Overview of LSA and PLSA and their effectiveness in few shot learning

Kerui Zhu
keruiz2
keruiz2@illinois.edu

Abstract

Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (PLSA) are two commonly used unsupervised topic models based on term-document information. Both models can well make use of the term occurrence in each document to learn to distinguish documents into different topics. The performance benefits from the large corpus, where great amount of redundancy in terms can provide statistically distinguishable features for the model to learn. However, in real life, human can easily separate a small set of documents into several clusters. This few shot setting rises an interesting question: can the topic model captures important features in different topics by seeing only a few number of documents? In this article, we provide a general overview to the mechanism of the LSA and the PLSA model and analysis to their performance in the few shot setting.

1 Introduction

Topic modeling is a task that takes in a set of documents and returns a set of topic, where each topic is represented by a set of words that contain essential meaning about the topic. For example, topic "computer science" may be represented by terms "programming", "coding", "Java", etc. Each document may contain information of more than one topic and different topics may share some common terms. Therefore, topic model should try to find the features that can imply the topic of the document in an unsupervised manner.

To solve this problem, many useful models have been invented. The two commonly used are LSA and PLSA. This two models make use of the co-occurrence information of the terms in each document.

2 Background

In this section, we will introduce the basic mechanism of the LSA and PLSA.

2.1 LSA Model

For the LSA model, we first collect the term occurrence table across all the documents. The term occurrence table has size M $\times$ N, where M is the number of document and N is the number of terms. Then we perform Singular Value Decomposition (SVD) to the occurrence table. The SVD analysis will decompose the table as the multiplication of three matrices. The left matrix has the size of M $\times$ M, where each row represents information for a document. The right matrix has the size of N $\times$ N, where each column represents information for a term. The middle matrix has the size of M $\times$ N and is a diagonal matrix. The singular value on the diagonal line can tell the importance of each feature in the document representation and the term representation. We select the top k singular value and the corresponding features in both left and right matrices. After that, each document and term representation has k features. Since both the document representation and term representation are in the same dimension space, we can compare their similarity using cosine similarity between vectors. That means we can do clusters between document and document, term and term, and document and term. In this way, we can treat each cluster as a topic, where documents in the cluster are considered as under the same topic and the terms in the cluster are considered as the topic terms.

2.2 PLSA Model

For the PLSA model, we assume that each document may have information from more than one topic. Each word in the document may be generated by one of the topics and each topic has a probability to be chosen. The probability of all topics in a document is fixed but unknown at

the beginning. The algorithm first collects a term document table as the LSA algorithm. Then we assume that there are l topics, Z = $\{z_1, z_2, ..., z_l\}$, in total among all the documents. For each word w in document d, we have

$$P(d, w) = P(d) * P(w|d)$$

where

$$P(w|d) = \sum_z \in ZP(w|z) * P(z|d)$$

3 Evaluation on small corpus

4 Conclusion

4.1 Electronically-available resources

We strongly prefer that you prepare your PDF files using LaTeX with the official ACL 2015 style file (acl2015.sty) and bibliography style (acl.bst). These files are available at `http://acl2015.org`. You will also find the document you are currently reading (acl2015.pdf) and its LaTeX source code (acl2015.tex) on this website.

You can alternatively use Microsoft Word to produce your PDF file. In this case, we strongly recommend the use of the Word template file (acl2015.dot) on the ACL 2015 website (`http://acl2015.org`). If you have an option, we recommend that you use the LaTeX2e version. If you will be using the Microsoft Word template, we suggest that you anonymize your source file so that the pdf produced does not retain your identity. This can be done by removing any personal information from your source document properties.

4.2 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 LaTeX using the *pdflatex* command. If your version of 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 paper. When working with `dvips`, for instance, one should specify `-t a4`. Or using the command `\special{papersize=210mm,297mm}` in the latex preamble (directly below the `\usepackage` commands). Then using `dvipdf` and/or `pdflatex` which would make it easier for some.

Print-outs of the PDF file on A4 paper should be identical to the hardcopy version. If you cannot meet the above requirements about the production of your electronic submission, please contact the publication chairs as soon as possible.

4.3 Layout

Format manuscripts two columns to a page, in the manner these instructions are formatted. The exact dimensions for a page on A4 paper 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 paper size. If you cannot meet the above requirements about the production of your electronic submission, please contact the publication chairs above as soon as possible.

4.4 Fonts

For reasons of uniformity, Adobe's **Times Roman** font should be used. In LaTeX2e this is accomplished by putting

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

in the preamble. If Times Roman is unavailable, use **Computer Modern Roman** (LaTeX2e's default). Note that the latter is about 10% less dense than Adobe's Times Roman font.

Type of Text	Font Size	Style
paper title	15 pt	bold
author names	12 pt	bold
author affiliation	12 pt	
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.

4.5 The First Page

Center the title, author's name(s) and affiliation(s) across both columns. Do not use footnotes for affiliations. Do not include the paper ID number assigned during the submission process. 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), and the affiliation 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. The affiliation should contain the author's complete address, and if possible, an electronic mail address. Start the body of the first page 7.5 cm from the top of the page.

The title, author names and addresses should be completely identical to those entered to the electronical paper submission website in order to maintain the consistency of author information among all publications of the conference. If they are different, the publication chairs may resolve the difference without consulting with you; so it is in your own interest to double-check that the information is consistent.

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 paper 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 paper. It should be no longer than 200 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.

4.6 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 subsubsections.

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 LaTeX and BibTeX style files, you can use the command `\newcite` to get "author (year)" citations.

As reviewing will be double-blind, the submitted version of the papers should not include the

authors' names and affiliations. Furthermore, self-references that reveal the author's identity, e.g.,

> "We previously showed (Gusfield, 1997) ..."

should be avoided. Instead, use citations such as

> "Gusfield (1997) previously showed ... "

Please do not use anonymous citations and do not include acknowledgements when submitting your papers. Papers that do not conform to these requirements may be rejected without review.

References: Gather the full set of references together under the heading **References**; place the section before any Appendices, unless they contain 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).

The LaTeX and BibTeX style files provided roughly fit the American Psychological Association format, allowing regular citations, short citations and multiple citations as described above.

Appendices: Appendices, if any, directly follow the text and the references (but see above). Letter them in sequence and provide an informative title: **Appendix A. Title of Appendix**.

4.7 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.[1] Footnotes should be separated from the text by a line.[2]

4.8 Graphics

Illustrations: Place figures, tables, and photographs in the paper near where they are first discussed, rather than at the end, if possible. Wide illustrations may run across both columns. Color illustrations are discouraged, unless you have verified that they will be understandable when printed in black ink.

[1]This is how a footnote should appear.

[2]Note the line separating the footnotes from the text.

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.

5 XML conversion and supported LaTeX packages

Following ACL 2014 we will also we will attempt to automatically convert your LaTeX source files to publish papers in machine-readable XML with semantic markup in the ACL Anthology, in addition to the traditional PDF format. This will allow us to create, over the next few years, a growing corpus of scientific text for our own future research, and picks up on recent initiatives on converting ACL papers from earlier years to XML.

We encourage you to submit a ZIP file of your LaTeX sources along with the camera-ready version of your paper. We will then convert them to XML automatically, using the LaTeXML tool (`http://dlmf.nist.gov/LaTeXML`). LaTeXML has *bindings* for a number of LaTeX packages, including the ACL 2015 stylefile. These bindings allow LaTeXML to render the commands from these packages correctly in XML. For best results, we encourage you to use the packages that are officially supported by LaTeXML, listed at `http://dlmf.nist.gov/LaTeXML/manual/included.bindings`

6 Translation of non-English Terms

It is also advised to supplement non-English characters and terms with appropriate transliterations and/or translations since not all readers understand all such characters and terms. Inline transliteration or translation can be represented in the order of: original-form transliteration "translation".

7 Length of Submission

Long papers may consist of up to 8 pages of content, plus two extra pages for references. Short papers may consist of up to 4 pages of content, plus two extra pages for references. Papers that do not conform to the specified length and formatting requirements may be rejected without review.

Acknowledgments

The acknowledgments should go immediately before the references. Do not number the acknowl-

edgments section. Do not include this section when submitting your paper for review.

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.