**Note:**

**This is not a resubmission**

**It is entirely new**

**My first proposal was DOA**

**BUILDING A CLIENT/DOMAIN SPECIFIC SUMMARIZER**

There are a lot of text summarizers available but the well-known ones are general text summarizers – not domain specific. I will build a domain specific summarizer for Covid-19 research/information. This will start with building a domain specific vocabulary and then create a word-embedding for the vocabulary.

This is of value to any researcher, doctor or the general public. Papers are being published at a phenomenal rate since this worldwide crisis started. No one has the time to read even a fraction of them, let alone find them.

My source of data is the web. I will scrape about 40 journal level articles to start building my summarizer.

I will first amass and tokenize my data.

I will then normalize it.

Next step will be vectorized word-embeddings using Bert.

Next, I will build and train an RNN model using Keras with cross-entropy-loss as the performance metric

Then I will take an unseen text(s) and summarize it using:

1 - My summarizer

2 - Gensim summarizer

3 – spaCy summarizer

I will compare/contrast all the summarizations

This is fundamentally a comparison between an abstractive model (mine) and extractive models based on TextRank (Gensim & spaCy)

**The biggest challenge will be evaluating the summaries!**

There is no best way.

Human judgement is the Gold Standard.

You can use BLEU or ROUGE but those are primarily intended for translation.

You can use one (or more) of the following metrics:

Language Diversity

Correct Style

Semantic similarity

Length

**Milestones:**

1 – Amass the data

2 – Normalize and tokenize

3 – Vectorize the tokens

4 – Build RNN model using Keras

5 – Train the model

6 – Run my summarizer and preexisting summarizers on unseen text

7 – Evaluate the summaries

**What kind of input will I take from the user?**

I am not clear as to what this question is asking?

You can receive or use input -- you can’t take input.

I assume it means what might a client ask me to do?

A client may have many documents that need to be summarized for online access or internal use.

Say they have dozens of blogs or whitepapers.

Maybe a database of technical instructions.

I can build a custom summarizer.