
Customer Feedback Topic Modelling Using Distributed Online Latent Dirichlet Allocation

Yuhou Zhou

Department of Computer Science & Electrical Engineering
Jacobs University Bremen
Bremen, Germany
yu.zhou@jacobs-university.de

Abstract

1 Introduction

Customer feedback is information provided by clients about their general experience with a product or a service. In CEWE, one method to collect such feedback is asking customers to complete online survey.

By collecting customer feedback, the extracted information can act as a guide to improve products. Feedback gives the company the clue of customers' general experience and the satisfaction rate. Except the description provided by sellers, the publicly posted feedback is useful information to other consumers, imparting them an overview how the product really is. This information also helps managers to make correct business decisions.

The current way of processing user feedback requires that working staff reads the feedback and later extracts topics. This process can be costly and inefficient, because of manual intervention. At one time, people can only process a small batch of text from a recent period. Thus, the information, which concerning long lasting but less frequently appeared issues, may lost. This proposal will present a workflow, which can extract topics from customer feedback without manual effort.

2 Related Work

Discovering "topics" from a collection of documents has been a long-standing problem in natural language understanding. Latent Semantic Analysis (LSA) acts as the base of recent topic models [2]. Hofmann proposed Probabilistic Latent Semantic Indexing (pLSA) [3], bringing topic models from deterministic to probabilistic. In 2003, Blei described Latent Dirichlet Allocation (LDA) which is a Bayesian alternative to pLSA [1]. LDA makes use of word co-occurrence information in documents, to generate document-topic distribution.

add oneline LDA

cite distributed and multicore LDA

However, the lack of word co-occurrence in short texts prevents LDA yielding satisfying results [9]. Such short texts are news headlines, tweets, user feedback, questions and answers, and so on. One strategy to overcome this data sparsity problem is to concatenate a set of short texts according to available metadata [6, 7, 8]. For example, aggregating texts by timestamps, locations or users. The evaluation plans for topic models include perplexity, topic coherence, etc. [4, 5].

3 Approach

Data preparation will be performed to filter useless information and concatenate entries of feedback according to their metadata. Afterwards a LDA model will be trained on previous customer feedback. Topic coherence will be picked to measure the function of the model.

The tools expected to be used and compared are Google Cloud Platform (GCP) Natural Language, Python spaCy, and Gensim. GCP Natural Language provides German support and offers easy to use APIs. Spacy also supports German and is highly optimized. It has very good interfaces with other machine learning libraries, such as TensorFlow, scikit-learn. Gensim is a python library which is highly efficient and scalable towards topic and semantic modelling.

Two obstacles are about to be solved. The feedback is written in German is one of the difficulties. Nowadays, many natural language processing libraries have poor support to German language. Either they cannot train models based on German, or they only provide limited features to German. The other difficulty is the variation of text is dramatic and unpredictable. This is reflected on two aspects. First, the quality of the feedback is varying. Some feedback contains totally useless information, such as "Nein", the copy of the survey email. The length of the feedback varies from one word to a long text containing hundreds of words. Second, the content of new feedback is unforeseeable. It may include new products, new services, etc.

For model deployment, the model will be retrained based on a time period. Outdated observations will be removed, and new observations will be included into the training process. This process can be scheduled using an orchestrator tool, such as Apache Airflow.

3.1 System Pipeline

Second level headings are lower case (except for first word and proper nouns), flush left, bold and in point size 10. One line space before the second level heading and 1/2 line space after the second level heading.

3.2 LDA Model Architecture

3.2.1 Headings: third level

Third level headings are lower case (except for first word and proper nouns), flush left, bold and in point size 10. One line space before the third level heading and 1/2 line space after the third level heading.

4 Citations, figures, tables, references

These instructions apply to everyone, regardless of the formatter being used.

4.1 Citations within the text

Citations within the text should be numbered consecutively. The corresponding number is to appear enclosed in square brackets, such as [1] or [2]-[5]. The corresponding references are to be listed in the same order at the end of the paper, in the **References** section. (Note: the standard `BIBTEX` style `unsrt` produces this.) As to the format of the references themselves, any style is acceptable as long as it is used consistently.

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]", not "In our previous work [4]". If you cite your other papers that are not widely available (e.g. a journal paper under review), use anonymous author names in the citation, e.g. an author of the form "A. Anonymous".

Table 1: Sample table title

PART	DESCRIPTION
Dendrite	Input terminal
Axon	Output terminal
Soma	Cell body (contains cell nucleus)

4.2 Footnotes

Indicate footnotes with a number¹ in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).²

4.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction; art work should not be hand-drawn. The figure number and caption always appear after the figure. Place one line space before the figure caption, and one line space after the figure. The figure caption is lower case (except for first word and proper nouns); figures are numbered consecutively.

Make sure the figure caption does not get separated from the figure. Leave sufficient space to avoid splitting the figure and figure caption.

You may use color figures. However, it is best for the figure captions and the paper body to make sense if the paper is printed either in black/white or in color.



Figure 1: Sample figure caption.

4.4 Tables

All tables must be centered, neat, clean and legible. Do not use hand-drawn tables. The table number and title always appear before the table. See Table 1.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

5 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

¹Sample of the first footnote

²Sample of the second footnote

6 Preparing PostScript or PDF files

Please prepare PostScript or PDF files with paper size “US Letter”, and not, for example, “A4”. The `-t letter` option on `dvips` will produce US Letter files.

Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or Embedded TrueType fonts. Here are a few instructions to achieve this.

- You can check which fonts a PDF file uses. In Acrobat Reader, select the menu `Files>Document Properties>Fonts` and select `Show All Fonts`. You can also use the program `pdf fonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- The IEEE has recommendations for generating PDF files whose fonts are also acceptable for NIPS. Please see <http://www.emfield.org/icuwb2010/downloads/IEEE-PDF-SpecV32.pdf>
- LaTeX users:
 - Consider directly generating PDF files using `pdflatex` (especially if you are a MiKTeX user). PDF figures must be substituted for EPS figures, however.
 - Otherwise, please generate your PostScript and PDF files with the following commands:

```
dvips mypaper.dvi -t letter -Ppdf -G0 -o mypaper.ps
ps2pdf mypaper.ps mypaper.pdf
```

Check that the PDF file only contains Type 1 fonts.
 - `xfig` “patterned” shapes are implemented with bitmap fonts. Use “solid” shapes instead.
 - The `\bbold` package almost always uses bitmap fonts. You can try the equivalent AMS Fonts with command

```
\usepackage[psamsfonts]{amssymb}
```

or use the following workaround for reals, natural and complex:

```
\newcommand{\RR}{I\!\!R} %real numbers
\newcommand{\Nat}{I\!\!N} %natural numbers
\newcommand{\CC}{I\!\!C} %complex numbers
```
 - Sometimes the problematic fonts are used in figures included in LaTeX files. The ghostscript program `eps2eps` is the simplest way to clean such figures. For black and white figures, slightly better results can be achieved with program `potrace`.
- MSWord and Windows users (via PDF file):
 - Install the Microsoft Save as PDF Office 2007 Add-in from <http://www.microsoft.com/downloads/details.aspx?displaylang=en&familyid=4d951911-3e7e-4ae6-b059-a2e79ed87041>
 - Select “Save or Publish to PDF” from the Office or File menu
- MSWord and Mac OS X users (via PDF file):
 - From the print menu, click the PDF drop-down box, and select “Save as PDF...”
- MSWord and Windows users (via PS file):
 - To create a new printer on your computer, install the AdobePS printer driver and the Adobe Distiller PPD file from <http://www.adobe.com/support/downloads/detail.jsp?ftpID=204> *Note:* You must reboot your PC after installing the AdobePS driver for it to take effect.
 - To produce the ps file, select “Print” from the MS app, choose the installed AdobePS printer, click on “Properties”, click on “Advanced.”
 - Set “TrueType Font” to be “Download as Softfont”
 - Open the “PostScript Options” folder
 - Select “PostScript Output Option” to be “Optimize for Portability”
 - Select “TrueType Font Download Option” to be “Outline”

- Select “Send PostScript Error Handler” to be “No”
- Click “OK” three times, print your file.
- Now, use Adobe Acrobat Distiller or ps2pdf to create a PDF file from the PS file. In Acrobat, check the option “Embed all fonts” if applicable.

If your file contains Type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

6.1 Margins in LaTeX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the `graphicx` package. Always specify the figure width as a multiple of the line width as in the example below using `.eps` graphics

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.eps}
```

or

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

for `.pdf` graphics. See section 4.4 in the graphics bundle documentation (<http://www.ctan.org/tex-archive/macros/latex/required/graphics/grfguide.ps>)

A number of width problems arise when LaTeX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

References

References follow the acknowledgments. Use unnumbered third level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to ‘small’ (9-point) when listing the references. **Remember that this year you can use a ninth page as long as it contains *only* cited references.**

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauero, D. S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609-616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SIMulation System*. New York: TELOS/Springer-Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.