Creative Component

Taikgun Song

Abstract

Write abstract. It is known that the performance of Laten Dirichlet Allocation based topic models over short texts. In this paper, we would like to compare LDA methods with different 'parameter' under experimental setting.

1 Introduction

literature ~~to~~ Review. AIM of LDA.

Then the following R packages were utilized to conduct `Latent Dirichlet Allocation`(Blei, Ng, and Jordan 2003) method: `openNLP`(Hornik 2016a), `NLP` (Hornik 2016b), `topicmodels` (Bettina Grün 2016).

2 Data Cleaning

Datasets. In order to compare different LDA processes, a collection of short user-generated online reviews from a popular website, the Trip Advisor, was used for evaluation. Among all restaurants in Honolulu, Hawaii registered on the Trip Advisor, the latest 10 reviews were scrapped and read into R (R Core Team 2014) using the `RCurl` (Lang 2015) package. This dataset includes the following information of the 7700 Honolulu restaurants: restaurant name, number total reviews, average star rating, individual review title, individual review entry, individual star rating, and the date visited. In this paper, we are particularly interested in individual review entry. The raw data scrapped from the web are noisy and preperation process is necessary to minimize this noise. The initial step for this cleaning process is to remove all non-latin characters, and change all characters to lower case letters.

Good definition "Not meaningful

list of of stop words. bar chart.

2.1 **Removing Stop Words** The second step is to remove ~~meaningless words~~ and words with low frequency by utilizing the `tm` (Ingo Feinerer 2015) and the `RTextTools` (Timothy P. Jurka 2014) packages in R. The listed results in Table 1 shows the ~~importance~~ of stop words of removing stop ~~words~~ prior to running LDA method.

Bold Stop words

	Topic.1	Topic.2	Topic.3	Topic.4
1	the	the	the	the
2	and	and	and	and
3	for	you	was	with
4	you	for	were	was
5	food	this	for	for
6	this	are	had	but
7	are	they	our	this
8	but	was	but	they
9	good	great	that	had
10	with	have	not	you

(a) Table LDA output before removing the stop words

	Topic.1	Topic.2	Topic.3	Topic.4
1	food	food	good	food
2	service	good	place	great
3	good	place	food	good
4	restaurant	like	great	place
5	just	just	also	service
6	one	one	chicken	get
7	ordered	service	just	one
8	back	get	ordered	restaurant
9	like	restaurant	really	can
10	really	will	get	best

(b) Table LDA output after removing the stop words

Table 1: Difference of LDA output between with and without stop words

2.2 **Stemming.** histogram. 30-50 vertical.

- LDA, documents are repsented as random mixtures over latent topics, where each topic is characerized by a distirbution over words. (Need paraphrasing)

What stemming & Example

good example w/ stemming.

LDA assumse that words are generated by topics.

Therefore, retrieving information by reducing the inflected words to its original word stem

may increase the probability of the joint distribution of topic mixture leading (thus increasing the probability of a document and a corpus).

	Topic.1	Topic.2	Topic.3	Topic.4
1	food	food	good	food
2	service	good	place	great
3	good	place	food	good
4	restaurant	like	great	place
5	just	just	also	service
6	one	one	chicken	get
7	ordered	service	just	one
8	back	get	ordered	restaurant
9	like	restaurant	really	can
10	really	will	get	best

(a) Table LDA output before stemming

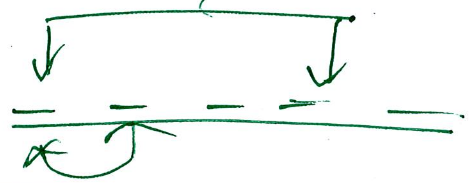
	Topic.1	Topic.2	Topic.3	Topic.4
1	good	food	good	food
2	food	place	food	good
3	place	good	order	great
4	great	restaur	servic	restaur
5	get	great	place	place
6	breakfast	servic	one	delici
7	servic	order	just	tri
8	time	price	great	like
9	price	eat	time	one
10	wait	time	like	love

(b) Table LDA output after stemming

Table 2: Difference of LDA output between with and without stop words

feature extraction

3 Sequential bigram



Wallach

~~nic~~

There is / isn't in R.

original

1. nice service → nice servic

2. I was served → ~~I was~~ servic

f

Creative Component

Taikgun Song

Abstract

Write abstract. It is known that the performance of Laten Dirichlet Allocation based topic models over short texts. In this paper, we would like to compare LDA methods with different 'parameter' under experimental setting.

1 Introduction

Lit·review
Wallack

Then the following R packages were utilized to conduct `Latent Dirichlet Allocation`(Blei, Ng, and Jordan 2003) method: `openNLP`(Hornik 2016a), `NLP` (Hornik 2016b), `topicmodels` (Bettina Grün 2016).

2 Data Set

2 Data ~~Cleaning~~ Preparation

Datasets. In order to compare different LDA processes, a collection of short user-generated online reviews ~~from a~~ popular website, the Trip Advisor, was used for evaluation. Among all restaurants in Honolulu, Hawaii registered on the Trip Advisor, the latest 10 reviews were scrapped and read into R (R Core Team 2014) using the `RCurl` (Lang 2015) package. This dataset includes the following information of the 7700 Honolulu restaurants: restaurant name, number total reviews, average star rating, individual review title, individual review entry, individual star rating, and the date visited. In this paper, we are particularly interested in individual review entry. The raw data scrapped from the web are noisy and preperation process is necessary to minimize this noise. The initial step for this cleaning process is to remove all non-latin characters, and change all characters to lower case letters.

definition

subsection ~~list of [illegible]~~ What are stop words ?

~~Removing~~ Stop Words The second step is to remove meaningless words and words with low frequency by utilizing the `tm` (Ingo Feinerer 2015) and the `RTextTools` (Timothy P. Jurka 2014) ~~packages in R. The~~ listed results in Table 1 shows the importance of removing stop words prior to running LDA method.

not meaningful

not stop words → later

bold all stop words

	Topic.1	Topic.2	Topic.3	Topic.4
1	the	the	the	the
2	and	and	and	and
3	for	you	was	with
4	you	for	were	was
5	food	this	for	for
6	this	are	had	but
7	are	they	our	this
8	but	was	but	they
9	good	great	that	had
10	with	have	not	you

(a) Table LDA output before removing the stop words

	Topic.1	Topic.2	Topic.3	Topic.4
1	food	food	good	food
2	service	good	place	great
3	good	place	food	good
4	restaurant	like	great	place
5	just	just	also	service
6	one	one	chicken	get
7	ordered	service	just	one
8	back	get	ordered	restaurant
9	like	restaurant	really	can
10	really	will	get	best

(b) Table LDA output after removing the stop words

Table 1: Difference of LDA output between with and without stop words

~~subsection~~

Stemming.

- LDA, documents are repsented as random mixtures over latent topics, where each topic is characerized by a distirbution over words. (Need paraphrasing)

- What is stemming
- Give an example
- code for stemming

serv*

feature extraction

Kogan – what is that, computational cost

LDA assumse that words are generated by topics.

Therefore, retrieving information by reducing the inflected words to its original word stem

may increase the probability of the joint distribution of topic mixture leading (thus increasing the probability of a document and a corpus).

	Topic.1	Topic.2	Topic.3	Topic.4
1	food	food	good	food
2	service	good	place	great
3	good	place	food	good
4	restaurant	like	great	place
5	just	just	also	service
6	one	one	chicken	get
7	ordered	service	just	one
8	back	get	ordered	restaurant
9	like	restaurant	really	can
10	really	will	get	best

(a) Table LDA output before stemming

	Topic.1	Topic.2	Topic.3	Topic.4
1	good	food	good	food
2	food	place	food	good
3	place	good	order	great
4	great	restaur	servic	restaur
5	get	great	place	place
6	breakfast	servic	one	delici
7	servic	order	just	tri
8	time	price	great	like
9	price	eat	time	one
10	wait	time	like	love

(b) Table LDA output after stemming

Table 2: Difference of LDA output between with and without stop words

3 Sequential bigram

Wallach