1

00:00:00,069 --> 00:00:07,429

[SOUND]

This lecture

2

00:00:07,429 --> 00:00:11,820

is a continuing discussion of Generative

Probabilistic Models for text clustering.

3

00:00:13,450 --> 00:00:17,620

In this lecture, we are going to continue

talking about the text clustering,

4

00:00:17,620 --> 00:00:20,910

particularly, the Generative

Probabilistic Models.

5

00:00:23,950 --> 00:00:28,320

So this is a slide that you have seen

earlier where we have written down

6

00:00:28,320 --> 00:00:32,735

the likelihood function for

a document with two

7

00:00:32,735 --> 00:00:38,049

distributions, being a two component

mixed model for document clustering.

8

00:00:39,800 --> 00:00:47,360

Now in this lecture, we're going to

generalize this to include the k clusters.

9

00:00:47,360 --> 00:00:51,670

Now if you look at the formula and think

about the question, how to generalize it,

10

00:00:51,670 --> 00:00:56,860

you'll realize that all we need is to add

more terms, like what you have seen here.

11

00:00:57,960 --> 00:01:04,020

So you can just add more thetas and

the probabilities of

12

00:01:04,020 --> 00:01:08,890

thetas and the probabilities of

generating d from those thetas.

13

00:01:08,890 --> 00:01:13,200

So this is precisely what we

are going to use and this is

14

00:01:13,200 --> 00:01:17,860

the general presentation of the mixture

model for document clustering.

15

00:01:19,810 --> 00:01:24,820

So as more cases would follow these

steps in using a generating model first,

16

00:01:24,820 --> 00:01:27,430

think about our data.

17

00:01:27,430 --> 00:01:30,360

And so in this case our data

is a collection of documents,

18

00:01:30,360 --> 00:01:33,740

end documents denoted by d sub i,

19

00:01:33,740 --> 00:01:37,310

and then we talk about the other models,

think of other modelling.

20

00:01:37,310 --> 00:01:41,410

In this case, we design a mixture

of k unigram language models.

21

00:01:41,410 --> 00:01:48,280

It's a little bit different from the topic

model, but we have similar parameters.

22

00:01:48,280 --> 00:01:52,396

We have a set of theta i's that

denote that our distributions

23

00:01:52,396 --> 00:01:55,810

corresponding to the k

unigram language models.

24

00:01:55,810 --> 00:02:01,260

We have p of each theta i as

a probability of selecting

25

00:02:01,260 --> 00:02:05,463

each of the k distributions

we generate the document.

26

00:02:05,463 --> 00:02:11,090

Now note that although our goal

is to find the clusters and

27

00:02:11,090 --> 00:02:16,450

we actually have used a more general

notion of a probability of each

28

00:02:16,450 --> 00:02:19,560

cluster and this as you will see later,

29

00:02:19,560 --> 00:02:25,610

will allow us to assign

a document to the cluster

30

00:02:25,610 --> 00:02:29,510

that has the highest probability of

being able to generate the document.

31

00:02:31,070 --> 00:02:35,530

So as a result,

we can also recover some other interesting

32

00:02:36,880 --> 00:02:40,520

properties, as you will see later.

33

00:02:42,390 --> 00:02:46,010

So the model basically would make

the following assumption about

34

00:02:46,010 --> 00:02:47,370

the generation of a document.

35

00:02:47,370 --> 00:02:51,130

We first choose a theta i according

to probability of theta i, and

36

00:02:51,130 --> 00:02:55,740

then generate all the words in

the document using this distribution.

37

00:02:55,740 --> 00:02:58,500

Note that it's important that we

38

00:02:58,500 --> 00:03:02,030

use this distribution all

the words in the document.

39

00:03:02,030 --> 00:03:04,770

This is very different from topic model.

40

00:03:04,770 --> 00:03:08,100

So the likelihood function would

be like what you are seeing here.

41

00:03:10,060 --> 00:03:16,620

So you can take a look

at the formula here,

42

00:03:16,620 --> 00:03:22,244

we have used the different notation

43

00:03:22,244 --> 00:03:28,810

here in the second line of this equation.

44

00:03:28,810 --> 00:03:33,837

You are going to see now

the notation has been changed

45

00:03:33,837 --> 00:03:39,102

to use unique word in the vocabulary,

in the product

46

00:03:39,102 --> 00:03:45,130

instead of particular

position in the document.

47

00:03:45,130 --> 00:03:50,750

So from X subject to W,

is a change of notation and

48

00:03:50,750 --> 00:03:58,580

this change allows us to show

the estimation formulas more easily.

49

00:03:58,580 --> 00:04:03,227

And you have seen this change also

in the topic model presentation, but

50

00:04:03,227 --> 00:04:08,191

it's basically still just a product of

the probabilities of all the words.

51

00:04:10,010 --> 00:04:10,900

And so

52

00:04:10,900 --> 00:04:15,100

with the likelihood function, now we can

talk about how to do parameter estimation.

53

00:04:15,100 --> 00:04:19,090

Here we can simply use

the maximum likelihood estimator.

54

00:04:19,090 --> 00:04:22,960

So that's just a standard

way of doing things.

55

00:04:22,960 --> 00:04:25,880

So all should be familiar to you now.

56

00:04:25,880 --> 00:04:27,890

It's just a different model.

57

00:04:27,890 --> 00:04:30,390

So after we have estimated parameters,

58

00:04:30,390 --> 00:04:34,060

how can we then allocate

clusters to the documents?

59

00:04:34,060 --> 00:04:37,740

Well, let's take a look at

the this situation more closely.

60

00:04:37,740 --> 00:04:41,850

So we just repeated the parameters

here for this mixture model.

61

00:04:43,030 --> 00:04:47,230

Now if you think about what we can

get by estimating such a model,

62

00:04:47,230 --> 00:04:52,640

we can actually get more information than

what we need for doing clustering, right?

63

00:04:52,640 --> 00:04:57,008

So theta i for example,

represents the content of cluster i,

64

00:04:57,008 --> 00:05:02,770

this is actually a by-product, it can help

us summarize what the cluster is about.

65

00:05:02,770 --> 00:05:06,020

If you look at the top

terms in this cluster or

66

00:05:06,020 --> 00:05:09,740

in this word distribution and they will

tell us what the cluster is about.

67

00:05:11,130 --> 00:05:16,010

p of theta i can be interpreted as

indicating the size of cluster because it

68

00:05:16,010 --> 00:05:21,310

tells us how likely the cluster would

be used to generate the document.

69

00:05:21,310 --> 00:05:24,750

The more likely a cluster is

used to generate a document,

70

00:05:24,750 --> 00:05:28,240

we can assume the larger

the cluster size is.

71

00:05:30,280 --> 00:05:32,880

Note that unlike in PLSA and

72

00:05:32,880 --> 00:05:36,640

this probability of theta

i is not dependent on d.

73

00:05:37,640 --> 00:05:41,520

Now you may recall that the topic

you chose at each document

74

00:05:41,520 --> 00:05:42,750

actually depends on d.

75

00:05:42,750 --> 00:05:48,720

That means each document can have

a potentially different choice of topics,

76

00:05:48,720 --> 00:05:54,260

but here we have a generic choice

probability for all the documents.

77

00:05:54,260 --> 00:05:58,950

But of course, even a particular document

that we still have to infer which

78

00:05:58,950 --> 00:06:01,840

topic is more likely to

generate the document.

79

00:06:01,840 --> 00:06:02,770

So in that sense,

80

00:06:02,770 --> 00:06:08,890

we can still have a document

dependent probability of clusters.

81

00:06:10,020 --> 00:06:14,890

So now let's look at the key problem

of assigning documents to clusters or

82

00:06:14,890 --> 00:06:16,320

assigning clusters to documents.

83

00:06:17,940 --> 00:06:22,587

So that's the computer c sub d here and

this will take one of the values in

84

00:06:22,587 --> 00:06:27,560

the range of one through k to indicate

which cluster should be assigned to d.

85

00:06:28,690 --> 00:06:32,985

Now first you might think about

a way to use likelihood on it and

86

00:06:32,985 --> 00:06:37,939

that is to assign d to the cluster

corresponding to the topic of theta i,

87

00:06:37,939 --> 00:06:41,090

that most likely has

been used to generate d.

88

00:06:42,450 --> 00:06:46,530

So that means we're going to choose

one of those distributions that

89

00:06:46,530 --> 00:06:49,500

gives d the highest probability.

90

00:06:49,500 --> 00:06:50,734

In other words,

91

00:06:50,734 --> 00:06:56,580

we see which distribution has the content

that matches our d at the [INAUDIBLE].

92

00:06:56,580 --> 00:07:01,870

Intuitively that makes sense,

however, this approach

93

00:07:01,870 --> 00:07:06,980

does not consider the size of clusters,

which is also a available to us and

94

00:07:06,980 --> 00:07:12,140

so a better way is to use

the likelihood together with the prior,

95

00:07:12,140 --> 00:07:16,038

in this case the prior is p of theta i.

96

00:07:16,038 --> 00:07:20,880

And together, that is, we're going to

use the base formula to compute

97

00:07:20,880 --> 00:07:24,230

the posterior probability of theta,

given d.

98

00:07:25,650 --> 00:07:30,058

And if we choose theta .based

on this posterior probability,

99

00:07:30,058 --> 00:07:36,010

we would have the following formula that

you see here on the bottom of this slide.

100

00:07:36,010 --> 00:07:42,390

And in this case, we're going to choose

the theta that has a large P of theta i,

101

00:07:42,390 --> 00:07:47,610

that means a large cluster and

also a high probability of generating d.

102

00:07:47,610 --> 00:07:51,690

So we're going to favor

a cluster that's large and

103

00:07:51,690 --> 00:07:54,982

also consistent with the document.

104

00:07:54,982 --> 00:08:01,090

And that intuitively makes

sense because the chance of

105

00:08:01,090 --> 00:08:05,720

a document being a large cluster is

generally higher than in a small cluster.

106

00:08:07,640 --> 00:08:13,000

So this means once we can estimate

the parameters of the model,

107

00:08:13,000 --> 00:08:16,930

then we can easily solve

the problem of document clustering.

108

00:08:16,930 --> 00:08:20,850

So next, we'll have to discuss how to

109

00:08:20,850 --> 00:08:25,512

actually compute

the estimate of the model.

110

00:08:25,512 --> 00:08:35,512

[MUSIC]