LING 410X: Language as Data

Semester: Spring '18

Instructor: Sowmya Vajjala

Iowa State University, USA

27 March 2018

Class Outline - Topic Modeling in detail

- Continuing from last class: how does a topic model "learn"?
- Sensitivity of topic model to its parameters (e.g., num. iterations)
- Evaluating topic models automatically
- Human evaluation of topic models

Class Outline - Topic Modeling in detail

- Continuing from last class: how does a topic model "learn"?
- Sensitivity of topic model to its parameters (e.g., num. iterations)
- Evaluating topic models automatically
- Human evaluation of topic models
- Reminder: A5 submission due on 3/31. If topicmodels is slow, or you don't want to learn an extra library, you can also do the assignment with mallet instead.

Exercise on Thursday

- How many of tried this out?
- How many of you successfully make this work on your computers?

Exercise on Thursday

- How many of tried this out?
- How many of you successfully make this work on your computers?
- The task was to "learn" a topic model, based on a dataset.

"Learning" a topic model: intuitive explanation of LDA

- Aim: organize a topic collection into topics, where each topic is a collection of words.
- Assumptions: each document is a mixture of topics, each topic is a mixture of keywords, a keyword can exist in multiple topics.

"Learning" a topic model: intuitive explanation of LDA

- Aim: organize a topic collection into topics, where each topic is a collection of words.
- Assumptions: each document is a mixture of topics, each topic is a mixture of keywords, a keyword can exist in multiple topics.
- Learning:
  - Decide on the number of topics K.
  - Go through each document, randomly assign each word in a document to one of the K topics.
  - At that point, you have on poorly represented topic model already, where each document is represented as a collection of topics.

after random initialization

- Again, we start revisiting all documents.
- For each document d, for each word w in it, we compute two probabilities:
  $P(t|d)$ i.e, proportion of words in the document d assigned topic t
  $P(w|t)$ i.e., proportion of documents with topic t, containing word w.

after random initialization

- Again, we start revisiting all documents.
- For each document d, for each word w in it, we compute two probabilities:
  $P(t|d)$ i.e, proportion of words in the document d assigned topic t
  $P(w|t)$ i.e., proportion of documents with topic t, containing word w.
- We reassign w a new topic t with a probability $P(t|d) * P(w|t)$
- Assumption made: each time we assign a word a new topic like this, we are assuming all other assignments are correct.

after random initialization

- Again, we start revisiting all documents.
- For each document d, for each word w in it, we compute two probabilities:
  $P(t|d)$ i.e, proportion of words in the document d assigned topic t
  $P(w|t)$ i.e., proportion of documents with topic t, containing word w.
- We reassign w a new topic t with a probability $P(t|d) * P(w|t)$
- Assumption made: each time we assign a word a new topic like this, we are assuming all other assignments are correct.

Finally,

- If you keep doing this process several times (iterations), you will reach a state where there is not much change in the topic-word distributions between iterations.
- At that point, one final time, you go through all documents and assign topic probabilities to them.

Note: Read `https://goo.gl/jbgCKq` for more detailed explanation of this.

Some name dropping

- The topics and words are assumed to have "Dirichlet distribution" which is one of the several probability distributions studied in probability and statistics (D of LDA comes from there!)
- The iterative process I conceptually explained is generally implemented by an algorithm called "Gibbs Sampling".

"Learning" a topic model: steps in the code

1. Pre-processing of the corpus (in author's version: splitting it into chunks, pre-processing of chunks)
2. Getting it into a two column format (id, text).
3. Using mallet package in R to build a topic model.
4. use mallet.import() function to convert a dataset into mallet format
5. use MalletLDA() function to create a topic model, setting its parameters.
6. observe and analyze the output in different ways.

(Let me go through this taking our movie review data from text classification weeks)

Step 1: Pre-processing of the corpus

```
setwd("~/Dropbox/ClassroomSlides-BothCourses/LING410X/21Mar2018/data/reviews")

get_text_string <- function(file_path)
{
  fulltext <- scan(file_path, what = "character", sep = "\n")
  fulltext_as_string <- tolower(paste(fulltext, collapse = " "))
  return (fulltext_as_string)
}

files.v <- dir(path=getwd(), pattern=".*txt")
documents = c()
for(i in 1:length(files.v)){
  documents = c(documents, get_text_string(files.v[i]))
}
```

Step 2: Getting it into a two column format (id, text)

```
docids = seq(1:length(documents))

fortopicmodels <- cbind(docids,documents)
finaldocuments <- as.data.frame(fortopicmodels, stringsAsFactors=F)
```

Step 3: Import data into mallet

```
stoplistfile = "/ClassroomSlides-BothCourses/LING410X/21Mar2018/stoplist.csv"
mallet.instances <- mallet.import(finaldocuments$docids,
                                  finaldocuments$documents,
                                  stoplistfile,FALSE,token.regexp="[\\p{L}']+")
```

- What does that regular expression mean?

Step 3: Import data into mallet

```
stoplistfile = "/ClassroomSlides-BothCourses/LING410X/21Mar2018/stoplist.csv"
mallet.instances <- mallet.import(finaldocuments$docids,
                                finaldocuments$documents,
                                stoplistfile,FALSE,token.regexp="[\\p{L}’]+")
```

- What does that regular expression mean? It means: match one or more occurrences of any Unicode character or a " ' ".
- What is that "FALSE"?

Step 3: Import data into mallet

```
stoplistfile = "/ClassroomSlides-BothCourses/LING410X/21Mar2018/stoplist.csv"
mallet.instances <- mallet.import(finaldocuments$docids,
                                finaldocuments$documents,
                                stoplistfile,FALSE,token.regexp="[\\p{L}']+")
```

- What does that regular expression mean? It means: match one or more occurrences of any Unicode character or a " ' ".
- What is that "FALSE"? - It is the argument "preserve case".

More on regular expressions:
http://www.regular-expressions.info
More on unicode in regular expressions:
http://www.regular-expressions.info/unicode.html

Setting up the topic model training process

```
topic.model <- MalletLDA(num.topics=XX) #You have to specify this.
topic.model$loadDocuments(mallet.instances) #from previous slide.
topic.model$setAlphaOptimization(40, 80) #Optional
topic.model$train(NUM) #Actual training happens here.
```

What are all those parameters?

- Best place to look for that= ?MalletLDA
- setAlphaOptimization(40,80) means: perform model optimization after every 40 iterations, starting first after 80 iterations.
- number of iterations: the more the better, generally. However, trade off is that training becomes very slow.
- Every 50 iterations, R prints a summary of your topic model showing top 7 words per topic.
- Every 10 iterations, R also shows a log likelihood of this topic model (the closer to 0, the better).

Output screenshots for Thursday's Data, when I trained it last year

Output screenshot - at the beginning

```
Console ~/
> #This starts the training process
> topic.model$train(400)
Mar 28, 2017 8:55:31 AM cc.mallet.topics.ParallelTopicModel estimate
INFO: <10> LL/token: -10.34105
Mar 28, 2017 8:55:33 AM cc.mallet.topics.ParallelTopicModel estimate
INFO: <20> LL/token: -9.85248
Mar 28, 2017 8:55:35 AM cc.mallet.topics.ParallelTopicModel estimate
INFO: <30> LL/token: -9.6863
Mar 28, 2017 8:55:36 AM cc.mallet.topics.ParallelTopicModel estimate
INFO: <40> LL/token: -9.59737
Mar 28, 2017 8:55:38 AM cc.mallet.topics.ParallelTopicModel estimate
INFO:
0       0.11628 doctor made thought great mind letter long
1       0.11628 irish ireland mr family native father called
2       0.11628 sir reilly man robert squire folliard mr
3       0.11628 d'arcy bejabers camp gold dermod man se
4       0.11628 rachael warren magsie miss george gregory woman
5       0.11628 abellino time men thou long frazier ground
6       0.11628 lady young read sort fair table mr
7       0.11628 ' 'i miss 'and wylder 'you rachel
8       0.11628 property man justice church principles fact judge
9       0.11628 don parker farrel horse pablo mike miguel
10      0.11628 heart love father mind tears girl child
11      0.11628 father heart mother god connor son man
12      0.11628 bryce cardigan colonel shirley pennington sequoia woods
13      0.11628 god good poor day father make home
14      0.11628 nell tim sheila danny children prairie wild
15      0.11628 door room night hand madame window house
16      0.11628 men lord people country cumber friend good
17      0.11628 ye mr story wi' eagle gilbert henry
18      0.11628 peter mary dinner day great emily made
19      0.11628 men people country state land indians english
20      0.11628 day young life children years death family
21      0.11628 sir mr replied good hycy man bryan
22      0.11628 night face life arms clarence found long
23      0.11628 man country time house fact blood found
24      0.11628 uncle lady cousin milly time maud good
25      0.11628 day long work office hand asked home
26      0.11628 high river small large green trees years
27      0.11628 charles maria french washington great place make
28      0.11628 mr sir captain lake man larkin mark
29      0.11628 wid good man ould mane priest night
30      0.11628 night truth matter felt poor replied make
31      0.11628 big mother flurry water house family back
32      0.11628 replied poor god man exclaimed ay good
33      0.11628 room eyes marston mrs face stood passed
34      0.11628 john money good dollars years pay day
35      0.11628 susan martie mrs billy sally asked miss
36      0.11628 doctor sir mrs toole puddock sturk dangerfield
37      0.11628 gerald san city years race young francisco
38      0.11628 mother young margaret lydia good aunt mrs
39      0.11628 replied miss harry woodward family mr mother
40      0.11628 night made men man good knew life
41      0.11628 spirit denis light father deep spoke appearance
42      0.11628 school van time man boys job horses

Mar 28, 2017 8:55:38 AM cc.mallet.topics.ParallelTopicModel estimate
INFO: <50> LL/token: -9.54878
Mar 28, 2017 8:55:39 AM cc.mallet.topics.ParallelTopicModel estimate
INFO: <60> LL/token: -9.50984
```

Output screenshot - after 400 iterations

```
INFO: <360> LL/token: -9.2627
Mar 28, 2017 8:56:30 AM cc.mallet.topics.ParallelTopicModel estimate
INFO: <370> LL/token: -9.25952
Mar 28, 2017 8:56:32 AM cc.mallet.topics.ParallelTopicModel estimate
INFO: <380> LL/token: -9.25652
Mar 28, 2017 8:56:33 AM cc.mallet.topics.ParallelTopicModel estimate
INFO: <390> LL/token: -9.25065
Mar 28, 2017 8:56:35 AM cc.mallet.topics.ParallelTopicModel estimate
INFO:
0	0.2753	doctor letter time read great thought long
1	0.01875 irish ireland family mr john county race
2	0.02243 sir reilly robert mr folliard replied whitecraft
3	0.01584 d'arcy bejabers don camp dermod gold jos
4	0.01545 rachael warren magsie billy clarence gregory alice
5	0.01169 abellino rosabella flodoardo andreas venice thou doge
6	0.3172	mr miss young good lady man great
7	0.11139 ' 'i 'and poor 'you 'well looked
8	0.10621 judge gentlemen law court man justice protestant
9	0.01702 don parker farrel pablo mike miguel kay
10	0.29702 heart love father god mother tears girl
11	0.06827 connor son mother fardorougha una heart bartle
12	0.01467 bryce cardigan shirley colonel pennington timber sequoia
13	0.08262 father priest church denis bishop poor holy
14	0.01356 nell tim sheila danny children prairie johann
15	0.32824 door room night house window bed time
16	0.07018 sir mr phil darby lord m'clutchy val
17	0.01095 ye wi' a' hae na gilbert sae
18	0.05661 susan miss peter mary sue emily ella
19	0.0693	men irish san california state city land
20	0.34224 life man country people young day years
21	0.02738 hycy bryan m'mahon burke kathleen replied cavanagh
22	0.3506	eyes face night life voice heart back
23	0.16512 country men house purcel night people party
24	0.0184	madame uncle milly maud cousin silas mary
25	0.38849 day time home back long told make
26	0.20601 water trees road river green long sun
27	0.04143 charles maria washington french great indians thou
28	0.02129 lake wylder rachel larkin stanley captain brandon
29	0.05588 sir wid good boy poor m'carthy mogue
30	0.59271 man time good replied make made matter
31	0.07142 john mother big van flurry farm family
32	0.12677 replied god man father ha poor good
33	0.11696 sir marston mrs ma'am mademoiselle room de
34	0.22014 money man business good dollars thousand pay
35	0.03537 martie sally billy lydia wallace monroe teddy
36	0.02801 sir toole puddock sturk dangerfield nutter doctor
37	0.01237 gerald ye mr ffrench young answered man
38	0.20356 mrs mother room girl woman girls margaret
39	0.02555 woodward harry replied sir barney mother goodwin
40	0.29675 man head back dead hand men horse
41	0.31523 time felt spirit eye heart eyes length
42	0.05723 school city paper editor wichita story job

Mar 28, 2017 8:56:35 AM cc.mallet.topics.ParallelTopicModel optimizeBeta
INFO: [beta: 0.01314]
Mar 28, 2017 8:56:35 AM cc.mallet.topics.ParallelTopicModel estimate
INFO: <400> LL/token: -9.24646
Mar 28, 2017 8:56:35 AM cc.mallet.topics.ParallelTopicModel estimate
INFO:
Total time: 1 minutes 6 seconds
```

Output screenshot - 10 topics instead of 43

```
Mar 28, 2017 9:01:07 AM cc.mallet.topics.ParallelTopicModel estimate
INFO: <350> LL/token: -9.20658
Mar 28, 2017 9:01:08 AM cc.mallet.topics.ParallelTopicModel optimizeBeta
INFO: [beta: 0.04899]
Mar 28, 2017 9:01:08 AM cc.mallet.topics.ParallelTopicModel estimate
INFO: <360> LL/token: -9.20642
Mar 28, 2017 9:01:09 AM cc.mallet.topics.ParallelTopicModel estimate
INFO: <370> LL/token: -9.20205
Mar 28, 2017 9:01:10 AM cc.mallet.topics.ParallelTopicModel estimate
INFO: <380> LL/token: -9.20313
Mar 28, 2017 9:01:12 AM cc.mallet.topics.ParallelTopicModel estimate
INFO: <390> LL/token: -9.20172
Mar 28, 2017 9:01:13 AM cc.mallet.topics.ParallelTopicModel estimate
INFO:
0       0.09582 irish men ireland washington french great people
1       0.23133 sir man mr replied good reilly country
2       0.16219 replied wid god man father good poor
3       0.28464 time mr man day years work good
4       0.13631 john mother big nell tim flurry farm
5       0.12586 ' sir 'i mr good lake poor
6       0.66484 night door room house back hand man
7       0.14838 susan martie mrs rachael miss mother billy
8       0.71019 heart father time love life day felt
9       0.05437 don d'arcy bryce man cardigan bejabers parker

Mar 28, 2017 9:01:13 AM cc.mallet.topics.ParallelTopicModel optimizeBeta
INFO: [beta: 0.04945]
Mar 28, 2017 9:01:13 AM cc.mallet.topics.ParallelTopicModel estimate
INFO: <400> LL/token: -9.20225
Mar 28, 2017 9:01:13 AM cc.mallet.topics.ParallelTopicModel estimate
INFO:
Total time: 47 seconds
> |
```

More functionalities in Mallet

- mallet.doc.topics function returns topic weights for each document in the training data.
- mallet.read.dir function takes a folder path with .txt files and converts it into id-text format for mallet.
- mallet.topic.hclust function performs a clustering of topics.
- mallet.topic.labels function gives a string with most probable words for a topic.

More details:
`https://cran.r-project.org/web/packages/mallet/mallet.pdf`

Getting Topics and Words

```
doc.topics <- mallet.doc.topics(topic.model, smoothed=T, normalized=T)
```

```
> doc.topics[10,]
 [1] 0.0307282777 0.0004646110 0.0003096116 0.0027386575 0.0831497737 0.0002558009 0.1737143446 0.0002883194 0.0037027725 0.7046478311
> doc.topics
              [,1]         [,2]         [,3]         [,4]         [,5]         [,6]         [,7]         [,8]        [,9]        [,10]
 [1,] 0.2381523988 0.0004566427 0.0003043016 0.0069793400 0.0988743131 0.0002514138 0.1342899945 0.0024272006 0.085104659 4.331597e-01
 [2,] 0.0176794271 0.0004656267 0.0003102885 0.0246046821 0.1532836606 0.0002563601 0.1587920583 0.0002889496 0.040872931 6.034460e-01
 [3,] 0.0002048295 0.0004983056 0.0003320653 0.1573392180 0.1195924769 0.0026137756 0.2377797941 0.0003092289 0.062456892 4.188734e-01
 [4,] 0.0119860682 0.0005018276 0.0003344123 0.0006020728 0.0874543268 0.0002762912 0.3431225524 0.0003114145 0.060542371 4.948687e-01
 [5,] 0.0025383147 0.0028311059 0.0003312903 0.0262700484 0.1333171323 0.0002737118 0.2372248247 0.1076708197 0.146333799 3.432090e-01
 [6,] 0.0002197372 0.0231217736 0.0003562334 0.0558545166 0.0906511999 0.0002943198 0.1446593257 0.0680933372 0.149822301 4.669273e-01
 [7,] 0.0002057924 0.0005006481 0.0003336263 0.0640620166 0.0919496137 0.0002756418 0.1049235792 0.0050115240 0.161468161 5.712694e-01
 [8,] 0.0002395791 0.0005828438 0.0003884006 0.0745796265 0.0085386309 0.0003208963 0.0646872844 0.0113069276 0.007381352 8.319745e-01
 [9,] 0.0002138340 0.0005202115 0.0052311959 0.0836611852 0.0027365393 0.0002864129 0.1163503995 0.0223032201 0.031010817 7.376862e-01
[10,] 0.0307282777 0.0004646110 0.0003096116 0.0027386575 0.0831497737 0.0002558009 0.1737143446 0.0002883194 0.003702773 7.046478e-01
[11,] 0.0410083141 0.0004576237 0.0067502512 0.1273065235 0.0024073003 0.0002519539 0.2656328560 0.0711822375 0.117513980 3.674890e-01
[12,] 0.0002043514 0.0121669592 0.0003312903 0.1476361408 0.0002812234 0.0026076751 0.2255550082 0.0026424705 0.085650753 5.229241e-01
[13,] 0.0002087361 0.0028918512 0.0217947740 0.0006092496 0.0741925503 0.0074317098 0.3090679774 0.0336917105 0.163777842 3.863336e-01
[14,] 0.0002072538 0.0005042033 0.0027031073 0.0077062586 0.0026523293 0.0002775993 0.3707852014 0.0003128888 0.098702781 5.161484e-01
[15,] 0.0076717889 0.0005292597 0.0003526928 0.0702078510 0.0077536270 0.0002913945 0.1804927378 0.1320199347 0.185604399 4.150763e-01
[16,] 0.0002449409 0.0033934371 0.0311701330 0.0510708062 0.0506929650 0.0003280781 0.2200002103 0.0003697847 0.088675472 5.540542e-01
[17,] 0.0002164774 0.0252512262 0.0003509487 0.1242547631 0.0942513288 0.0002899536 0.2438841337 0.0003268137 0.046228932 4.649454e-01
[18,] 0.0002170140 0.0054851210 0.0003518186 0.0080691706 0.0002986493 0.0002906723 0.0263730037 0.3002365984 0.105829592 5.528484e-01
[19,] 0.0121289446 0.0005078095 0.0074905237 0.1889485445 0.0050553408 0.0002795847 0.1016563489 0.0050832100 0.097024674 5.818250e-01
[20,] 0.0099828994 0.0005202115 0.0052311959 0.0006241291 0.0735622637 0.0002864129 0.1163503995 0.0003228229 0.031010817 7.621088e-01
[21,] 0.0002159435 0.0054580644 0.0003500832 0.1066837833 0.0002971761 0.0027555989 0.0657046793 0.0496532155 0.014052230 7.548292e-01
[22,] 0.0001996907 0.0050472678 0.0003237345 0.0074250441 0.0504509105 0.0002674692 0.2683060725 0.0094243984 0.019836796 6.387186e-01
[23,] 0.0002214042 0.0005386282 0.0003589358 0.1599560847 0.0786952569 0.0002965525 0.2039174805 0.0053917074 0.001763933 5.488600e-01
[24,] 0.0002010665 0.0257500415 0.0003259648 0.0488122004 0.1059131539 0.0002693119 0.2632651888 0.0003035479 0.088866796 4.662927e-01
[25,] 0.0048609328 0.0028245136 0.0003305189 0.0331944637 0.0887646559 0.0002730745 0.2017445107 0.0026363174 0.090108370 5.752626e-01
[26,] 0.0002208457 0.0005372695 0.0003580305 0.0410021883 0.0003039224 0.0002958045 0.1100761941 0.1390626364 0.180846305 5.272968e-01
[27,] 0.0002180951 0.0030215126 0.0003535713 0.1401289019 0.0177366790 0.0227105314 0.0713412127 0.0202567324 0.191048596 5.331842e-01
[28,] 0.0001983337 0.0004825027 0.0025867672 0.1614104096 0.0002729419 0.0002656516 0.1011207580 0.0410736109 0.105780834 5.868082e-01
[29,] 0.0002180951 0.0378945965 0.0003535713 0.1351470328 0.0177366790 0.0002921204 0.1535420532 0.0999666384 0.026646915 5.282023e-01
[30,] 0.0002143575 0.0274521870 0.0003475119 0.0079703940 0.0027432391 0.0002871141 0.2072203419 0.0811157188 0.160843757 5.118054e-01
[31,] 0.0002277389 0.0031551189 0.0003692057 0.0891014203 0.0601382391 0.0263158333 0.0094688140 0.0003438151 0.131868381 6.790114e-01
[32,] 0.0028214796 0.0005526020 0.0003682478 0.0758986040 0.0003125956 0.0003042461 0.0613308792 0.0107202492 0.061479320 7.862118e-01
[33,] 0.0001930851 0.0092908814 0.0003130256 0.0049741420 0.0576031770 0.0024639083 0.2947152843 0.0002914985 0.025796469 6.043585e-01
[34,] 0.0001978854 0.0027415252 0.0138814859 0.0005775793 0.0273936813 0.0002650511 0.4444293930 0.0002987455 0.033218141 4.769965e-01
```

Sensitivity of a topic model to its parameters

- Note: Topic models are very sensitive to their parameters.
- How to choose the number of topics?: Intuitive - needs an understanding of the corpus. There are some ways of using probability based estimates (which are beyond the scope of this course), but there is no "formula" for choosing the best number.

Sensitivity of a topic model to its parameters

- Note: Topic models are very sensitive to their parameters.
- How to choose the number of topics?: Intuitive - needs an understanding of the corpus. There are some ways of using probability based estimates (which are beyond the scope of this course), but there is no "formula" for choosing the best number.
- How to choose the number of iterations?: Looking at log likelihoods is a good way.

Useful reference: ldatuning library in R
`https://cran.r-project.org/web/packages/ldatuning/ldatuning.pdf`

Human evaluation of topic models -1

- Word intrusion test: If an irrelevant word is shown along with related words for a given topic, a human evaluator should be able to identify that word as an intruder.

Human evaluation of topic models -1

- Word intrusion test: If an irrelevant word is shown along with related words for a given topic, a human evaluator should be able to identify that word as an intruder.
- If a human is having a difficulty doing that, the topic is not coherent.

Human evaluation of topic models -1

- Word intrusion test: If an irrelevant word is shown along with related words for a given topic, a human evaluator should be able to identify that word as an intruder.
- If a human is having a difficulty doing that, the topic is not coherent.
- Experimental results: models with a larger number of topics may improve automated evaluations but will reduce human interpretability, because they will not be coherent.

Human evaluation of topic models -2

- Topic intrusion test: If a document is given, and its most appropriate topics (according a topic model) are shown along with a random topic, human evaluator should be able to identify the intruder topic correctly.

Human evaluation of topic models -2

- Topic intrusion test: If a document is given, and its most appropriate topics (according a topic model) are shown along with a random topic, human evaluator should be able to identify the intruder topic correctly.
- If topic assignments from the model are intuitive, then the human intruder should not have difficulties in doing this.

Human evaluation of topic models -2

- Topic intrusion test: If a document is given, and its most appropriate topics (according a topic model) are shown along with a random topic, human evaluator should be able to identify the intruder topic correctly.
- If topic assignments from the model are intuitive, then the human intruder should not have difficulties in doing this.
- topic here means top-N keywords from the topic model for that topic.

Human evaluation of topic models -2

- Topic intrusion test: If a document is given, and its most appropriate topics (according a topic model) are shown along with a random topic, human evaluator should be able to identify the intruder topic correctly.
- If topic assignments from the model are intuitive, then the human intruder should not have difficulties in doing this.
- topic here means top-N keywords from the topic model for that topic.
- Experimental results: humans do well documents which have focussed discussions since it is easy to assign a topic in such cases (easy for machines too!)

Automatic evaluation of topic models-1

1. Approximating the word intrusion test automatically by comparing semantic relatedness between words in a topic.

Automatic evaluation of topic models-1

1. Approximating the word intrusion test automatically by comparing semantic relatedness between words in a topic.
2. hold out a few words from each document from training data and use it to evaluate the topic model.

Note: Inference from a topic model is available in original Mallet tool, but does not seem to be available in the R library version.
`http://mallet.cs.umass.edu/topics.php`
Enthusiasts can have a look at this R code for the same:
`https://gist.github.com/agoldst/edcfd45b5ac371296b76`

Automatic evaluation of topic models -2

- Use the topic model as a part of some other application scenario (e.g., using it to retrieve similar documents to a given document).
- Check if the application performance has gotten better because of the use of this topic model.
- If it does, we say the topic model is good for that application. Else, not good.

Automatic evaluation of topic models -2

- Use the topic model as a part of some other application scenario (e.g., using it to retrieve similar documents to a given document).
- Check if the application performance has gotten better because of the use of this topic model.
- If it does, we say the topic model is good for that application. Else, not good.
- Semi-automatic: comparing topics in terms of overlapping words in Top-25 words for each topic.

Evaluation: general remarks

- A general conclusion from topic modeling research has been that human evaluations and machine evaluations do not agree with each other.
- While machine evaluations are good for certain tasks (e.g., search engines etc.) we may need topic models that are optimized to human interpretations for some other tasks (e.g., analyzing literary documents for themes).

Evaluation: general remarks

- A general conclusion from topic modeling research has been that human evaluations and machine evaluations do not agree with each other.
- While machine evaluations are good for certain tasks (e.g., search engines etc.) we may need topic models that are optimized to human interpretations for some other tasks (e.g., analyzing literary documents for themes).
- One thing to keep in mind: human evaluation is expensive.
- How to consider human interpretation as a part of the mathematical modeling process? - is an ongoing research question.

Next Class

- Wrapping up topic modeling
- Topic modeling with or without mallet - Assignment 6 practice