Expo Talking Points

**Intro Slide**

Okay it is about 8:40, I’m going to go ahead and get started. Welcome everyone! Thank you for attending the intern expo. I am so excited to share what I’ve been working on this summer!

I am Laura Federline. You can see that I go to NC State. I am graduating this November – so with the schedule adjustments the semester will be wrapped up before thanksgiving. It will be here before I know it.

and I am majoring in Statistics with a minor in biology. Some fun facts about me is I am on the Pack Clogging Dance team at NC State and I studied abroad in the Czech republic earlier this year.

Here at SAS, I am currently a year-round intern in Health and Life Sciences R&D. This is my 3rd summer at SAS. You can read through some of the other things I’ve gotten to work on during my time here. If any of those interest you feel free to connect with me after the expo.

So now I’m going to dive into my summer project, if you have any questions or comments throughout feel free to throw those in the chat and I will answer those afterwards.

**Poster**

This summer, I have been working on a project that demonstrates how text analytics can be used to understand large amounts of content released about COVID-19. As we all know, a lot of information and research has been released on this topic at a really fast rate and there is a lot for us to parse through and keep up with.

* I’ll be showing how text analytics can aid in summarizing and understanding large amounts of text as well as bring patterns to light that may not have been found otherwise
* I’ll be going over where the text for my analysis comes from, how I prepared this text, and the results from my analysis.

\*ZOOM IN\*

Text Origin

* The text that I used for my analysis comes from the situation reports released by the WHO
* These are released daily, and are still being released, released one today
* Mine is 1st 100, this covers time period from January 21st – April 29th

Report Structure

* The idea is that there is a lot of important info to read through
* The text is semi-structured- which means the authors of these reports have a deliberate structure and content in mind. To an extent, we know what to expect from these reports. Also, all of the text falls under a header which allows the text to be categorized in this way which will be helpful later on
* Screenshot of one of the pages to give you an idea of what they look like
  + They are just like they sound – provide the most recent situation for coronavirus all over the world, including things like highlights, focus subjects, and recommendations for the public

Project Goals

Data Prep

* The first step in my Data Prep was to write a Python script using the PYPDF2 package
  + This is a really useful tool to automatically extract text from PDFs
  + And this also where I did a little bit of formatting to get a data structure
* Next, I imported this data into SAS Studio
  + 1st 2 bullets
  + And the next step was an iterative process between SAS Studio and VA. VA was used to explore the contents and check for any discrepancies in the text which I would then go back to my code and fix these issues

So now that the data is ready for analysis, we can start learning about the text. I’m going to show a couple screenshots to give an idea of what I’ll show in the demo right after

Text Topics

* This is a screenshot of the text topics object in Visual Analytics
* Text topics are what I use for my entire analysis
* A text topic is essentially a set of terms with different weights
  + Used to identify advisory themes
  + Really useful in quantifying the content in text
  + Each text doc is assigned a relevance for each text topic
    - This relevance represents to what extent does that topic represent the text document
      * This is helpful in identifying content change and even emphasis changes

Content Change

This is an example of how the relevance values can be used to identify content changes

Emphasis Change

Word Cloud

1st week. Vs. last week

Emphasis Change

Rec & Advice

Strategic Objectives

Next Tab

Counting cases

prevention