Topic: Contextual Text Analysis using ML/AI with Python

One of the most challenging tasks in NLP is the identification of the features of data that predict our target. Text data provides many opportunities to extract features and parse the texts to extract morphological, syntactical and semantic representations from the data. There are two ways to use text analytics.

1. Analyzing text that exists, such as customer reviews, gleaning valuable insights
2. Structure the text so that it can used in ML models to predict future events

The business problem trying to tackle here is regarding context aware text analysis.This text generations can be thought of having wide applications. This could be helpful in solving business problems in the following areas and more. Some areas that I can think of are

- Regenerating contents that has been damaged or lost
- Machine translation
- Chatbots
- Smart autocomplete
- Understanding the patterns of behavioral disorders in social networks

Business Problem / Questions

I am planning on using and building on top of the course project work done as part of our DSC650. This is the Enron email text analysis. I am interested to classify the emails that was released for general public. My intention is to bring out the answers to the following key questions:

a. What was the sentiments expressed in the emails?
b. How much of pessimistic or optimistic mood do they show?
c. Is there any mention of fraudulent activities mentioned in the emails?
d. How to classify the emails, if possible to classify?

Datasets

- https://www.kaggle.com/code/scollins/enron-email-analysis/data
- https://new.pythonforengineers.com/blog/analysing-the-enron-email-corpus/
- dsc650/assignment04
- https://www.kaggle.com/code/scollins/enron-email-analysis

Methods

I plan to start by exploring the linguistic properties of a specific language can give us the quick ability to perform statistical computation on text. I will also try to understand how context modifies interpretation.

Data engineering steps involved in the process

- Importing dependencies
- Loading data
- Creating character/ word mappings
- Data preprocessing
- Modelling
- Generating text

Data Modelling steps in more details:

- Use grammar to extract key phrases from the documents
- Explore n-grams and discover significant collocations of words to enhance the bag-of-words model

Ethical Considerations

As a data scientist and data practitioner I am interested in building language aware data products. For that I am interested to build applications that accept text data as input, parse it into composite parts, compute upon these composite parts and ten recombine them in a way that delivers a meaningful and tailored results.

Identifying gender biasedness in News Articles – This analysis was done by Neal Caren in 2013, an Asst. Prof of Sociology at University of North Carolina Capel Hill.

Semantic information extraction and enrichment from social documents can serve as a vital asset to explore early signs of behavioral disorders and help prevent serious issues like cyber-bullying, suicidal related behavior and radicalization.

Challenges / Issues

So far as I can understand I can think of the following challenges that I would like to solve with this approach

- Data cleaning
- Extracting the free form email messages from the corpus
- Contextual feature extraction
- Augment models with grammars to capture meaning of specific type of phrases

References

1. https://www.analyticsvidhya.com/blog/2018/03/text-generation-using-python-nlp/
2. Applied Text Analysis with Python – Benjamin Bengfort, Rebecca Bilbro & Tony Ojeda
3. Neal Caren, Using Python to see how the Times writes about men and women, (2013) http://bit.ly/2GJBGfV
4. https://new.pythonforengineers.com/blog/analysing-the-enron-email-corpus/