**Not the Corona Virus Team Member’s Information**

|  |  |  |
| --- | --- | --- |
| **Team Member Name** | **Email Address** | **Challenge** |
| Cesar Diaz | [calexdiaz143@gmail.com](mailto:Calexdiaz143@gmail.com) | 1 |
| Rhodora Villanueva | [rhodora33.rv@gmail.com](mailto:Rhodora33.rv@gmail.com) | 1 |
| Jaimie Gillette | [jaimie.gillette@gmail.com](mailto:Jaimie.gillette@gmail.com) | 2 |
| Timmy Ngo | [timmyngo021@gmail.com](mailto:Timmyngo021@gmail.com) | 2 |
| Randy Rodriguez | [randy.renezz@gmail.com](mailto:Randy.renezz@gmail.com) | 3 |
| Carlos Sanchez | [sanchezcj95@gmail.com](mailto:Sanchezcj95@gmail.com) | 3 |
| Annie Voigt | [a.voigt1121@gmail.com](mailto:a.voigt1121@gmail.com) | 3 |

**Overview of Approach**

The CDC guidance document was parsed to break the document into sections by topic. A series of data pre-processing steps were performed to transform the text data into vectorized representations. Tf-idf encoding with cosine similarity was used to find the most relevant section in the document for each prompt. An algorithm was used to identify the part of speech for each sentence within the document. The corpus was reduced to sentences starting with a verb, because these are actionable instructions or directives. A Latent Semantic Indexing model was trained to help create a merged set of general and specific instructions for each situation, reducing to a complete set by removing largely similar phrases.

**Technical Description of Solution**

Our approach can be summarized in 5 steps:

1. Parse the CDC guidance text document using “\*\*\*” to delineate headers and divide the document into topic-specific sections.
2. Prepare the text for follow-on machine learning steps. This pre-processing step involves:
   1. Removing common words that provide little meaning (aka stop words, such as the, it, and, of, to, he, she)
   2. Removing punctuation
   3. Removing capitalization
   4. Remove pluralization and words with 2 or fewer letters
   5. Tokenizing the sentences into words
3. Vectorize the titles and the scenario prompts using a term frequency–inverse document frequency (tf-idf) bag-of-words approach. Then compare each of the vectorized prompts with the titles using a cosine similarity measure to find the most relevant topic for each prompt.
4. Filter the resulting set of text to actionable instructions using a part-of-speech algorithm to only return sentences that start with a verb.
5. For each of the prompts (asthma, older adults, etc.) find a union set of instructions between the set of rules found in step 4 with the general guidance. Each situation-specific instruction was compared to each of the general instructions using a Latent Semantic Indexing model to measure the phrase similarities. A threshold was used to identify cases where the phrases were semantically similar enough to be considered redundant. When a general rule was not found to be semantically similar to the specific instructions, the general rules were appended to generate the final merged list of instructions for each scenario.

This pipeline was written in python with heavy reliance on the sklearn, nltk, and gensim python packages for natural language processing.

**Reason(s) for Choosing the Approach Taken**

With respect to the document parsing, we decided not to rely on the “@” tags within the text document because we felt that would not generalize well for future problems and did not require any Machine Learning approahes. We also interpreted the readme instructions to mean that the “@” symbol provided some examples, but was not an exhaustive list of instructions within the CDC guidance that was important for each situation. For example, for older adults, there is an instruction to “Visit with your friends and family outdoors, when possible.” This is not annotated with an “@” but is important guidance that older adults should know. We felt it could be considered unethical to not return all actionable guidance.

Instead, our team attempted to create a more robust algorithm to find instructions amongst all the text, without relying on these special annotations. We did this using a part-of-speech tagging algorithm and filtered the text to only those sentences that start with a verb. Sentences that start with a verb are called imperatives and represent a command or directive. We used this to limit all the text to sentences that provided actionable information.

As with any ML problem data pre-processing is an essential first step and can make or break all follow-on steps in the pipeline. NLP text preparation is especially important and we followed common standards of practice to remove stop words, punctuation, capitalization, and vectorize via tf-idf. Given more time we would have also lemmatized or stemmed the words to translate the words into their root word (i.e., running -> run).

When finding the most relevant title for each prompt, we relied on the cosine similarity measure to compare the phrases after vectorizing via tf-idf. This is an appropriate similarity measure because the number of words in each phrase does not impact the distance measure (there is no need to normalize based on sentence length). The tf-idf weights less frequent words higher thereby giving more weight to the important keywords in the encoding. For example, the word “newborn” will be appropriately weighted higher than “sick” for the tf-idf vectorizer trained on all the text in the CDC guidance document because sick is a relatively common word in this corpus compared to newborn.

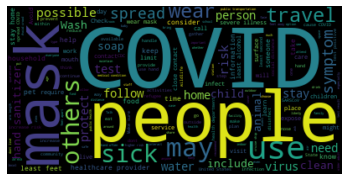


Figure 1: Word cloud generated from CDC guidance text document.

To find the unique union of both specific and general instructions we relied on the Latent Semantic Indexing (LSI) model. We felt this would be more effective at capturing the semantic meaning of phrases rather than the literal words.

This approach should scale well to larger datasets and runs in 1.56 seconds on a standard laptop.

**Ethical Considerations**

If this algorithm were to be fielded in the real-world, there would need to be a massive and unavoidable disclaimer that says all recommendations should be thoroughly vetted by a medical professional. Each individual’s risk profile needs to be carefully considered before accepting the results from this algorithm. Your doctor is the best source of personalized guidance.