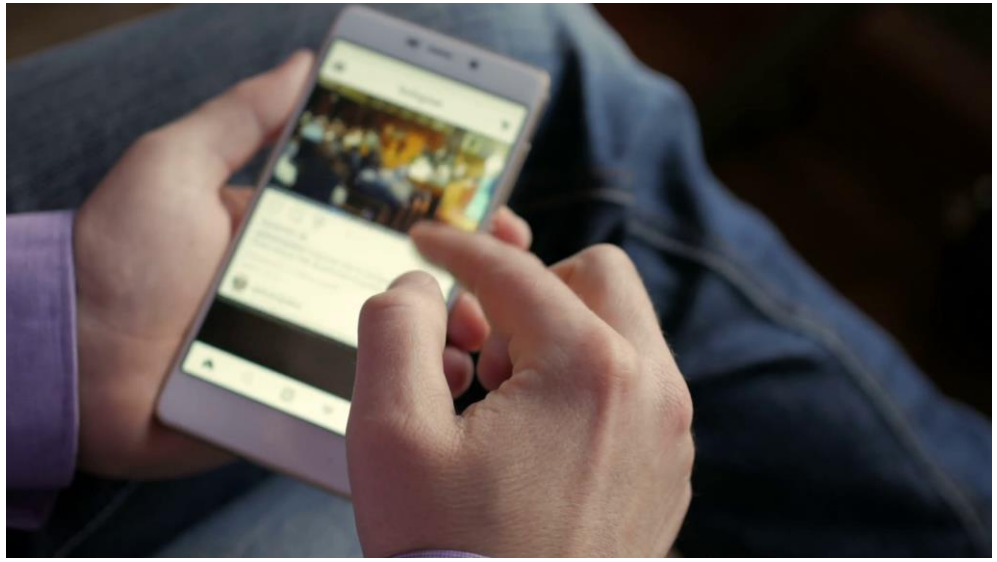
# Latent Dirichlet Allocation (LDA) Model Example

LEE WERNER

# The Problem:

We have a dataset that contains 50,120 articles from multiple news sources such as The New York Times, Atlantic, Business Insider, Breitbart, and CNN between 2015 and 2017. We want to ask one simple question:

**If we classified all of the articles into nine topics, what would they be?***



*If we wanted to instead find the top 10 discussed topics of the articles, we might instead set the number of topics to be 100 or 1000. However, for the purposes of this demonstration, we will classify all articles into just nine topics.

# The Solution:

Using Natural Language Processing & Machine Learning algorithms in Python, we can use our computer to find patterns recognizable to humans.

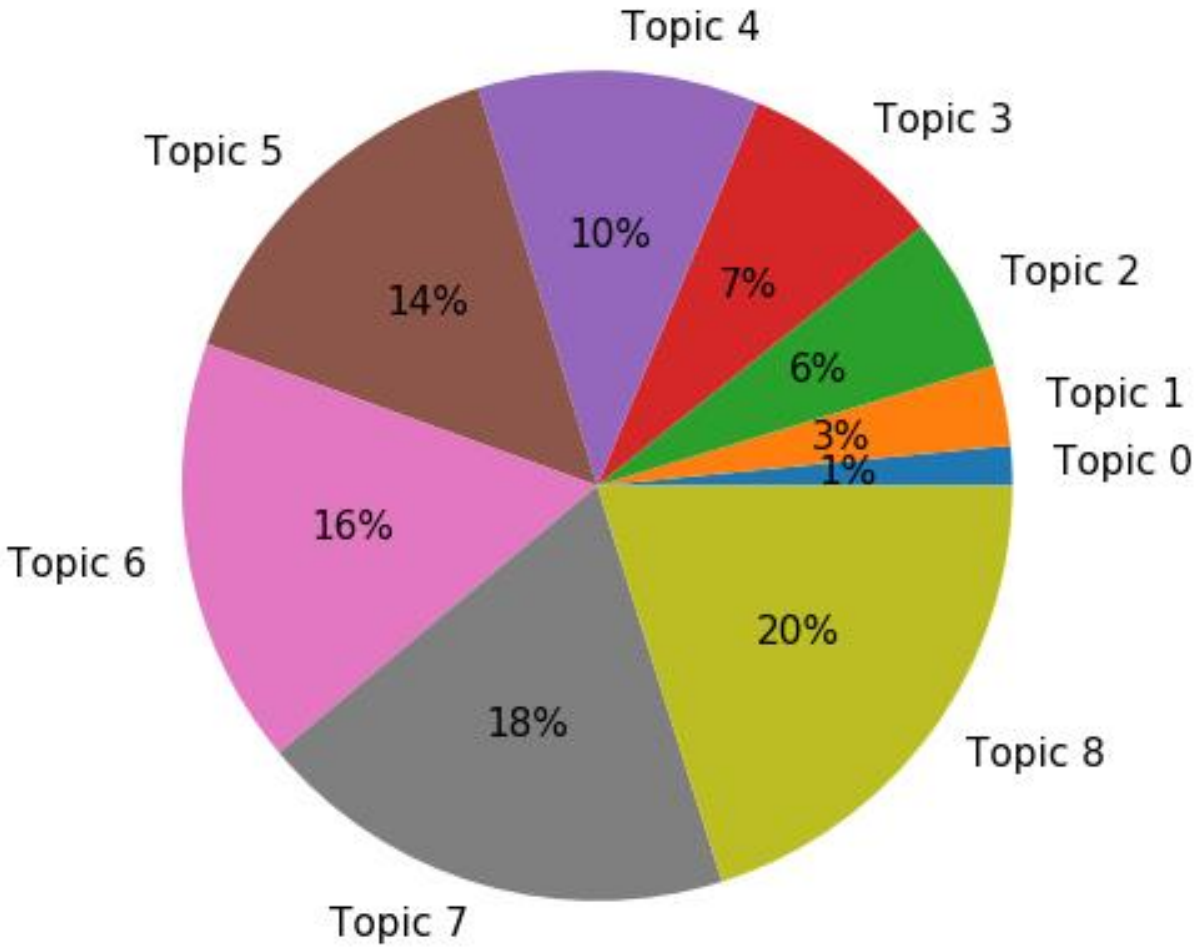
What algorithm should we use?

- There are many options, but here, we use Latent-Dirichlet Allocation (LDA) modeling.
  - LDA begins by assigning a random set of words found in the document to each topic, and then refines its classification through progressive iterations that tests itself against the document. The process continues until a “topic model” has been found that matches the data.

READ MORE ON LDA AT HTTPS://TOWARDSDATASCIENCE.COM/TOPIC-MODELING-AND-LATENT-DIRICHLET-ALLOCATION-IN-PYTHON-9BF156893C24

# The Results:

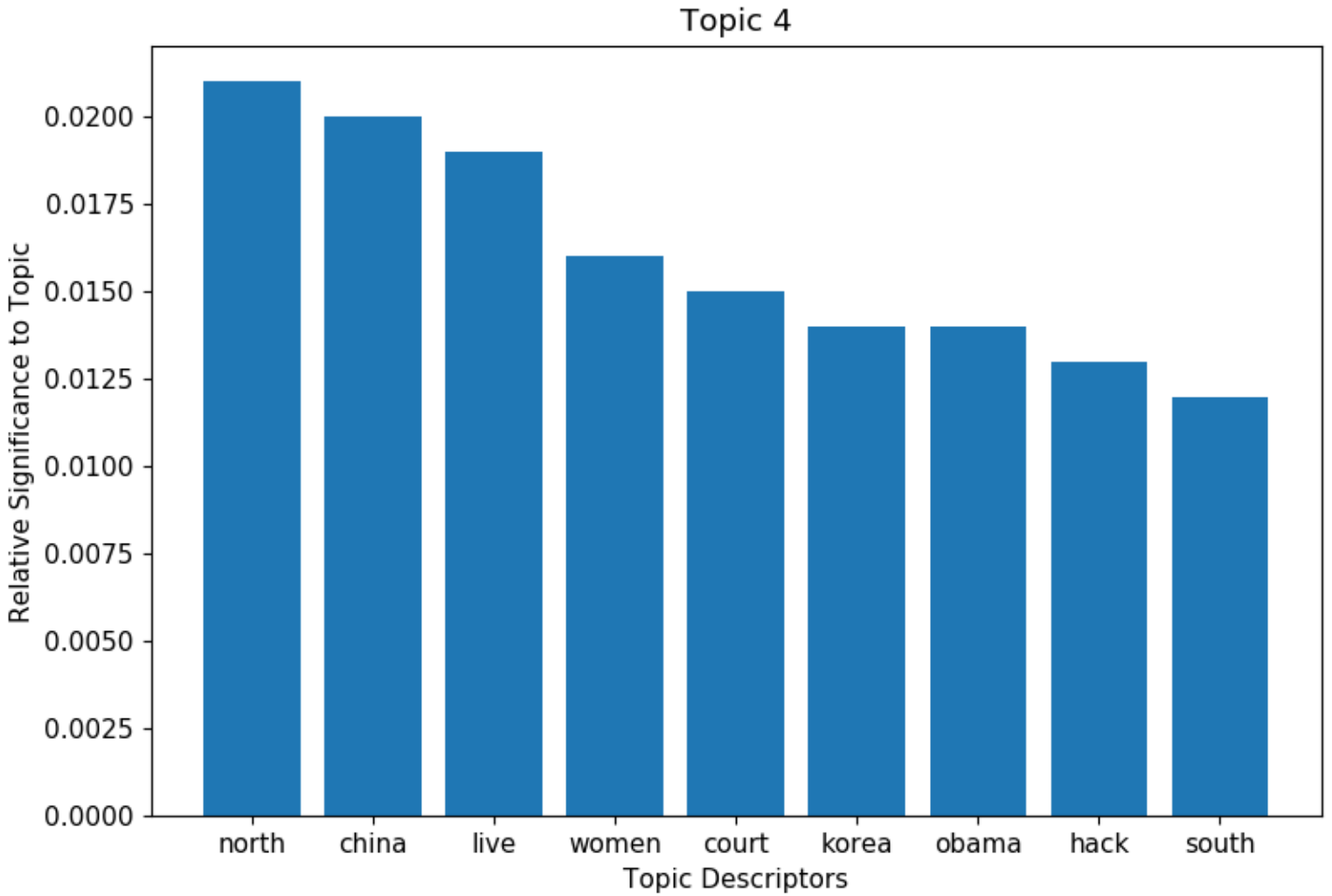
First, the model generates an un-intelligible distribution of topics in the form of a pie chart:



# Topic Descriptions

Then, we use the way that LDA has defined each topic via word associations with that topic to then define that topic in understandable terms.

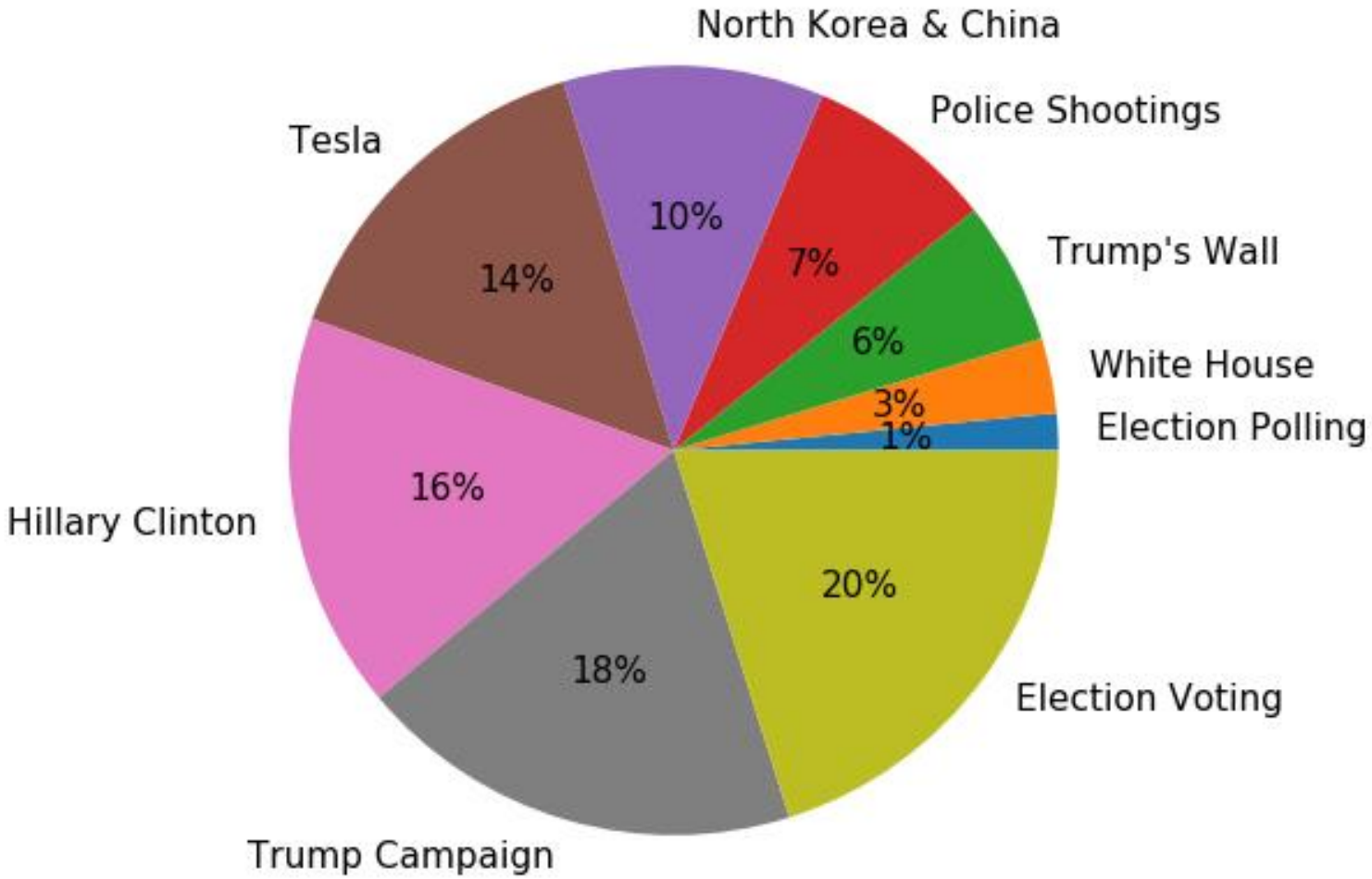
For example, topic 4 shown right: this might be something like “North Korea & China”.



# Rename the topics:

After looking through each topic definition, we can re-define our pie chart with topic definitions that make sense to us humans.

Now, we've transformed 50,120 article titles into nine topics – in just a few minutes.



# The Significance

Why is LDA important? Put more broadly, why is natural language processing & machine learning important?

**To find patterns in otherwise un-useable data.** Any scenario where there is a mountain of data that is in the form of language, natural language processing will be useful.

As one thinks about the opportunity & insight this might give their business, it's important to also recognize the ethical ramifications of using machine learning techniques on large sets of data. What if, instead of article titles, this was trans-scripted phone conversation data?