AUTOMATIC TEXT SUMMARIZATION

KDM Project Submission 2 - Spring 2021

Assignment 2 Data and Models

Team number : 4

Team Members :

Charles Scola, Bayard Rucker, Vyoma Desai, Claire Ndofor

Motivation:

Our motivation is driven by the interesting topics we came across during our knowledge and discovery management course and how convenient some of the tools we learned made work easy and time saving. With the text summarization idea , we will be able to accomplish a similar goal by making research easier for students and industries. Inorder to obtain our goal, we ask ourselves the following questions to get an objective

- Can we create a summary with the major points of an original document?
- Abstractive (write your own summary) and Extractive (select pieces of text from original) are two popular approaches

CHAPTER 2 (DATA) :

Data should guide the process of identifying data sets that are most relevant to the real-world problem in Assignment 1:

Please answer the following question about Data:

1. **Who** is the data set about?

   This dataset contains the documents and accompanying questions from the news articles of CNN. There are approximately 90k documents and 380k questions

2. **Who** were sampled in this data set?

   the sample is taken from the news organization CNN

3. **Who** were over sampled or under sampled? Are they representative of the main characters in Assignment 1?

   there does not appear to be any issue with over or under sampling

4. Is there any identifiable information or is there any risk of disclose identifiable information? This is fundamentally about the sampling issue, and anonymity.

   Yes, the author of the article

5. **What** events, activities, behaviors, and observations etc. are recorded by the data set?

   The data set is broken into two subsets: questions and stories.

6. Does the data set record the targeted events, activities, behaviors, etc. in Assignment 1? This is fundamentally about the variables.

   The data set contains the information needed to solve the main questions asked in assignment 1.

7. **When** did the event, activity, behavior, and observation, etc. take place? **When** were the data collected?

   this data set appears to have been created in 2015

8. Is it longitudinal or cross-sectional?

   no

9. Are they real time data?

   no

10. How old or fresh are the data? To what extent generalization can be made across time to inform Assignment 1? This is fundamentally about the

temporal structure of the data set, and the external validity of the data set across time.

The dataset is roughly 6 years old. The age of the dataset does not appear to be an issue. However if a dataset set of for example shakespeare plays. this would have an effect on our question because of the change in language over time.

11. **Where** did the event, activity, behavior, and observation, etc. take place? **Where** were the data collected if the information is available? What does the geographical coverage of the data set look like?

    no information could be found on this. Given the nature of the work CNN does its possible that the set covers global events but more likely that its focused on the United States.

12. Does the data set contain geographical information (GIS)? Is this a local, regional, national, or global data set?

    no

13. To what extent generalization can be made across settings to inform Assignment 1? This is fundamentally about geographic variables in the data set, and the external validity of the data set across settings.

    Generalization is not as prevalent in this data set (unless you count only english articles being a form of generalization) since it contains thousands of different

14. **Why** did the event, activity, behavior, or observation etc. take place? **Why** were the data collected?

    All of the articles in the dataset were written in response to real world events. This data was originally collected to build a different model.

15. **How:** If you would like, you can add a dimension of how. How did it happen? Sometimes, the answer to how can be covered by what, when and where.

nothing to add for how.

CHAPTER 3 (THE SCIENTIST AND AI) :

The objective of Chapter 3 is to select or create the most relevant ML algorithms given chapter 1 and Chapter 2. The data sets should be properly aligned so that they can be used in ML. Accordingly, Chapter 3 describes a story about data scientists (the students) and AI together help the people/community find the solution. The main story line is about iterative experimentation that data scientists and AI conduct collaboratively along the journey of "rescuing" people in need.

"Chapter 3 the Scientist and AI" is a story about the data scientist and AI' selecting the most relevant ML algorithms given Assignment 1 and Data:

1. **Who:** In this story, there are three main characters: 1) the people/the community who needs help, 2) the data scientist (that is you), and 3) AI.

   Please answer the following questions:

2. How much does the data scientist understand Assignment 1 (domain) and data?

   the data scientist are fairly new to this domain of and data

3. **What** models and analysis did the data scientist and AI apply to fulfill the needs of the people or the community?

   While we have considered a handful of different models. Our current approach is to use. use a seq2seq model.

4. Can the data scientist estimate and select data for their goals from Assignment 1? Can they map data sets from Assignment 2 onto appropriate ML models?

   Yes, data scientists have a well understood goal from chapter 1. and the selected dataset would map on to our model without a major issue.

5. Can the data scientist connect Story 1 with ML models/stories about what a ML model can do? To perform good ML research, what in-depth knowledge and experience with ML algorithms and ML stories does a data scientist need?

   Yes part 1 can be easily connected to general NLP problems. the experience and knowledge most useful to the data scientist for this model. its data wrangling and NLP

6. **When** has to do with the iterations, how much time did it take for experimentation? How efficient is the modeling/algorithm?

   We are still in the process of experimentation with our model.

7. Can the data scientist determine the acceptance level of the model (validation with accuracy and runtime performance) considering the targeted users?

   At this point we still can't determine the accuracy of our model.

8. **Where** has to do with the learning environment. Where did this experiential learning process take place? For example, it was part of an online Deep Learning course.

   Our experiential learning process has drawn a lot from kaggle.

9. **Why** explains the modeling. Explain the ML model you are using?

   Our current approach is to use. use a seq2seq model. This is a model that's called an encoder-decoder. It maps input sequences to output sequences. The basic idea is to use 2 recurrent neural networks that work together to try and predict the sequences.

10. **How**: If you would like, you can add a dimension of how. How did it happen? Sometimes, the answer to how can be covered by what, when and where.

    nothing to add for how