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 only need to tell the algorithm how many topics are concerned. On the right are some examples, Topic 19 is constituted by word on the first column each having the probability on the second column.

At the same time, the model explores the author’s interest into that topic. For example, Author Tresp has highest the probability to use topic 19 among other authors.

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 0.08 and the word learning with probability 0.04. And in the topic model, there is a 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 0.08.

Let’s explore how would the machine generate text. We shall begin from topic model. First, a topic z is sampled from a multinomial distribution theta which used Dirichlet distribution alpha as its prior. After we obtained topic z, we plug it into the word by topic distribution phi using beta as its prior. Finally, we generate w from phi.

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

Finally, 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 plug in x into topic by author distribution theta and obtain topic z. After we obtain topic z, the rest is just like the topic model.

So these are the three generative models, but how do we the distribution. 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.