Latent Dirichlet Allocation Models Considering Emojis

Taikgun Song

### abstract

XXX write later XXX

# Introduction

Text data contains valuable insights that may be useful for content recommendation, customer care service, social media analysis, and . However, these information are usually hidden underneath the plain text. Topic modeling is a text-mining method that extracts information from a text data by identifying latent semantic structures in the text body. One of the most widely used as a topic modeling method is the Latent Dirichlet Allocation(LDA). LDA is a popular hierarchical Bayesian model which assumes that each of the documents in a collection consist of a mixture of topics, and these topics are reponsible for the establishment of words in each document. Topics, however, are the latent part of the document set and one can only observe words collected into documents. LDA exploits statistical inference to discover structure given the words and documents by calculating the relative importance of topics in documents and words in topics.

The rapid growth in internet and telecommunication technology triggered the development of Social Network Services(SNS) platform such as Tweeter, Facebook, and blog posts. The SNS messages often include individual's perceptions, feelings, and opinions. Therefore, evaluating this primary data may be meaningful for policy makers, social science researchers, and business entrepreneurs. This electronic word-of-mouth heavily uses text data as the medium of communication. Thus, topic modeling including LDA may be ideal method for analyzing SNS text data for information retrieval tasks.

The use of emoji - a pictogram that expresses the author's feeling and emotion - mixed in with other text is a unique characteristic of SNS messages that distinguishes itself from other text data. As shown in , many SNS messages can be found with emoji embedded in the content. Conventionally, emoji characters have been considered as a noise and were deleted prior to applying LDA techniques and other topic modeling methods. Nevertheless, one should focus on the richness of information that emoji characters can provide. Especially consider the emotional and symbolic representation of emoji that cannot be better expressed with alphabet characters. Therefore, in contrast to the typical topic modeling procedure, this paper propose the idea of incorporating emoji characters to enhance the performance of the LDA method on SNS text data.

The use of emoji characters have three main benefits. First, it may reduce the systematic problem of LDA with data sparcity. All emoji characters have name and keywords associated with the contextual meaning that it conveys. By translating emoji characters into its English name or related keywords will increase the observation, and thus lead to better LDA results. Second, each emoji character has a couple of pre-determined topic dimension set by the official organization. This information could be used as an auxiliary information during the topic matching process. Lastly, emoji character itself is an abstract of emotion and symbolic representation. Thus, it is natural to take the output of LDA containing emoji translation to sentiment analysis.

# LDA

Let be the word in the document. We assume that the topic of is , a topic associated with document . Assume $z\_m \sim Multinomial(\boldsymbol\theta\_m)$, where $\boldsymbol\theta\_m \sim Dirichlet(\boldsymbol\alpha)$ for all and . For a given topic , we assume that $w\_{mn} \sim Multinomial(\boldsymbol\phi\_k), n=1, \dots, n\_m, m=1, \dots, M$, where $\boldsymbol\phi\_k \sim Dirichlet(\boldsymbol\beta)$, .

The summarization of the assumptions are written below.

The graphical display of LDA is given in .

Then, the total probability of the model is

The marginal distribution of word w given hyper parameter and is obtained by the following equation:

On the application perspective, solving the posterior distribution is the key which is given as the following equation.

$$p(\boldsymbol\theta, \textbf{z} | \textbf{w}, \alpha, \beta)= \frac{p(\theta, z, w | \alpha, \beta)}{p(w|\alpha, \beta)}$$

However, this posterior distribution is intractable for exact inference.

# Data preparation

## Removing Stop Words

A natural language can be categorized as two distinctive set of words: content/lexical words and function/structure words. Content/lexical words are words with substantive meanings. Function/structure words on the other hand have little lexical meaning, but establish grammatical structure between other words within a sentence.

LDA models a document as a mixture of topics, and then each word is drawn from one of its topic. Therefore, the method depends on the frequency of observed words in a given text data set. This makes LDA method vulnerable when meaningless words such as function/structural words are present in the data set with high frequency. Thus, any group of non-informative words including the function/structural words should be filtered out before doing an analysis, and this group of words are called the stop words. For example, prepositions(of, at, in, without, between), determiners(the, a, that, my), conjunctions(and, that, when), pronouns(he, they, anybody, it) are common examples of the stop words. For the work done in the paper, the tm package in R was used to delete the stop words.

## Stemming

Due to structural and grammatical reasons of English, a family of words that are driven from a single root word is used in different forms. For example, words such as "stems", "stemmer", "stemming", and "stemmed" are all based on a root word "stem". Words with same meaning but different in forms contribute to data sparsity, reducing the performance of the LDA method. The stemming procedure cuts inflectional forms of a word to its root form eventually increasing the frequency of word observations.

The stemming process has two disadvantages. First, there are possibility of over stemming. For example, three different words "universal", "university", and "universe" have the same stemmed word "univers". The accuracy of the LDA method may decrease by putting words with different meanings into a single topic. Moreover, when the LDA output is given as a stemmed word, it is difficult to trace the stemmed word to its original form. To overcome this problem, this paper matched the stemmed word to the most frequently used original word. Example of stemming using the tm is provided in .

\ \

## n-gram

n-gram is a neighboring sequence of n items from a collection of text data set. This item could be anything from phonemes or syllables to letters or words based on the application. Applying the concept of n-gram is important in computational linguistics

An example of word-level-n-gram with text "he is a nice person" is given in .

# Application

## Data Set and exploratory data analysis

$\\$

Two samples of twitter messages with the following hash-tag #inlove and #hateher were scraped. The data set contains 944 #inlove messages, 1145 #hateher messages, and 1195 #marchscience messages. The proportion of Twitter messages containing emoji characters per hashtag is illustrated in . 52.7% of the #inlove tweets, 29.3% of the #hateher tweets, and 7.8% of #marchscience tweets make use of one or more emojis.

Proportion of Twitter messages with emoji

|  |  |  |  |
| --- | --- | --- | --- |
|  | #inlove | #hateher | #marchscience |
| **Proportion** | 0.5275 | 0.2926 | 0.07782 |

For the hashtag #inlove, a total number of 1188 emojis were used, consisting of 182 unique emojis. For hashtag #hateher, 695 emojis from 112 unique emojis were used. For hashtag #sciencemarch, 202 emojis from 102 unique emojis were used (Note that there may be multiple emojis per Twitter message). Top 5 frequently used emojis per hashtag is given in .

It is interesting to see "Face with tear of joy" as the most popular emoji for hashtag #hateher. Although the name itself contains the word "joy", some users of this emoji adopted this pictogram to express their mixed feeling of love and hate at the same time.

## Results

LDA was performed on the following three difference cases:

## LDA on a raw data set

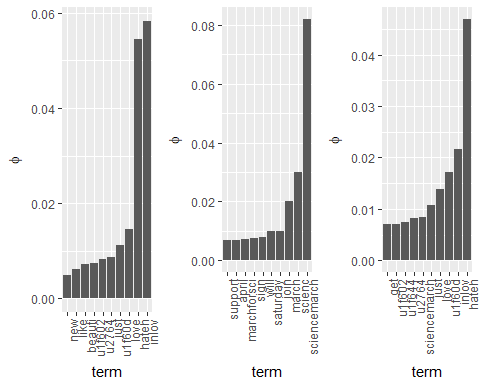
The second case was to run LDA on a raw data set. Stemming and stop word deletion were performed. Different number of topic dimensions were tested and the result of 4 topic dimension with 10 terms are provided in . Describe the output.

Output LDA with the raw data

|  |  |  |
| --- | --- | --- |
| Topic 1 | Topic 2 | Topic 3 |
| inlov | sciencemarch | hateh |
| hateh | scienc | inlov |
| love | march | u1f60d |
| u1f60d | join | love |
| just | saturday | just |
| u2764 | will | sciencemarch |
| u1f602 | sign | u2764 |
| beauti | marchforsci | u1f644 |
| like | april | u1f602 |
| new | support | get |

Word prob. given topic

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1.term | 1.phi | 2.term | 2.phi | 3.term | 3.phi |
| inlov | 0.05835 | sciencemarch | 0.08207 | hateh | 0.04696 |
| hateh | 0.05447 | scienc | 0.03015 | inlov | 0.02171 |
| love | 0.01456 | march | 0.02004 | u1f60d | 0.01725 |
| u1f60d | 0.01129 | join | 0.009815 | love | 0.01386 |
| just | 0.008727 | saturday | 0.009815 | just | 0.01077 |
| u2764 | 0.008333 | will | 0.00774 | sciencemarch | 0.00839 |
| u1f602 | 0.007405 | sign | 0.007646 | u2764 | 0.008208 |
| beauti | 0.007329 | marchforsci | 0.007232 | u1f644 | 0.007431 |
| like | 0.00617 | april | 0.006819 | u1f602 | 0.007126 |
| new | 0.004958 | support | 0.006819 | get | 0.006999 |



## LDA without Unicode

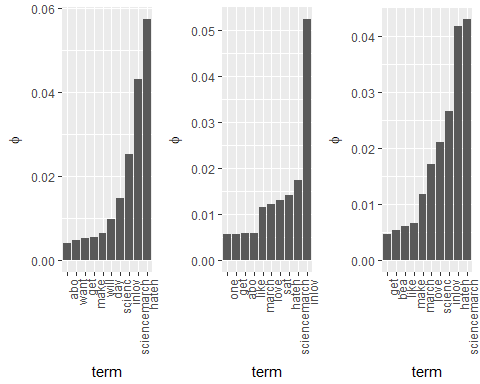
In most text mining examples, LDA is performed after removing the Unicode information. For the first case, therefore, Unicode characters were removed from the raw text data set. Then, the standard procedure of stemming and stop word deletion was performed to enhance the accuracy of LDA. tm package was used to conduct the above procedure.

Output of LDA with the raw data without the Unicode

|  |  |  |
| --- | --- | --- |
| Topic 1 | Topic 2 | Topic 3 |
| hateh | inlov | sciencemarch |
| sciencemarch | sciencemarch | hateh |
| inlov | hateh | inlov |
| scienc | sat | scienc |
| day | love | love |

Word prob. given topic

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1.term | 1.phi | 2.term | 2.phi | 3.term | 3.phi |
| hateh | 0.05744 | inlov | 0.05235 | sciencemarch | 0.04296 |
| sciencemarch | 0.04302 | sciencemarch | 0.01736 | hateh | 0.04164 |
| inlov | 0.02517 | hateh | 0.01418 | inlov | 0.0266 |
| scienc | 0.01491 | sat | 0.01315 | scienc | 0.02107 |
| day | 0.009753 | love | 0.0123 | love | 0.01708 |
| will | 0.006564 | march | 0.01165 | march | 0.01188 |
| make | 0.005566 | like | 0.005814 | make | 0.006575 |
| get | 0.005268 | abo | 0.005804 | like | 0.006116 |
| want | 0.004803 | get | 0.005786 | bea | 0.005424 |
| abo | 0.004104 | one | 0.005752 | get | 0.004701 |



## LDA with name translated

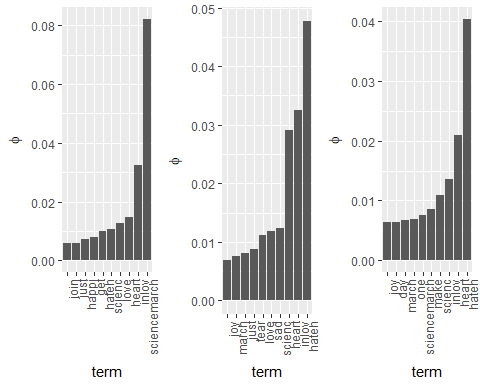
The last case was to perform LDA after translating the Unicode emoji characters in English. unicode package was used to match the Unicode to its name. Then the standard process of stemming and deletion of stop words where performed.

Output of LDA with translated Unicode

|  |  |  |
| --- | --- | --- |
| Topic 1 | Topic 2 | Topic 3 |
| sciencemarch | hateh | hateh |
| inlov | inlov | heart |
| heart | heart | inlov |
| love | scienc | scienc |
| scienc | sad | make |

Word prob. given topic

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1.term | 1.phi | 2.term | 2.phi | 3.term | 3.phi |
| sciencemarch | 0.0822 | hateh | 0.04776 | hateh | 0.04039 |
| inlov | 0.03252 | inlov | 0.03253 | heart | 0.02099 |
| heart | 0.01474 | heart | 0.02906 | inlov | 0.01367 |
| love | 0.01284 | scienc | 0.01242 | scienc | 0.01087 |
| scienc | 0.01065 | sad | 0.01179 | make | 0.008641 |
| hateh | 0.009862 | love | 0.01119 | sciencemarch | 0.007538 |
| get | 0.008055 | tear | 0.008803 | one | 0.006861 |
| happi | 0.007167 | just | 0.008046 | march | 0.006769 |
| just | 0.005986 | march | 0.007515 | day | 0.006423 |
| join | 0.005945 | joy | 0.006964 | joy | 0.006384 |



# Conclusion

As the result of the exploratory analysis indicates, user-generated-contents may contain Unicode emoji characters. These emoji characters sometimes carry mixture of condensed information that is difficult to express in words. The result of the output from the LDA indicates that words such as "heart" that would have been neglected using the traditional method may be saved when the Unicode characters are translated into meanings.

# Appendix

# emoji package in R

## Description of the emoji package

The emoji package contains information of the emoji v5.0 from its official publisher the . The illustration of the web page is shown in .

The data set emoji in the emoji package contains 8 variables:

uni\_no: Official number of emojis$\\$ uni\_code: Formal Unicode of emojis$\\$ uni\_name: Official name of emojis$\\$ cat1: Official category of emojis$\\$ cat2: Official sub-category of emojis from cat1$\\$ cat3: Official sub-category of emojis from cat2$\\$ uni\_keyws: Official keyword(s) of emojis$\\$ uni\_png: Image of emojis in PNG format represented in a matrix format$\\$

The package has a function emoji\_info\_table that summarizes all emoji and their information used in a single character string.

## Scoring of Sentiment

The characteristic of emoji (effectively delivers feelings and moods), naturally leads text mining with emoji to sentiment analysis. tidytext package in R has three general purpose lexicon sets. The AFINN score words from -5 to 5 scale, bing assigns words in binary category(positive and negative), and nrcassigns words with more categories.

# More work

1. Technical details The tm, topicmodels, emoji, tidytext, and tidyverse package in R was written to help the above analysis.
2. Check Stemming - scienc vs. science
3. Check output again. Also, a check aggregation of short messages to avoid data sparsity.
4. LDA explanation
5. Description of the emoji package