

# Name of your project

**First Author**  
email@domain

**Second Author**  
email@domain

**Third Author**  
email@domain

## Abstract

This document contains the instructions for preparing a report for ADA 2017. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. This document is based on the ACL 2014 paper format.

## 1 Credits

This document has been adapted from the instructions for earlier ACL proceedings, including those for ACL-2012 by Maggie Li and Michael White, those from ACL-2010 by Jing-Shing Chang and Philipp Koehn, those for ACL-2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, those for ACL-2005 by Hwee Tou Ng and Kemal Oflazer, those for ACL-2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats. Those versions were written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence*.

## 2 Introduction

The following instructions are directed to the ADA students who decided to prepare a report. Authors are required to provide a Portable Document Format (PDF) version of their reports. It must be maximum 4-page long, excluding references.

## 3 General Instructions

Manuscripts must be in two-column format. Exceptions to the two-column format include the title, authors' names and complete addresses, which must be centered at the top of the first page, and any full-width figures or tables (see the guidelines in Subsection 3.4). **Type single-spaced.** Start

all pages directly under the top margin. See the guidelines later regarding formatting the first page.

### 3.1 Format of Electronic Manuscript

For the production of the electronic manuscript you must use Adobe's Portable Document Format (PDF). PDF files are usually produced from  $\text{\LaTeX}$  using the *pdflatex* command. If your version of  $\text{\LaTeX}$  produces Postscript files, you can convert these into PDF using *ps2pdf* or *dvipdf*. On Windows, you can also use Adobe Distiller to generate PDF.

Please make sure that your PDF file includes all the necessary fonts (especially tree diagrams, symbols, and fonts with Asian characters). When you print or create the PDF file, there is usually an option in your printer setup to include none, all or just non-standard fonts. Please make sure that you select the option of including ALL the fonts. **Before sending it, test your PDF by printing it from a computer different from the one where it was created.** Moreover, some word processors may generate very large PDF files, where each page is rendered as an image. Such images may reproduce poorly. In this case, try alternative ways to obtain the PDF. One way on some systems is to install a driver for a postscript printer, send your document to the printer specifying "Output to a file", then convert the file to PDF.

It is of utmost importance to specify the **A4 format** (21 cm x 29.7 cm) when formatting the report. When working with *dvips*, for instance, one should specify `-t a4`.

Print-outs of the PDF file on A4 report should be identical to the hardcopy version.

### 3.2 Layout

Format manuscripts two columns to a page, in the manner these instructions are formatted. The exact dimensions for a page on A4 report are:

- Left and right margins: 2.5 cm

- Top margin: 2.5 cm
- Bottom margin: 2.5 cm
- Column width: 7.7 cm
- Column height: 24.7 cm
- Gap between columns: 0.6 cm

Papers should not be submitted on any other report size, no exceptions.

### 3.3 Fonts

For reasons of uniformity, Adobe’s **Times Roman** font should be used. In  $\text{\LaTeX}2\text{e}$  this is accomplished by putting

```
\usepackage{times}
\usepackage{latexsym}
```

in the preamble. If Times Roman is unavailable, use **Computer Modern Roman** ( $\text{\LaTeX}2\text{e}$ ’s default). Note that the latter is about 10% less dense than Adobe’s Times Roman font.

Type of Text	Font Size	Style
report title	15 pt	bold
author names	12 pt	bold
the word “Abstract”	12 pt	bold
section titles	12 pt	bold
document text	11 pt	
captions	11 pt	
abstract text	10 pt	
bibliography	10 pt	
footnotes	9 pt	

Table 1: Font guide.

### 3.4 The First Page

Center the title and author’s name(s) across both columns. Do not use footnotes for affiliations. Use the two-column format only when you begin the abstract.

**Title:** Place the title centered at the top of the first page, in a 15-point bold font. (For a complete guide to font sizes and styles, see Table 1) Long titles should be typed on two lines without a blank line intervening. Approximately, put the title at 2.5 cm from the top of the page, followed by a blank line, then the author’s names(s) on the following line. Do not use only initials for given names (middle initials are allowed). Do not format

surnames in all capitals (e.g., use “Schlangen” not “SCHLANGEN”). Do not format title and section headings in all capitals as well except for proper names (such as “BLEU”) that are conventionally in all capitals. Start the body of the first page 7.5 cm from the top of the page.

**Abstract:** Type the abstract at the beginning of the first column. The width of the abstract text should be smaller than the width of the columns for the text in the body of the report by about 0.6 cm on each side. Center the word **Abstract** in a 12 point bold font above the body of the abstract. The abstract should be a concise summary of the general thesis and conclusions of the report. It should be no longer than 150 words. The abstract text should be in 10 point font.

**Text:** Begin typing the main body of the text immediately after the abstract, observing the two-column format as shown in the present document. Do not include page numbers.

**Indent** when starting a new paragraph. Use 11 points for text and subsection headings, 12 points for section headings and 15 points for the title.

### 3.5 Sections

**Headings:** Type and label section and subsection headings in the style shown on the present document. Use numbered sections (Arabic numerals) in order to facilitate cross references. Number subsections with the section number and the subsection number separated by a dot, in Arabic numerals. Do not number subsections.

**Citations:** Citations within the text appear in parentheses as (Gusfield, 1997) or, if the author’s name appears in the text itself, as Gusfield (1997). Append lowercase letters to the year in cases of ambiguity. Treat double authors as in (Aho and Ullman, 1972), but write as in (Chandra et al., 1981) when more than two authors are involved. Collapse multiple citations as in (Gusfield, 1997; Aho and Ullman, 1972). Also refrain from using full citations as sentence constituents. We suggest that instead of

“(Gusfield, 1997) showed that ...”

you use

“Gusfield (1997) showed that ...”

If you are using the provided  $\text{\LaTeX}$  and  $\text{BibTeX}$  style files, you can use the command `\newcite` to get “author (year)” citations.

**Please do not use anonymous citations** and do not include acknowledgements when submitting your reports..

**References:** Gather the full set of references together under the heading **References**. Arrange the references alphabetically by first author, rather than by order of occurrence in the text. Provide as complete a citation as possible, using a consistent format, such as the one for *Computational Linguistics* or the one in the *Publication Manual of the American Psychological Association* (American Psychological Association, 1983). Use of full names for authors rather than initials is preferred. A list of abbreviations for common computer science journals can be found in the *ACM Computing Reviews* (Association for Computing Machinery, 1983).

### 3.6 Footnotes

**Footnotes:** Put footnotes at the bottom of the page and use 9 points text. They may be numbered or referred to by asterisks or other symbols.<sup>1</sup> Footnotes should be separated from the text by a line.<sup>2</sup>

### 3.7 Graphics

**Illustrations:** Place figures, tables, and photographs in the report near where they are first discussed, rather than at the end, if possible. Wide illustrations may run across both columns.

**Captions:** Provide a caption for every illustration; number each one sequentially in the form: “Figure 1. Caption of the Figure.” “Table 1. Caption of the Table.” Type the captions of the figures and tables below the body, using 11 point text.

## 4 Analysis

### 4.1 Basic correlations

### 4.2 Correlations over time

### 4.3 Data visualization

### 4.4 Ratio experiment

### 4.5 Number of swear words experiment

### 4.6 Machine learning experiment

## 5 Outview

In this section we will look at what we could have done differently but didnt do simply because it was too complicated or it didnt make a lot of sense.

---

<sup>1</sup>This is how a footnote should appear.

<sup>2</sup>Note the line separating the footnotes from the text.

## 5.1 The hotness problem

In our dataset we have this hotness ( hotttnesss ) attribute for each song which is supposed to be a measure of popularity. However, a lot of songs have a hotness equal to 0 or NaN, which makes our final dataset smaller. But another problem is that we dont really know how this parameter was generated. In our analysis we simply trusted it without being sure how it is calculated. so we tried to come up with other ways of calculating the popularity of songs :

### 5.1.1 Youtube

The main problem is that Youtube was founded in 2005 so we would have no data for music before that year. This would leave us with a lot of songs we cant even talk about. And, since one of the core ideas of our project is to do our analysis based on time (compute the correlations between offensiveness and popularity over time), this would have left us with only a few years between 2005 and 2010 (the data stops at 2011), which is probably not enough to do get interesting data. Another issue tightly related to this one is that the popularity of Youtube itself is variable. For example Despacito is the most viewed video on Youtube today with over 4.5 billion views but, 7 years ago, it was Justin Biebers Baby with only 1.7 billion views. Its probably not that Despacito is 2.6 times more popular than Baby was but just that Youtube was not used as much as it is used today. A lot less people had a connection to the Internet and there were also other means of listening to music (there are also other means today but they are not the same as before). Simply said, there is a variable we might call Youtube usage (or poularity or acces you can call it however you want) which varies over time and we would have to scale the number of views according to this variable, which is quite difficult or maybe even non feasible.

One other problem is the way data is structured on Youtube. The way we would gather our info would be to simply try to match the names of our songs with the name on Youtube since the track id used in the Million Songs dataset simply doesnt exist on Youtube. But, however, we know there can be many videos for one song. We often have the official video and additional videos with lyrics posted by fans (and there can be many of them). So the total number of views is scattered all around and we have no way to find a guarantee that what

we are scraping is enough. We might take only official videos but even that is not such a good idea because we would probably lose a lot of info and we would also fall for fake official videos. And the view count is not a good measure because a video can be removed and reuploaded because of artist rights or other things.

### 5.1.2 Spotify

One other idea we had in mind was to use the Spotify API. Spotify is one of the most popular apps today for listening to music so it can probably be used to gather interesting information. The Spotify API is very complete since it provides us with all the values we would need in our analysis, especially with the track info such as play counts for users or popularity. However, this popularity attribute seems to vary over time because the documentation specifies that a song which has more play counts now will have a better popularity than a song which has the same number of play counts in the past. We might still use the play count or other attributes in our analysis so it would probably give some interesting results. However, the effort to put in is too high because the process for registering an app for being able to use that API seems to be quite long. It is probable that this method would not give better results than the hotness/playcounts analysis we already did. But it might be interesting for some future work.

## References

- Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.
- American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.
- Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503–512.
- Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.
- Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.