Good evening. We are Doctrina Apparatus. Our topic today is the author-topic model for authors and documents. The model essentially answers one question. If the text we see is generated by a machine, what does the machine possibly rely on? Or as you can see this a probabilistic model exploring the relationships between authors, topics and words.

You should keep in mind that topic here is a distribution over words. They are discovered from unsupervised learning, so that we don’t have to tell the algorithm which topic but rather how many topics are concerned in a document. Below are some examples. Topic 19 is constituted by likelihood, mixture, density and etc. each having the probability to be written when topic 19 is used. Topic 19 is identified by these words.

At the same time, the model explores the author’s interest into that topic, which can also be seen in the picture. Author Singer\_Y has probability 0.0281 to use topic 19.

This paper also compares the author-topic model with its ancestors. Namely the topic model and author model. In the author model, author’s interest is directly modelled with word for example author John may use the word machine with probability 0.0283 and the word learning with probability 0.0497. And in the topic model, there is topic distribution for a document and word distribution within that topic for example for document d the probability to use topic 1 is 0.05 and the probability to use the word machine when using topic 1 is 0.0823.

Let’s explore how would the machine generate text. We shall begin from topic model. First, a topic denote as z is sampled from a multinomial distribution denote as theta. Theta used Dirichlet distribution alpha as its prior. After we obtained topic z, we plug it into the word distribution phi which used beta as its prior to know how likely a word is used in z. Hence, we are ready to sample word w from phi.

Similarly, for author models, we interchange topic with author. Now an author is chosen randomly from author list Ad. The the chosen author x is then plug into the word distribution phi to generate word w.

And the author-topic model, kind of combined the previous two models. We first choose an author x just like the author model and then we have a new step, plug in x into topic distribution theta so that we know how likely an author will choose topic z. After we obtain topic z, the rest is just like the topic model.

So these are the three generative models, but keep in mind that we are not using the model for writing but for exploring the interest which is the distribution metioned. In order to estimate these parameters. This paper used gibbs sampling. The first step would be assigning words to random topic and authors. Next, apply equation 4 to every word in the text and make new assignment accordingly. We repeat this step 2000 times and finally we save word by topic matrix and author by topic matrix which is what we have seen.