

Latent Dirichlet Allocation Models Considering Emojis

Taigun Song

abstract

XXX write later XXX

Contents

abstract	1
Introduction	1
Latent Dirichlet Allocation (LDA)	2
Data preparation	4
Removing Stop Words	4
Stemming	4
n-gram	5
Application	5
Data Set and exploratory data analysis	5
Results	6
LDA on a raw data set	6
LDA without Unicode	7
LDA with name translated	8
Conclusion	9
Appendix	9
emoji package in R	9
Description of the emoji package	9
Scoring of Sentiment	10
More work	10

Introduction

Text data contains valuable insights that is useful for content recommendation, customer care service, social media analysis, and others. However, the information is usually hidden within the text and has to be extracted using a modeling approach. Topic modeling is a text-mining method that extracts information from a text by identifying latent semantic structures in the text body. One of the most widely used topic modeling methods is the Latent Dirichlet Allocation(LDA). LDA is a hierarchical Bayesian model which assumes that each of the documents in a collection consists of a mixture of topics, and these topics are responsible for the choice of words in each document. Topics, are the latent part of the document set and one can only observe words collected in the documents. LDA uses 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. Evaluating this 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 Figure 1, 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 proposes the idea of incorporating emoji characters to enhance the performance of the LDA method on SNS text data.



Figure 1: Example of Twitter Messages

The use of emoji characters has three main benefits. First, it reduces the systematic problem of LDA with data sparsity. All emoji characters have name and keywords associated with the contextual meaning that it conveys. By translating emoji characters into English text and related keywords increases the amount of the text observed, and thus leads to better LDA results. Second, each emoji character has a set of pre-determined topic dimension assigned to it by the official organization. This information can be used as auxiliary information during the topic matching process. Lastly, the 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.

Latent Dirichlet Allocation (LDA)

At its core Latent Dirichlet Allocation is a generative statistical model, that identifies posterior probabilities of words belonging to previously unidentified topics, and topics belonging to documents. The underlying model is generally not analytically tractable, but uses a Gibbs sampling approach instead. Here, we are deriving the posterior distributions involved in more detail. A graphical overview of LDA is given in Figure 2.

Let M be the total number of documents in the data set, and N_m be the number of words in the m^{th} document. Let K be the total number of topics in the data. Define w_{mn} be the n^{th} word in the m^{th} document. LDA assumes the distribution of w_{mn} to follow a Multinomial distribution with parameter $\phi_{z,w}$. $\phi_{z,w}$ is a probability of observing word w in topic z . The model assumes that the distribution of words in topic z , i.e., ϕ_z , follows a Dirichlet distribution with prior $\beta = [\beta_1 \cdots \beta_N]$. Let z_{mn} be the topic of assigned to the word w_{mn} . Then, the model assumes z_m to follow a Multinomial distribution with parameter θ_m , where θ_m is the distribution of topics in document m . The distribution of θ_m is assumed to follow a Dirichlet distribution with a prior $\alpha = [\alpha_1 \cdots \alpha_K]$.

I would like to keep the below paragram and bullet points for now.

Let w_{mn} be the n^{th} word in the m^{th} document. We assume that the topic of w_{mn} is z_m , a topic associated with document m . Assume $z_m \sim \text{Multinomial}(\theta_m)$, where $\theta_m \sim \text{Dirichlet}(\alpha)$ for all $m = 1, \dots, M$ and $\alpha > 0$. For a given topic $z_m = k$, we assume that $w_{mn} \sim \text{Multinomial}(\phi_k)$, $n = 1, \dots, n_m$, $m = 1, \dots, M$, where $\phi_k \sim \text{Dirichlet}(\beta)$, $k = 1, \dots, K$.

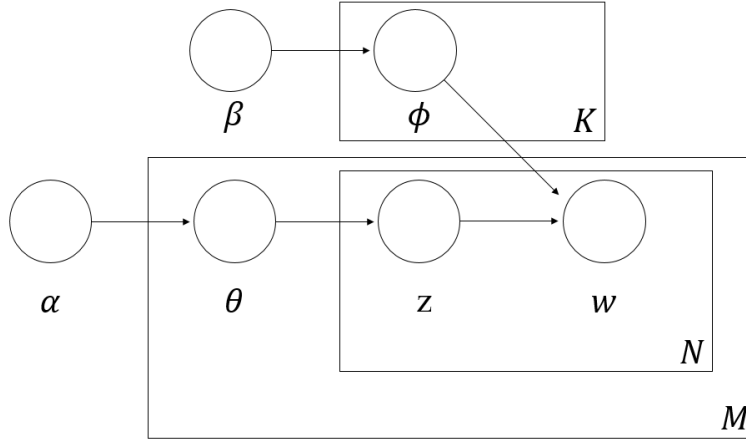


Figure 2: Graphical Model representation of LDA

The summarization of the assumptions are written below.

1. M : The total number of documents in the data set
2. N_m : The number of words in the m^{th} document
3. K : The total number of topics in the data set
4. w_{mn} : n^{th} word in document m , $m \in \{1, \dots, M\}$ and $n \in \{1, \dots, N_m\}$
5. z_{mn} : The topic of the w_{mn} , $z_{mn} \in \{1, \dots, K\}$

6. α : A vector of prior weights for each topic in a document
 $\alpha = [\alpha_1 \dots \alpha_K]$

7. $\theta_{m,k}$: The probability of observing topic k in document m
 $\theta_m \sim Dir(\alpha)$: The distribution of topics in document m

$$\theta_{M \times K} = \begin{bmatrix} \theta_1 = (\theta_{1,1}, \theta_{1,2}, \dots, \theta_{1,K}) \\ \theta_2 = (\theta_{2,1}, \theta_{2,2}, \dots, \theta_{2,K}) \\ \vdots \\ \theta_M \end{bmatrix}$$

8. β : A vector of prior weights of the word distribution for each topic
 $\beta = [\beta_1 \dots \beta_N]$

9. $\phi_{z,w}$: The probability of observing word w in topic z
 $\phi_z \sim Dir(\beta)$: The distribution of words in topic z

$$\phi_{K \times N} = \begin{bmatrix} \phi_1 = (\phi_{1,1}, \phi_{1,2}, \dots, \phi_{1,N}) \\ \phi_2 = (\phi_{2,1}, \phi_{2,2}, \dots, \phi_{2,N}) \\ \vdots \\ \phi_K \end{bmatrix}$$

10. $z_{mn} \sim Multinomial(\theta_m)$

11. $w_{mn} \sim Multinomial(\phi_{z_{mn}})$

Then, the total probability of the model is given as the product of the conditional probabilities

$$p(W, Z, \theta; \phi, \alpha, \beta) = \prod_{i=1}^K P(\phi_i; \beta) \prod_{j=1}^M P(\theta_j; \alpha) \prod_{t=1}^N P(Z_{j,t} | \theta_j) P(W_{j,t} | \phi_{Z_{j,t}})$$

The marginal distribution of word w given hyper parameter α and β is then obtained by integrating the below equation:

$$p(w | \alpha, \beta) = \int p(\theta | \alpha) \left(\prod_{v=1}^V \sum_{z_v} p(z_v | \theta) p(w_v | z_v, \beta) \right) d\theta$$

The posterior distribution is given as the following equation, however, it is intractable for exact inference and Gibbs sampling is used to infer the variables.

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

Data preparation

data comes with text format. need data preparation.

stop words, stemming, ... should go into a separate section called '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 vulnerable to high frequency function/structural words. Thus, any group of non-informative words including the function/structural words should be filtered out before doing an analysis. This group of words is called **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 analysis done here, the **tm** package in **R** was used to delete the stop words.

give some examples following the tweets or again go back to the xkcd example

	Original Tweet	Tweet with Stopword Removed
1	loving this misty weather this sweater and my favorite couple	loving misty weather, sweater favorite couple
2	fairytale atmosphere in alberobello Let's go for a walk	fairytale atmosphere alberobello Let's go walk
3	Me when ashleytisdale puts a New music session on YouTube	Me ashleytisdale puts New music session YouTube

Table 1: Example of removing stop words using the Twitter data

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 the root “stem”. Words with the same meaning but different forms contribute to data sparsity, reducing the performance of the LDA method. **Stemming** cuts inflectional forms of a word to its root form and increases the frequency of observed stems.

Stemming has two disadvantages. First, there is the 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 back 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 Table 2.

Is there something we can do about overstemming?

include a couple of examples Example of stemming provided below

	Original Tweet	Tweet after Stemming
1	loving this misty weather, this sweater and my favorite couple	love this misti weather, this sweater and my favorit coupl
2	fairytale atmosphere in alberobello Let's go for a walk	fairytal atmospher in alberobello Let go for a walk
3	Me when ashleytisdale puts a New music session on YouTube	Me when ashleytisdal put a New music session on YouTub

Table 2: Before and after Stemming

n-grams and just generally features of documents\ feature extraction. n-grams is just one of them and I have used uni-gram. justify why.\

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 is important especially with LDA, since n-gram is used as part of the prior distribution.

An example of word-level-n-gram with text “he is a nice person” is given in Table 3.

1-gram (unigram)	2-gram	3-gram	4-gram	5-gram
he	he is	he is a	he is a nice	he is a nice person
is	is a	is a nice	is a nice person	
a	a nice	a nice person		
nice	nice person			
person				

Table 3: Example of word-level-n-gram

Moreover, n-gram approach can help identify misspelled words or out-of-vocabulary words that commonly exist on the online platform. For example, the distance of the letter-level n-gram could be used to match strings.

Application

Data Set and exploratory data analysis

more info on the data: use dates - should we wrap this into a shiny app down the road?

Shiny had a problem with instant web scraps last year. I am not certain if that problem is fixed now.

Update the the dataset? I can always scrape a new sets of data and reflect the dates information. 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 Table 4. 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.

Table 4: 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 Table 5.
















#inlove	emoji	Count	#hateher	emoji	Count	#marchscience	emoji	Count
U+1F60D		297	U+1F602		154	U+1F52C		13
U+2764		164	U+1F644		88	U+1F30E		11
U+1F495		47	U+1F621		40	U+1F44D		9
U+1F618		40	U+1F612		38	U+1F680		8
U+2728		26	U+1F62D		36	U+1F30D		7

Table 5: Five most popular emoji for each hashtag

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:

1. LDA on a raw data set
2. LDA on a data set with Unicode removed
3. LDA on a data set with emoji translated to text

LDA on a raw data set

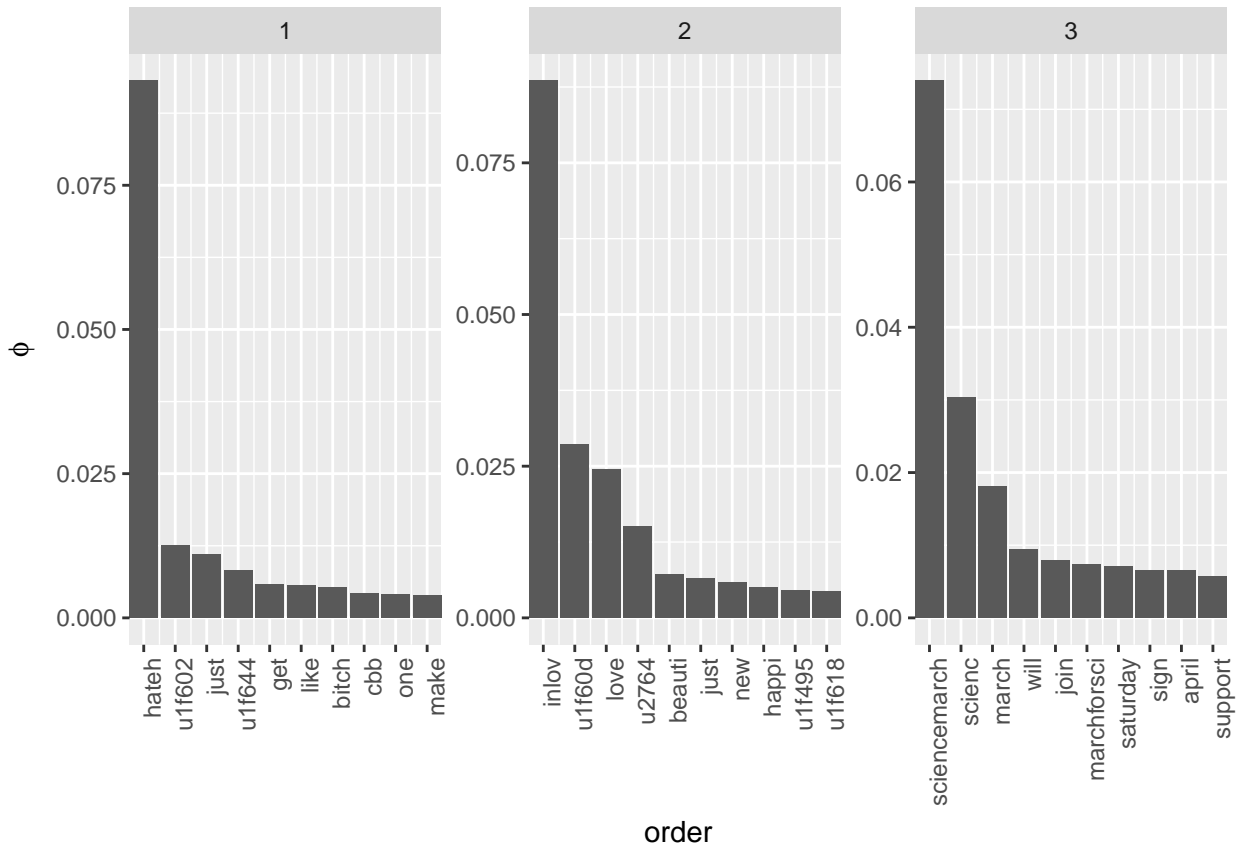
The second case was to run LDA on a raw data set. Stemming and stop word deletion were performed. Different number of topic dimensions were tested and the result of 4 topic dimension with 10 terms are provided in Table 6. Describe the output.

Table 6: Output LDA with the raw data

Topic 1	Topic 2	Topic 3
hateh	inlov	sciencemarch
u1f602	u1f60d	scienc
just	love	march
u1f644	u2764	will
get	beauti	join
like	just	marchforsci
bitch	new	saturday
cbb	happi	sign
one	u1f495	april
make	u1f618	support

Table 7: Word prob. given topic

1.term	1.phi	2.term	2.phi	3.term	3.phi
hateh	0.09315	inlov	0.08856	sciencemarch	0.07395
u1f602	0.01253	u1f60d	0.02867	scienc	0.03033
just	0.011	love	0.02449	march	0.0181
u1f644	0.008171	u2764	0.0151	will	0.009414
get	0.005883	beauti	0.007232	join	0.00795
like	0.005556	just	0.00647	marchforsci	0.007322
bitch	0.005229	new	0.005836	saturday	0.007113
cbb	0.004249	happi	0.004948	sign	0.00659
one	0.004031	u1f495	0.004567	april	0.006485
make	0.003922	u1f618	0.004314	support	0.005649



LDA without Unicode

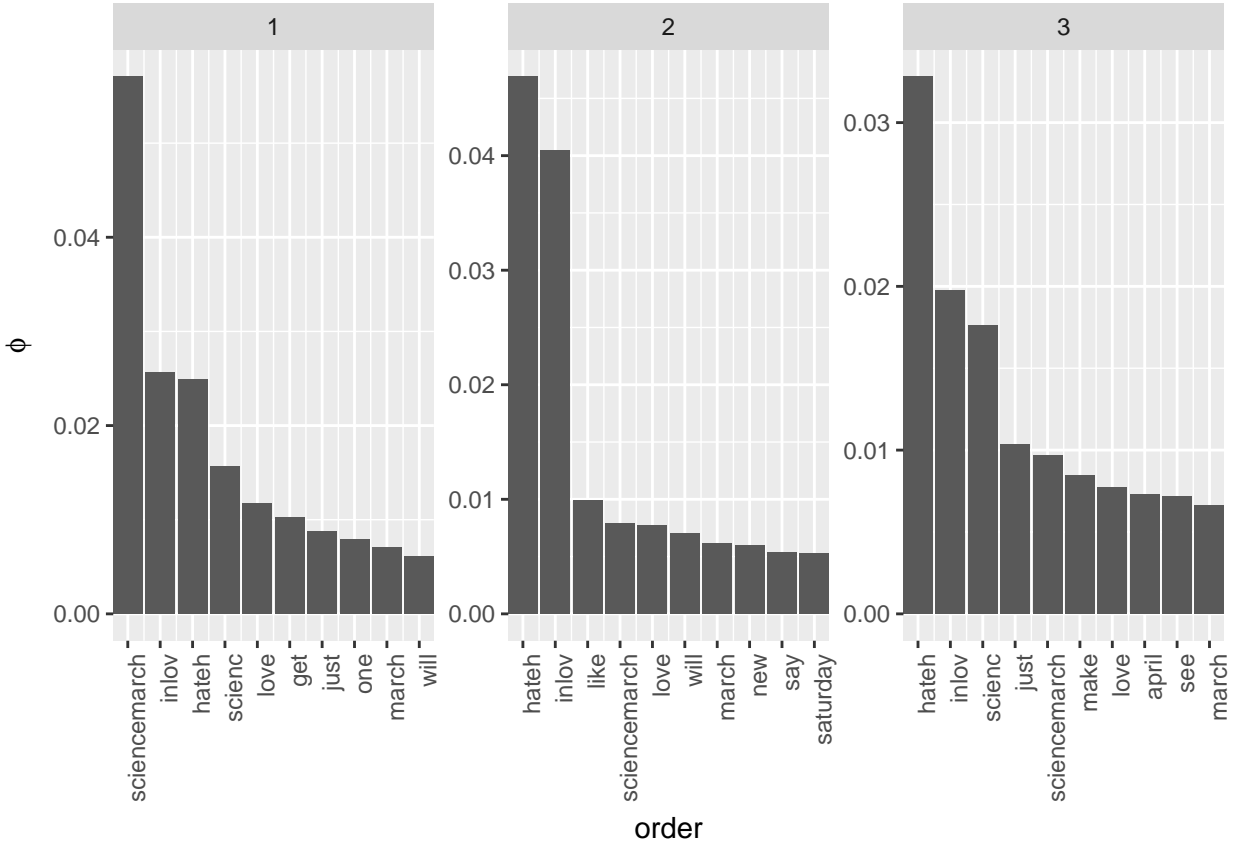
In most text mining examples, LDA is performed after removing the Unicode information. For the first case, therefore, Unicode characters were removed from the raw text data set. Then, the standard procedure of stemming and stop word deletion was performed to enhance the accuracy of LDA. `tm` package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

Table 8: Output of LDA with the raw data without the Unicode

Topic 1	Topic 2	Topic 3
sciencemarch	hateh	hateh
inlov	inlov	inlov
hateh	like	scienc
scienc	sciencemarch	just
love	love	sciencemarch

Table 9: Word prob. given topic

1.term	1.phi	2.term	2.phi	3.term	3.phi
sciencemarch	0.05707	hateh	0.0469	hateh	0.03281
inlov	0.02568	inlov	0.04045	inlov	0.01975
hateh	0.02496	like	0.009897	scienc	0.01763
scienc	0.01562	sciencemarch	0.007883	just	0.01038
love	0.01171	love	0.007728	sciencemarch	0.009702
get	0.01029	will	0.007053	make	0.008466
just	0.008804	march	0.006155	love	0.007715
one	0.007862	new	0.005943	april	0.007321
march	0.007064	say	0.005355	see	0.00717
will	0.006142	saturday	0.005299	march	0.006609



LDA with name translated

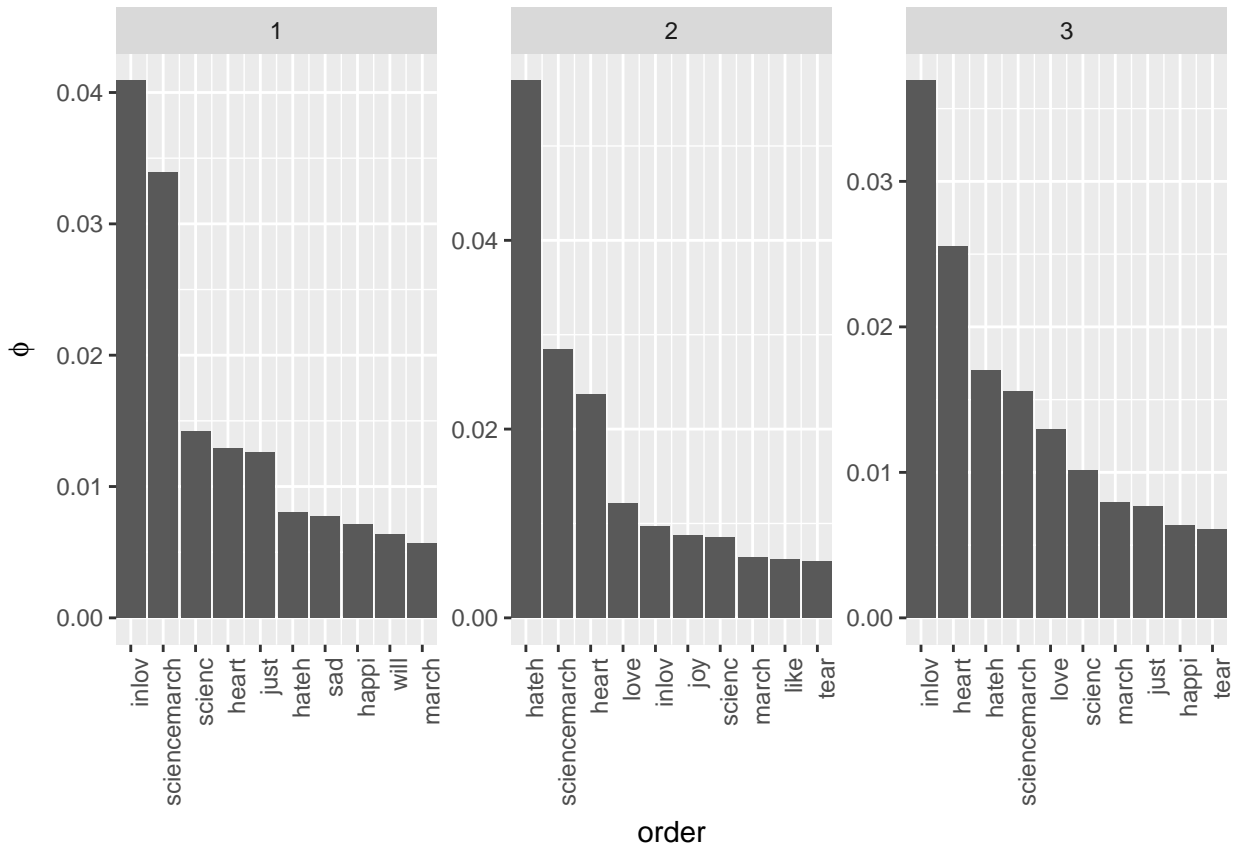
The last case was to perform LDA after translating the Unicode emoji characters in English. `unicode` package was used to match the Unicode to its name. Then the standard process of stemming and deletion of stop words where performed.

Table 10: Output of LDA with translated Unicode

Topic 1	Topic 2	Topic 3
inlov	hateh	inlov
sciencemarch	sciencemarch	heart
scienc	heart	hateh
heart	love	sciencemarch
just	inlov	love

Table 11: Word prob. given topic

1.term	1.phi	2.term	2.phi	3.term	3.phi
inlov	0.04091	hateh	0.05693	inlov	0.03693
sciencemarch	0.03393	sciencemarch	0.02845	heart	0.0255
scienc	0.0142	heart	0.02368	hateh	0.01702
heart	0.01292	love	0.01218	sciencemarch	0.0156
just	0.0126	inlov	0.009716	love	0.01298
hateh	0.008063	joy	0.008755	scienc	0.01017
sad	0.007761	scienc	0.008572	march	0.007957
happi	0.007088	march	0.006432	just	0.007692
will	0.006396	like	0.006243	happi	0.006373
march	0.005702	tear	0.00594	tear	0.006117



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

Plan to change this part after posting the emoji package on CRAN

Description of the emoji package

The `emoji` package contains information of the emoji v5.0 from its official publisher the Unicode Consortium. The illustration of the web page is shown in Figure 3.

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

[Email Charts](#)

Full Emoji List, v5.0

[Index & Help](#) |
 [Images & Rights](#) |
 [Spec](#) |
 [Proposing Additions](#)

This chart provides a list of the Unicode emoji characters and sequences, with images from different vendors, CLDR name, date, source, and keywords. The ordering of the emoji and the annotations are based on [Unicode CLDR data](#). Emoji sequences have more than one code point in the **Code** column. New characters show as a group with "..." before and after.

While these charts use a particular version of the [Unicode Emoji data files](#), the images and format may be updated at any time. For any production usage, those data files should be consulted. For more information, see [Index & Help](#).

Smileys & People															
face-positive															
No	Code	Browser	Apple	Google	Twtr.	One	FB	FBM	Sams.	Wind.	GMail	SB	DCM	KDDI	CLDR Short Name
1	U+1F600														grinning face
2	U+1F601														beaming face with smiling eyes
3	U+1F602														face with tears of joy
4	U+1F603														rolling on the floor laughing

Figure 3: Glimpse of the table of emoji on the Unicode.org website

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 `nrc` assigns words with more categories.

More work

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