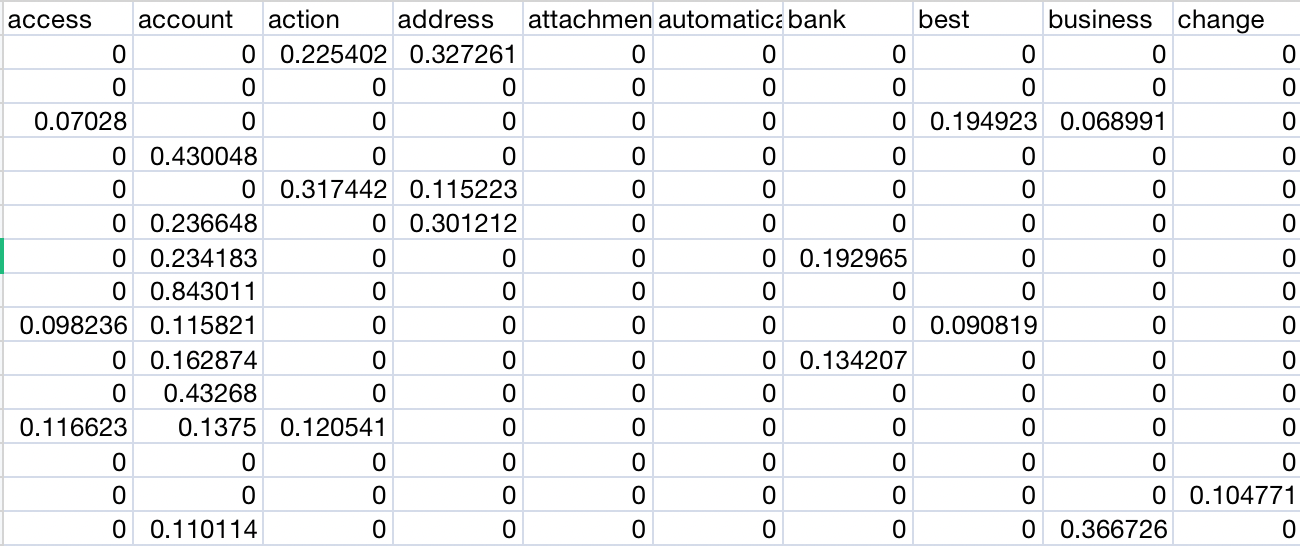
Data preprocessing:

1. Remove uninformative rows: drop rows with null Subject or Body
2. Text cleaning:
   1. Remove uninformative sentences, like “Notice: This message was sent from outside the University of Victoria email system. Please be cautious with links and sensitive information.”
   2. Remove /r, /n
   3. Handle email and url addresses, since it’s not strongly related to english words
3. Text tokenization
4. Remove stopwords (NLTK)
5. Lemmatization (root words)

Feature Extraction:

1. Combine ‘Subject’ and ‘Body’ into a single text for an email
2. Applied TF-IDF (Term Frequency-Inverse Document Frequency) vectorization to convert text into numerical features.

...