Slide # 1 – Title Slide (A Research Application of LDA Topic Modeling at Urban)

Thank you all for coming! I want to first thank Ian for helping me get on this meeting, Dan for allowing me to implement a model that I was really interested in for an Urban project, and Natalie for letting me practice my presentation in our office over and over again and giving me feedback.

I will be presenting on a Natural Language Processing algorithm called LDA Topic Modeling and an application of it on an Urban project.

Slide #1.b – NLP image (generic on google)

I will explain what LDA Topic Modeling is through the presentation, but I want to quickly explain what Natural Language Processing – or NLP - is a set of tools and algorithms used to analyze a large amount of texts. When I first began using NLP, I thought that my qualitative analysis of text would be entirely done with the machine. But what I learned while implementing and analyzing the results is that the machine is not doing the analysis for you, rather it is making the analysis easier and more efficient for you to do. With that being said, I’m going to go over a model I believe helps making qualitative analysis easier and more efficient.

Slide #2 – LDA Topic Modeling (title slide) – with full name of latent direclet allocation

* Fragment with: “or LDA”

Slide #2.b – What is LDA Topic Modeling

LDA Topic Modeling is a Natural Language Processing algorithm that seeks to transform a large amount of texts/documents into a comprehendible list of topics. Let’s see how this does that.

Slide #2.c – show that documents -> lda -> topics

Slide #2.d – documents

What does this look like? \*Show example of 10 basketball documents + 10 politics document\*

Slide #2.e – LDA

What does the model do when it reads a document? It count how many times the word appears in each document. It creates a term called tf-idf for each word. It looks at word co-occurrence to determine topics. Words that appear often with each other will be formed in topics.

In our example, we can imagine basketball, point guard, and 3-2 zone to appear In the same article.

Slide 2.f - Output

If we had 10 articles about basketball and 10 about politics and ran a topic model specyfing that we want 2 topics, we would get something like this: (Let’s actually do this, should be easy enough)

\*show 2 topic model with weights\*

Here we have the results of our topic model! The left hand columns are our weights associated with the words on the right hand column of each table. The weights represent the probability a word appears in a document that belongs to that respective topic. So, for example, if we have a basketball document with 100 words, we would expect \*\*\*\*.

Right now, I’m presenting the most heavily weighted words but every word is actually represented in a topic, just with varying weights.

Slide 2.g – visualizing the results

Show pyLDAvis! (with basketball and politics

Like I said, the topics are a series of probabilities associated with each word. So, technically we can run these probabilities with this formula to get a distance.

Then we can look at a distance matrix and apply principal component analysis on this to turn it into two dimensions

And we can plot the centers and adjust the bubble size for the prevelance of the document

So we see here are the basketball terms and here are the political terms!

Slide 3: Our application

One aspect of Dan’s STEM appreteinceship program is to see how the appreteinceship can supply sub-baculerate STEm fields. Community college can also do this.

So, naturally we want to compare them.

How do we go about doing this:

STEm Appreteinceship have RTI

STEm appreteicneships have OJT

We can easily compare the RTI

* OJT is much tougher

What if we use our topic modeling to help our analysis?

* Show documents then show lda then show topics

Explain analysis

Slide 4: Implementation

* Show how its easy

Slide 5: how to think about ML

* Imagine you have 1000 RA’s, what is the tasks you would ask them to do?

Slide 6: more math (show how the probability distribution and MDS work)

Slide 7: More NLP programs and what they can do

-word2vec

-doc2vec

-cosine sim

-sentiment analysis