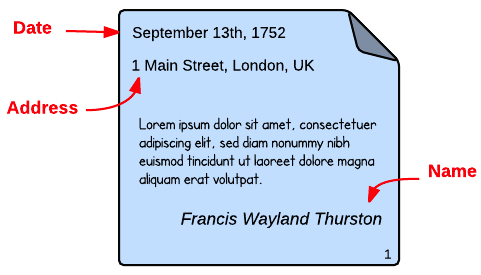
# ML Assignment

The data contains features extracted from text similar to the one shown below.



You have to create a ML model that predict the probability that a piece of text belongs to a particular class. Use techniques like Bag of Words, tf-idf vectorization and word embedding. Please use Hash field value and explain how you are going to use the Hash field.

# Data extraction

Fro the documents nGrams have been extracted, Each row in the Train.csvcorresponds to one such nGram.

# Features

For a given nGram several features have been extracted (145). These features have been saved in the train.csvand test.csv. They have parsing, spatial, content and relative information.

* Content: The cryptographic hash of the raw text.
* Parsing: nGram is a number, text, alphanumeric, etc.
* Spatial: Position and size of the nGram
* Relational: details of text nearby the nGram

The feature values can be:

* Numbers. Continuous/discrete numerical values.
* Boolean. The values include YES (true) or NO (false).
* Categorical. Values within a finite set of possible values.

# Labels

This are the labels corresponding to the probability that the current sample belongs to the given class. This is multilabel problem and hence a given sample can belong to more than one class.

# File descriptions

All the files are CSV.

**train.csv**- the features *x* x  of the training set. Each row corresponds to a different sample, while each column is a different feature.

* trainLabels.csv - the expected labels *y* y  for the training set. Each row corresponds to a different sample, while each column is a different label. The order of the rows is aligned with train.csv.
* test.csv - the features *x* x  of the test set. Each row corresponds to a different sample, while each column is a different feature.

sampleSubmission.csv - example of the expected probabilities *y* ̂   y ^    for the test set. Each row contains two columns, namely one string and the probability of each sample belonging to each label. For example, if the test.csv has 3 samples and 4 labels, the submission file must have 13 rows with these strings in the first column: *id\_label*, *1\_y1, 1\_y2, 1\_y3, 1\_y4, 2\_y1, 2\_y2, 2\_y3, 2\_y4, 3\_y1, 3\_y2, 3\_y3, 3\_y4, 4\_y1, 4\_y2, 4\_y3, 4\_y4*