xml markup present in this ad. In particular, we make the following substitutions: replace """ with a quotation, replace "'" with an apostrophe, replace "&" with an ampersand, and delete "&", "<", and ">". The entire pattern "</p>" denotes a line break. When a word

[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.

Figure 1: Raw Text and Markup: September 12, 1978 Wall Street Journal, Classified Ad #45


```
----------thisadendshere----------
<?xml version="1.0" encoding="UTF-8"?>
<record>
 <version>
  TDM_Record_v1.0.xsd
 </version>
 <recordid>
  4a667155d557ab68c878224bc3de0979
 </recordid>
 <recordtitle>
  Classified Ad 45 -- No Title
 </recordtitle>
 <alphapubdate>
  Sep 12, 1978
 </alphapubdate>
 <numericpubdate>
  19780912
 </numericpubdate>
 <objecttype>
  classified_ad
 </objecttype>
…
 <fulltext>
  &amp;lt;html&amp;gt; &amp;lt;head&amp;gt; &amp;lt;meta
name=&amp;quot;ValidationSchema&amp;quot; content=&amp;quot;http://www.w3.org/2002/08/xhtml/xhtml1-
strict.xsd&amp;quot;/&amp;gt; &amp;lt;title/&amp;gt; &amp;lt;/head&amp;gt;
&amp;lt;body&amp;gt; &amp;lt;p&amp;gt; Singer has long been one of the world s gr &amp;apos; pacesetters in
volume manufacturing of intricate, &amp;lt;/p&amp;gt; &amp;lt;p&amp;gt; precision machines that achieve
extreme reliability and durability. Our sewing machines are in use around the globe in every kind of climate. As pioneers in
electronic sewing machines, we have again set new standards. &amp;lt;/p&amp;gt; &amp;lt;p&amp;gt;
ELECTROMECHANICAL ENGINEERS, &amp;lt;/p&amp;gt; &amp;lt;p&amp;gt; Minimum of 6 eara experience
in developing of electromechanical consumer or atm&amp;amp;gt;lar products. BSME or BSEE degree required,
&amp;lt;/p&amp;gt; &amp;lt;p&amp;gt; advanced degree preferred. &amp;lt;/p&amp;gt;
&amp;lt;p&amp;gt; ELECTRONIC ENGINEERS MECHANICAL ENGINEERS &amp;lt;/p&amp;gt;
&amp;lt;p&amp;gt; A least 2+ years is needed in one of 2+ years of practical in mechanisms the following areas: and
machine design analysis. Working know! edga of computers as a design tool would be &amp;lt;/p&amp;gt;
&amp;lt;p&amp;gt; 1) Analog and digital industrial electron helpful. Experience in sophisticated , with microprocessor and
CAD knowl chanical products. Background should include edge desirable; mechanism or gear or machine design 2) Analog
sad digital circuitry, logic de and analysis. Knowledge of computers as , PC bond design, ISI and minicom neering ardes
helpful &amp;lt;/p&amp;gt; &amp;lt;p&amp;gt; puter ; &amp;lt;/p&amp;gt; &amp;lt;p&amp;gt; S)
Application of mini and micro-computers including , and hardware de- &amp;lt;/p&amp;gt; &amp;lt;p&amp;gt;
bugging of analog and digital circuitry. &amp;lt;/p&amp;gt; &amp;lt;p&amp;gt; DESIGNERS, JUNIOR
SPECIALIST AND SENIOR &amp;lt;/p&amp;gt; &amp;lt;p&amp;gt; Ezperience in fractional and AC 1-8 Years
experience in precision high toler _ and DC motors and motor control system as ante design of mechanical devices and/or
circuit well as other electromechanical devices. layout. Intricate detailing experience mandato- &amp;lt;/p&amp;gt;
&amp;lt;p&amp;gt; ry. Singer offers attractive salaries, benefits and professional working conditions, and very favorable
career . These positions are located at our Elizabeth, New Jersey facility and at our R&amp;amp;amp;D Laboratory in
Fairfield, New Jersey. &amp;lt;/p&amp;gt; &amp;lt;p&amp;gt; Please send resume stating position of interest in
confidence to: &amp;lt;/p&amp;gt; &amp;lt;p&amp;gt; Hosie Scott, Employment Manager &amp;lt;/p&amp;gt;
&amp;lt;p&amp;gt; or call (201) 527-6166 or 67 &amp;lt;/p&amp;gt; &amp;lt;p&amp;gt; SINGER
&amp;lt;/p&amp;gt; &amp;lt;p&amp;gt; DIVERSIFIED WORL. 321 First Street &amp;lt;/p&amp;gt;
&amp;lt;p&amp;gt; Elizabeth, New Jersey 07207 An Equal Opportunity Employer M/F &amp;lt;/p&amp;gt;
&amp;lt;/body&amp;gt; &amp;lt;/html&amp;gt;
 </fulltext>
</record>
----------thisadendshere----------
```


Notes: To fit this figure on a single page, we have excised some of the metadata. The triple dot, "...", indicates the point at which this excision occurs.

Figure 2: Cleaned Text: September 12, 1978 Wall Street Journal, Classified Ad #45

singer has long been one of the world s gr ' pacesetters in volume manufacturing of intricate , \n precision machines that achieve extreme reliability and durability . our sewing machines are in use around the globe in every kind of climate . as pioneers in electronic sewing machines , we have again set new standards \n electromechanical engineers \n minimum of 6 ear experience in developing of electromechanical consumer or atmlar products . bsme or b see degree required \n, advanced degree preferred \n electronic mechanical engineers \n a least 2+ years is needed in one of 2+ years of practical in mechanisms the following areas and machine design analysis . working know ! Edgar of computers as a design tool would be \n 1) analog and digital industrial electron helpful . experience in sophisticated , with microprocessor and cad knowledge mechanical products . background should include edge desirable ; mechanism or gear or machine design 2) analog sad digital circuitry , logic de and analysis . knowledge of computers as , pc bond design , is i and mini com sneering ares \n helpful punter \n ; s) \n application of mini and micro-computers including , and hardware debugging \n of analog and digital circuitry . \n designers junior specialist and senior \n experience in fractional and ac 1-8 years experience in precision high toiler and dc motors and motor control system as ante design of mechanical devices and or circuit well as other electromechanical devices . layout . intricate detailing experience \n mandatory . singer offers attractive salaries , benefits and professional working conditions , and very favorable Career . these positions are located at our Elizabeth , new jersey \n facility and at our r ; d laboratory in f airfield , new jersey . please send resume stating position of interest in confidence to \n: hosier scott , employment manager \n or call (201) 527-6166 or 67 \n singer \n diversified \n world . 321 first street \n elizabeth , new jersey 07207 an equal opportunity employer m f

Notes: A "\n" refers to a line a break.

appears with a piece of punctuation within a space, we include a separating space. (The reason for this last substitution is to make it such that an individual word uniquely defines a token. Otherwise, for example, "word" and "word!" would be identified as unique tokens.)

Second, we combine words which appear on two subsequent lines, separated by a hyphen. There are many such words due to the formatting of newspaper job ads in short columns of text. For example, in Figure 1, "de-" and "bugging" are on two subsequent lines. Our procedure searches for a hyphen at the end of a line, checks whether either of the two words which straddle a line are misspelled, and if the combined words are correctly spelled. If so, we concatenate the two hyphenated words. This procedure similarly replaces "mandato-" "ry" with "mandatory."

Third, we perform a spell check to undo any transcription errors induced by the OCR procedure. Our spell checker replaces "WORL" with "world," "Ezperience" with "experience."

In other cases, our spell checker was unable to correct the transcription error: "6 eara experience" should probably be changed replaced by "6 year experience," rather than "6 ear experience," while "atmlar products" should be replaced by "similar." We record the fraction of words which are incorrectly spelled. Eventually, after performing the procedures to determine which pages of ads represent job ads, to determine the boundaries between

individual job ads, and to identify the job title of each ad, we will exclude advertisements if the fraction of correctly spelled words is above a certain threshold, 35 percent.