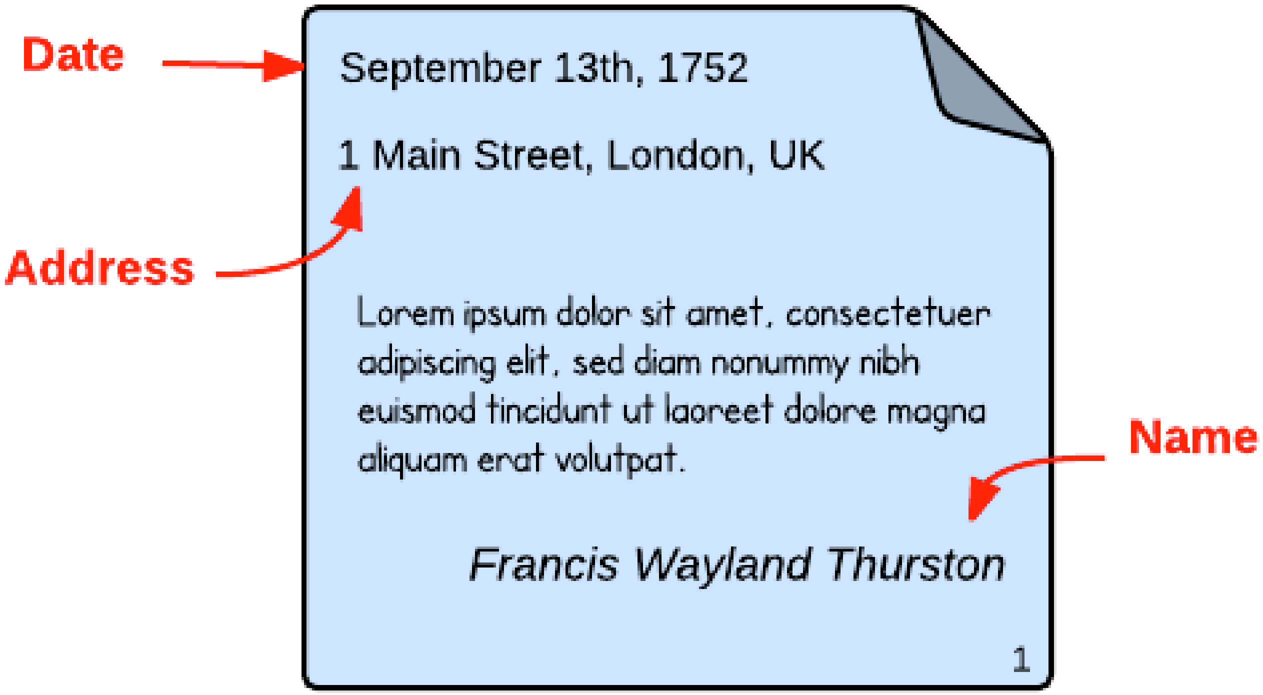
ML Assignment

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



September 13th, 1752

1 Main Street, London, UK

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Francis Wayland Thurston

1

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*□ˆ 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*