title: “Topic modeling project1”

author: “Haochen Pan Xiaoyanbin Cai Shengbo Wang”

date: “2022-11-14”

Topic Modeling

Part 1: Choosing Dataset

The data we choose is from the IMDB website. The dataset contains 50K movie reviews for natural language processing or text analysis and there is a dataset for binary sentiment classification. In other words, the website provides 25,000 highly polar movie reviews for training and 25,000 for testing. So, predict the number of positive and negative reviews using either classification or deep learning algorithms. We want to make some analysis of the reviews from the dataset.

Part 2: Preparation

We import the excel of the IMDB dataset and we change the CSV form in data. Frame. After we cleaned the dataset, we get a TermDocumentMatrix for the words from the excel as follows:

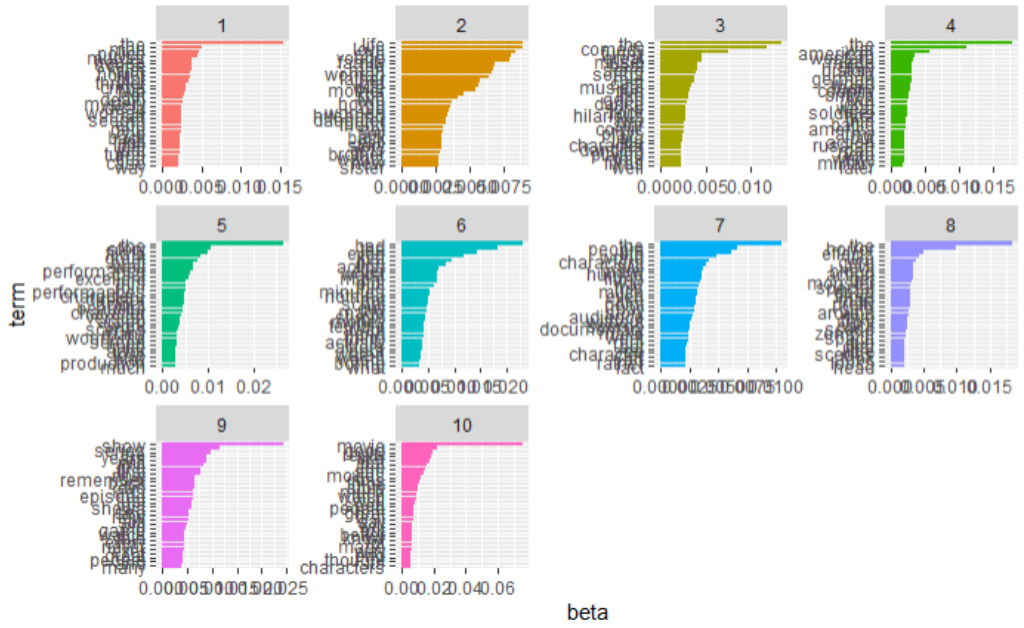
```
Warning: transformation drops documentsWarning: transformation drops documentsWarning: transformation
drops documentsWarning: transformation drops documents<<TermDocumentMatrix (terms: 162475, documents:
50000)>>
Non-/sparse entries: 4924101/8118825899
Sparsity           : 100%
Maximal term length: 72
Weighting          : term frequency (tf)
Sample             :
        Docs
Terms    12648 3025 31241 31437 31482 3655 40522 42947 43822 5709
  even       1    4     1     5     2    3     2     1     6    2
  good       0    2     0     7     1    2     2     4     1    2
  just       1    5     1     4     2    2     6     4     1    3
  like       5    4     7    15     3    4     3     7     2    8
  movie     12    2     1     0     0   14    22     2     4    0
  really     0    3     2     6     0    1     2     2     0    1
  see        1    3     3     0     1    1     7     6     1    2
  story      6    0     1     5     0    3     2     3     2    0
```

Part 3: LDA Code

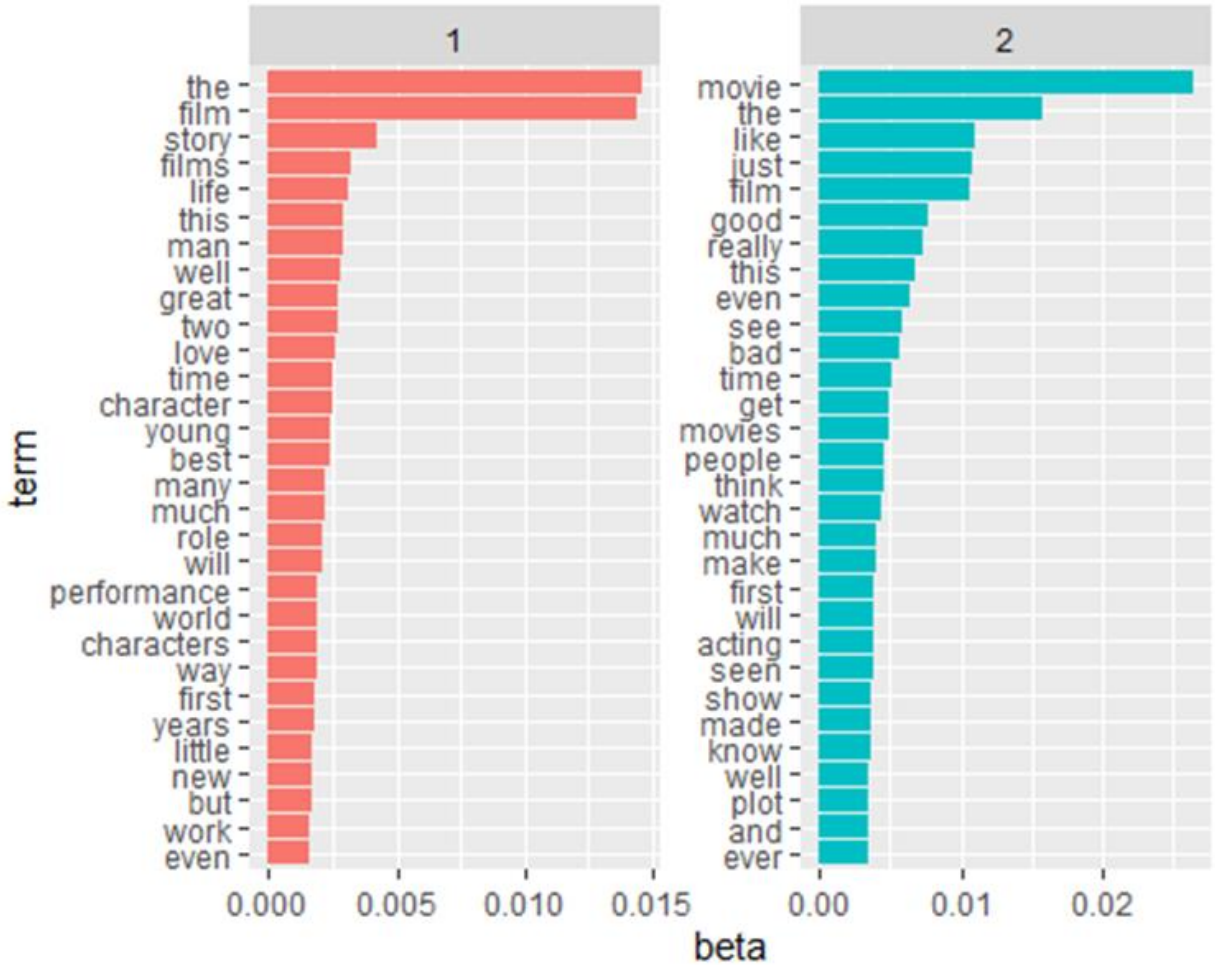
Then, according to the mining data's instruction, we start to use the LDA code to find useful information about words and make the visulization. In the beginning, we could not choose a large value of k due to incomplete data cleaning. But as we choose a better data cleaning code, we can also choose k=10, in other words, 10 topics.

```
A topic model with 10 topics, 50000 documents and a 162475 word dictionary.
Topic 1 Top Words:
         Highest Prob: the, man, police, murder, house, scene, killer
         FREX: hitchcock, holmes, murderer, wax, bruno, giallo, widmark
         Lift: ¡viva, ¨farenheit, 'humanization', 'lost, 'possessed', 'showdown', ""
         Score: murder, killer, horror, police, thriller, murders, holmes
Topic 2 Top Words:
         Highest Prob: life, love, she, young, family, the, man
         FREX: daughters, scarlett, natalie, mildred, edie, paulie, divorced
         Lift: scarlea, ¡¦s, ¡¨, ¡§, ¡§astronauts, ¡§but
         Score: mother, father, husband, she, wife, family, woman
Topic 3 Top Words:
         Highest Prob: the, comedy, funny, great, music, best, song
         FREX: songs, singing, batman, slapstick, keaton, broadway, sinatra
         Lift: batman, singing, ¨big, 'anna, "consider, "family", "food
         Score: comedy, funny, songs, batman, musical, singing, broadway
Topic 4 Top Words:
         Highest Prob: the, war, american, western, made, men, history
         FREX: war, western, russian, vietnam, westerns, hitler, civil
         Lift: "the, …, abdalla, abishag, abolitionists, adeline, adman
         Score: war, western, soldiers, westerns, german, military, russian
Topic 5 Top Words:
         Highest Prob: the, story, films, great, best, well, performance
         FREX: hamlet, austen, branagh, shakespeares, ponyo, eyre, miyazaki
         Lift: hamlet,     film,   with,  a,  as,  astounding,  journey
         Score: performance, performances, excellent, great, novel, wonderful, story
Topic 6 Top Words:
         Highest Prob: bad, the, even, just, like, acting, ever
         FREX: worst, awful, waste, crap, terrible, horrible, stupid
         Lift: absolutley, accomplishmentbr, actingwriting, actium, afterthoughts, ajooba, allred
         Score: bad, worst, waste, awful, stupid, crap, terrible
Topic 7 Top Words:
         Highest Prob: the, people, life, world, characters, will, many
         FREX: religion, beliefs, bettie, christianity, gandhi, herzog, israeli
         Lift: addicus, administrations, aeon, alexandres, anl, artiste, aschenbach
         Score: documentary, society, human, social, religion, people, political
Topic 8 Top Words:
         Highest Prob: the, horror, effects, like, gore, evil, blood
         FREX: monster, zombie, vampire, zombies, vampires, halloween, werewolf
         Lift: aaliyahs, aawip, abbeys, acuity, adamantium, aerobicide, afzel
         Score: horror, gore, zombie, monster, zombies, slasher, vampire
Topic 9 Top Words:
         Highest Prob: show, series, the, years, will, like, first
         FREX: episode, episodes, season, seasons, abc, spock, bam
         Lift:  usual, ¡gracias, £amazing, «battlestar, «bazar», «blakes», «blindpassasjer»
         Score: show, episode, episodes, series, season, kids, game
Topic 10 Top Words:
         Highest Prob: movie, good, really, like, just, the, see
         FREX: movie, movies, really, thought, disappointed, good, recommend
         Lift: achilless, afterdark, alittle, antonie, aquafresh, athenas, basilisk
         Score: movie, movies, really, think, watch, good, great
```

Subsequently, we completed the plotting of key topics using tidy and ggplot code.

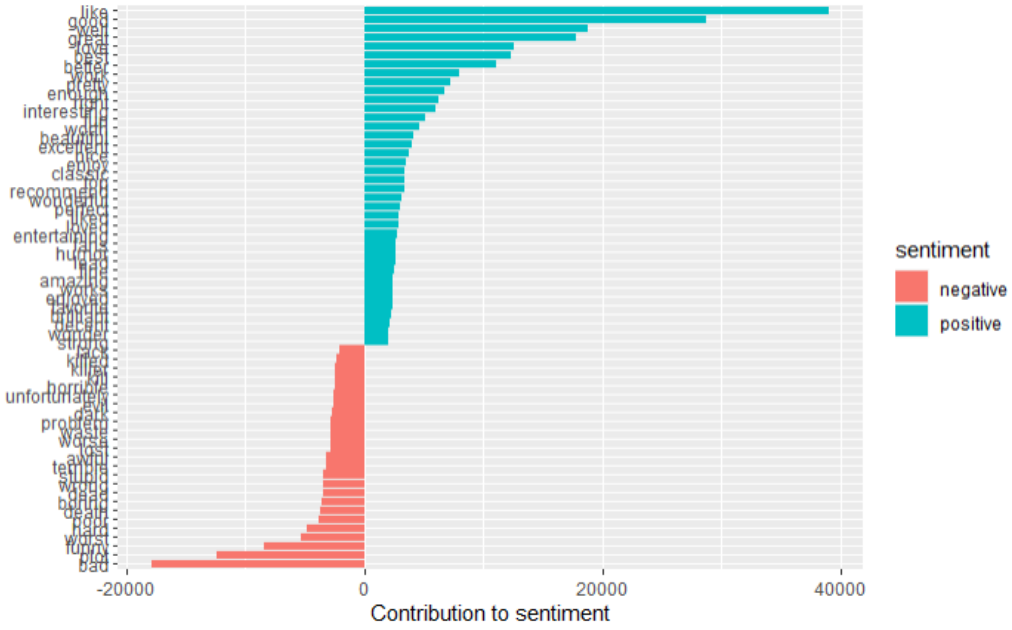


In contrast, the k=2 comparison map we made at the beginning is relatively simple.

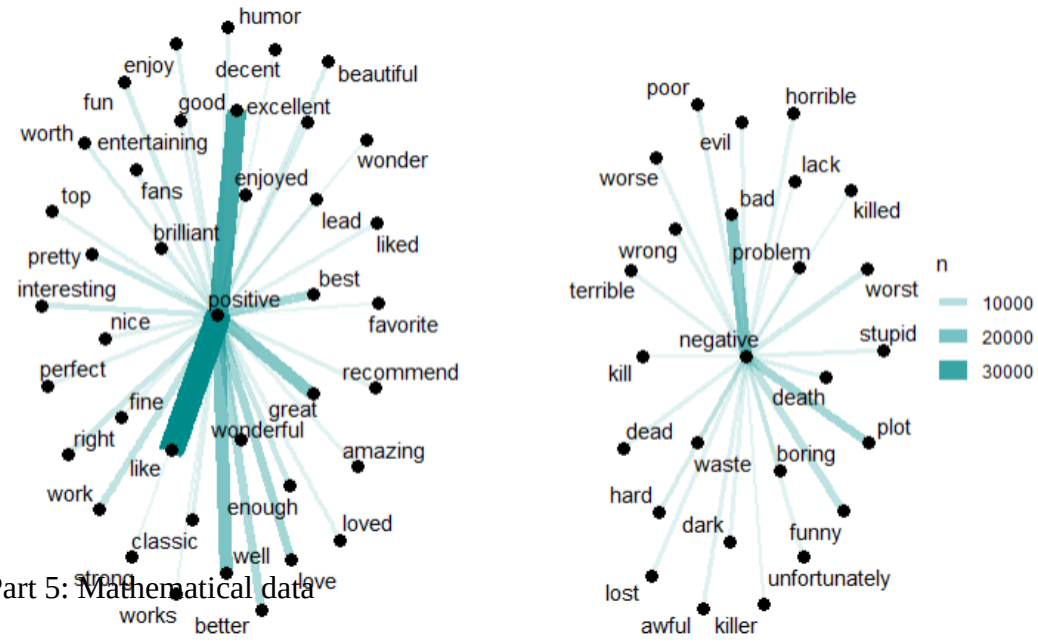


Part 4: Positive/Negative Sentiment

At the same time, we found that reviews with different evaluations have keywords with different frequencies. So we did the visulization too.



According to the content of Mining data chp8, we can also directly observe the association and occurrence frequency of different words through a more vivid method.



Part 5: Mathematical data

We consider the terms that had the greatest difference in β among topics. This can be estimated based on the log ratio of the two logs ratio is useful because it makes the difference symmetrical: being twice as large leads to a log ratio of 1 while being twice as large results in -1. To constrain it to a set of especially relevant words, we can filter for relatively common words, such as those that have a β greater than 1/1000 in at least one topic.

A tibble: 221 × 12

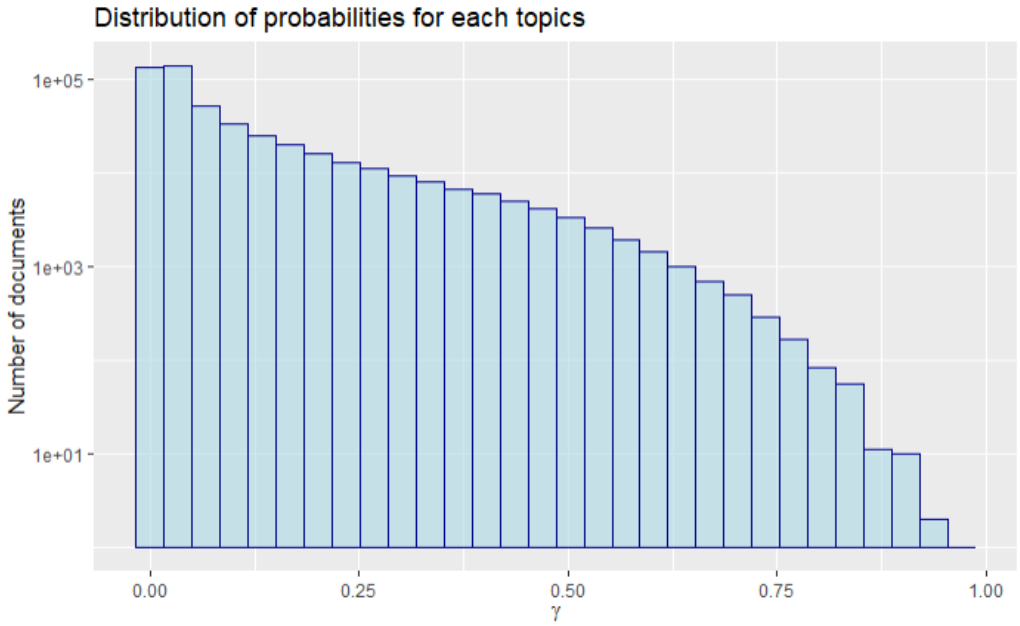
term <chr>	topic1 <dbl>	topic2 <dbl>	topic3 <dbl>	topic4 <dbl>	topic5 <dbl>
actress	3.629128e-17	1.779693e-03	4.410557e-04	7.198263e-32	1.163362e-03
after	1.004507e-03	1.043237e-03	3.007825e-04	6.042338e-04	5.778474e-07
age	4.802613e-15	1.013841e-03	2.892963e-04	2.125390e-04	2.768068e-04
along	1.087307e-03	8.971023e-04	1.077235e-03	9.072840e-04	7.392159e-04
always	7.550549e-04	1.033426e-03	1.486585e-03	4.935701e-04	1.889343e-03
and	6.134345e-04	1.468313e-03	1.718436e-03	1.175879e-03	7.861280e-04
another	1.854581e-03	1.760349e-03	1.144081e-03	1.649395e-03	5.907891e-04
around	1.753741e-03	1.850986e-03	5.899711e-04	7.321534e-04	1.474341e-05
atmosphere	1.105119e-03	3.128352e-39	1.684396e-33	5.236588e-33	1.077932e-03
away	1.393780e-03	1.620434e-03	1.027858e-04	4.355813e-04	3.426676e-05

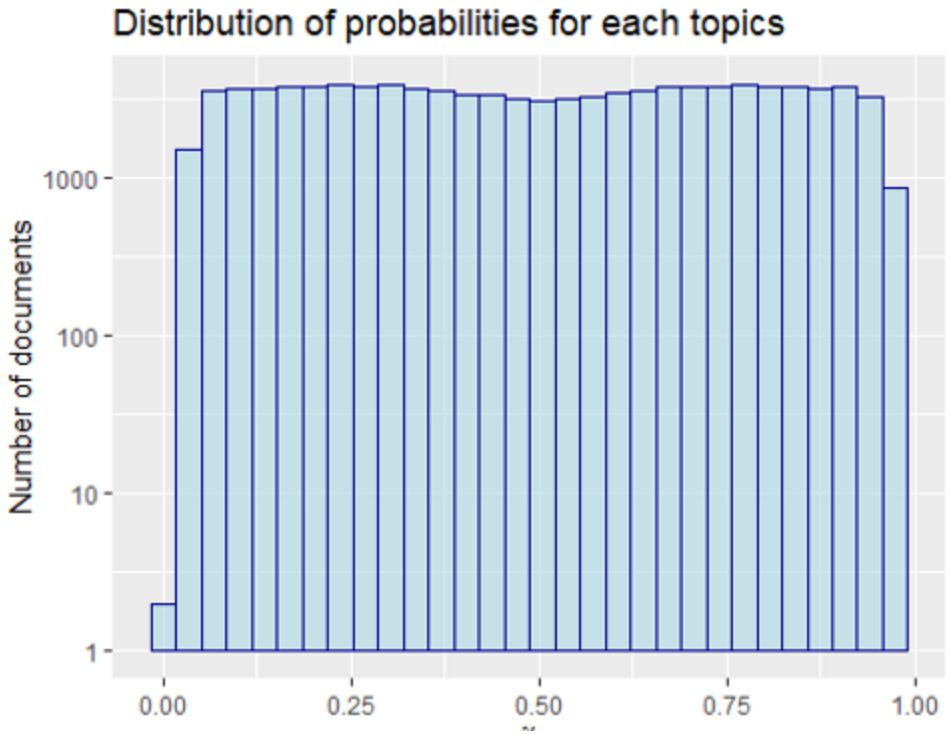
1-10 of 221 rows | 1-6 of 12 columns

Previous 1 2 3 4 5 6 ... 23 Next

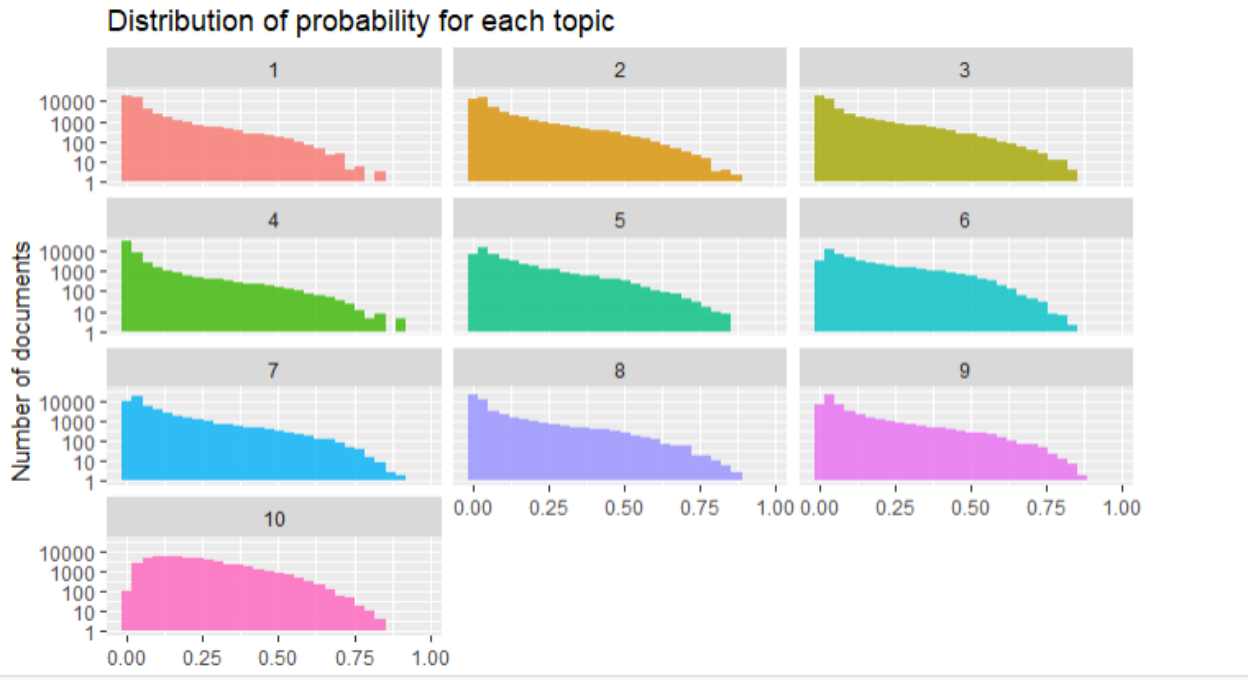
Part 6: Distribution

We noticed that some of the probabilities visible at the top of the data frame are low and some are higher. Our model has assigned a probability to each description belonging to each of the topics we constructed from the sets of words. Of course, we also found that when k=2 and k=10, the distribution graph is different. (Up : k = 10; Down k = 2)

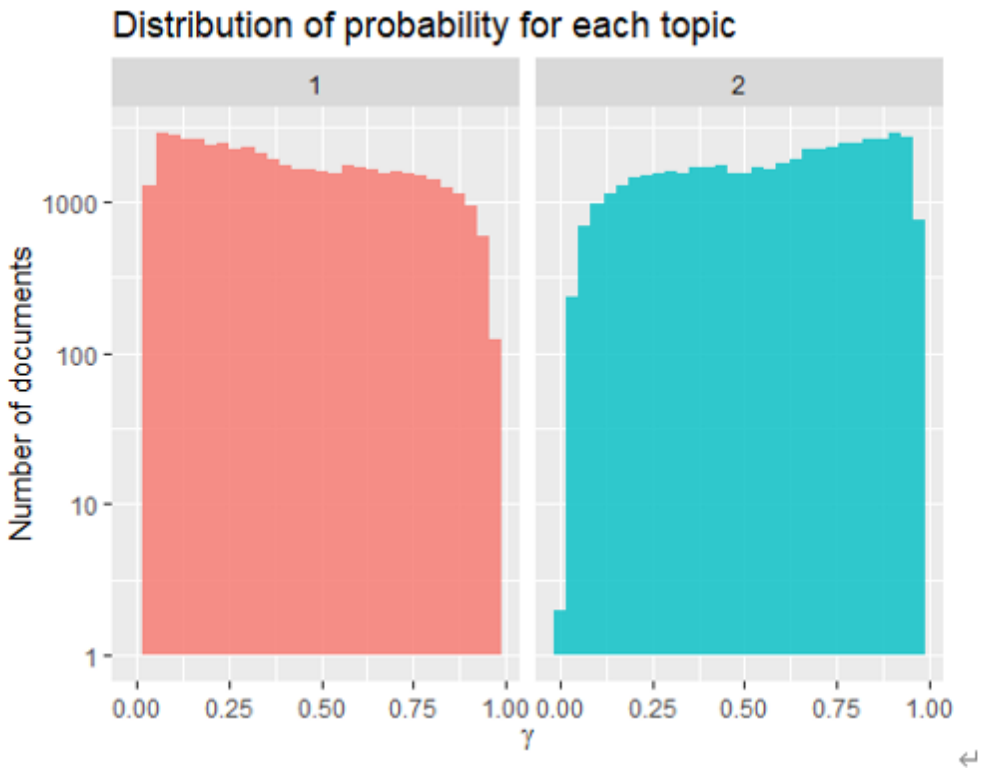




A similar thing happened in the distribution of probability.



All the documents with γ close to 0; This plot displays the type of information we used to choose how many topics for our topic modeling procedure.



This plot is the distribution of probability when we fit to K=2. Topic 1 with γ close to 0, but topic 2 with γ close to 1.