Getting Started with Text Classification

PyData Global Conference 2021

Nabanita Roy

Predict if Tweets are about Real Disasters

About the dataset

Exploratory Data Analysis

Text Pre-processing

Training a Binary Classifier

Model Evaluation

What is Natural Language?

What is Natural Language Processing?

Where does NLP fit in the AI ecosystem?

Do you want to build machines or computer applications?

Technological Goals

Do you want a machine to listen and act like humans do?

Cognitive Goals

Text Classification Using Supervised Machine Learning

Assigning **pre-defined** categories to text documents.

Target / Label is the term for the pre-defined categories.

Example:

- Is an email spam?
- Is a news article about politics, business, or sports?
- Gender Identification
- Sentiment Analysis – If a review is positive or negative?

Text Processing Techniques

- Regex and text extraction
- Tokenization
- Case Conversion
- Noise Removal
  - Accents
  - HTML Tags
  - Symbols/ Emojis
- Contractions (Example : I'll -> I will)
- Stopwords (Example : are, is, the)
- Normalization (Example: visited -> visit)
- Parts-of-Speech Tagging
- Named Entity Recognition

Text Processing Techniques

Example:

I just saw Barrack Obama in !

-> [‘I’, ‘ just’, ‘see’, ‘barrack’, ‘obama’]

<p> I’ll see you soon ☺ <3, mon chéri <p>

‘I’, ‘will’, ‘see’, ‘you’, ‘soon’, ‘mon’, ‘cherie’

Text Vectorization or Text Representations for Machine Learning

Machine Learning models quantifies everything.

Therefore, numeric representation for texts is required as input an ML model.

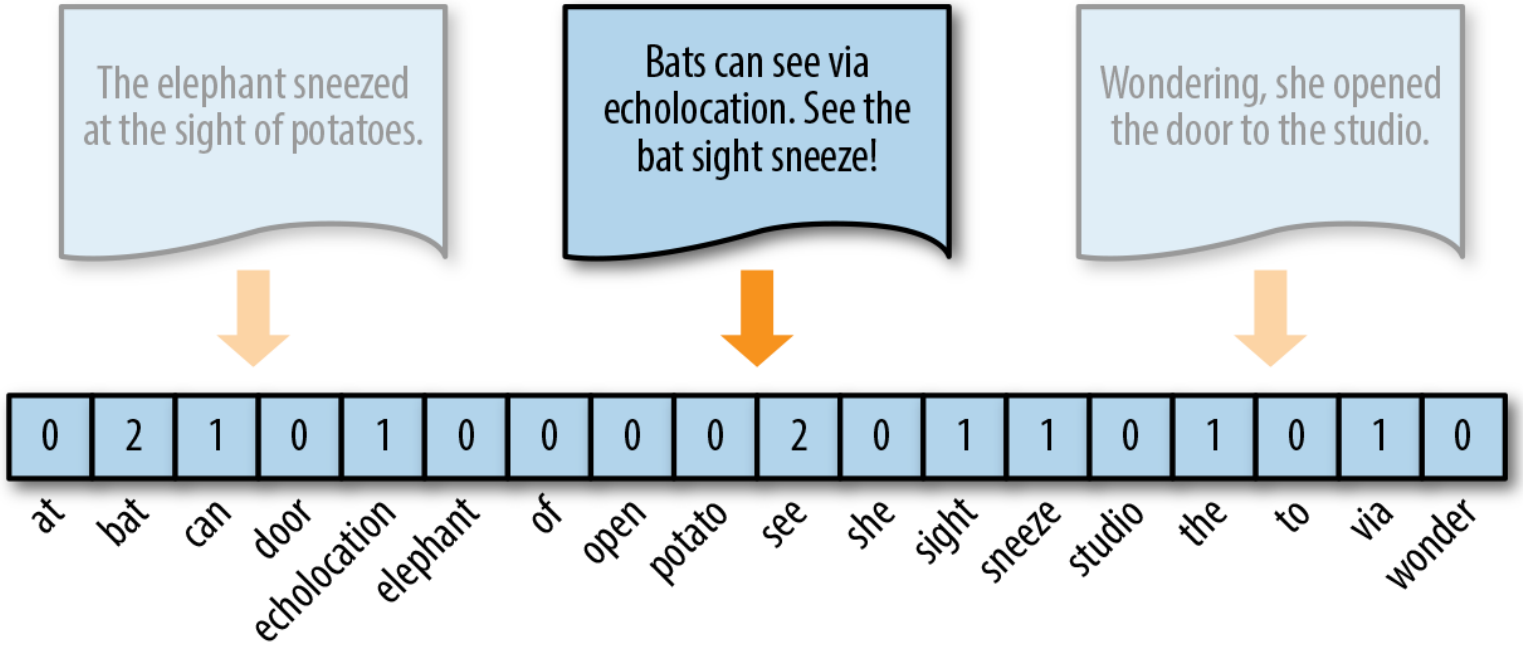

Image ref: https://towardsdatascience.com/from-word-embeddings-to-pretrained-language-models-a-new-age-in-nlp-part-1-7ed0c7f3dfc5

Statistical Models: Bag of Words

In a Bag of Words or BoW, bag refers to an unordered list of words which allows multiple occurrences of the words. Position of the words is ignored and Frequency (the number of occurrences) of a token is considered.

Image ref: https://sep.com/blog/a-bag-of-words-levels-of-language/

Statistical Models: N-Grams

Takes into account N tokens occurring in a sequence.

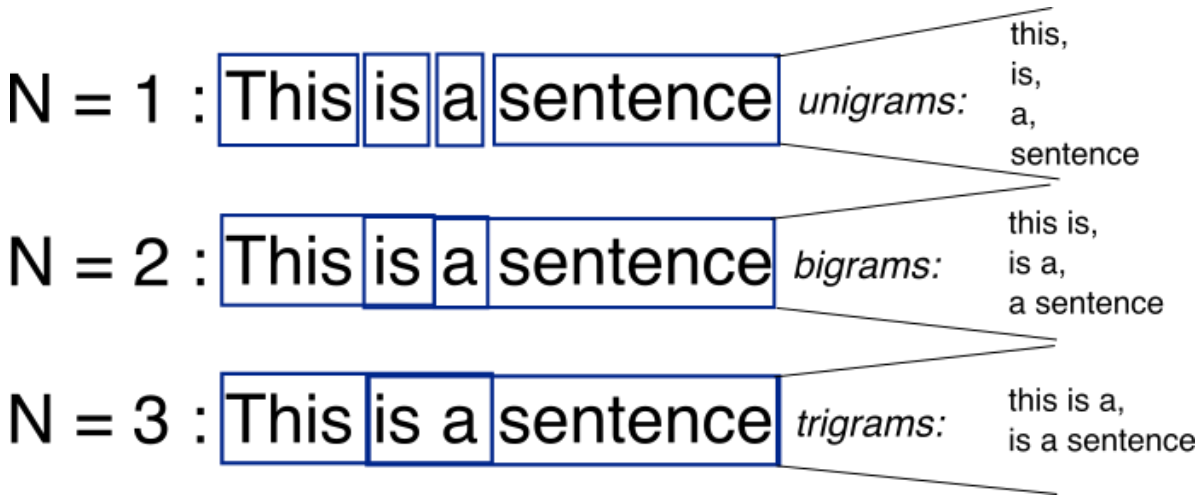

Image ref: http://rstudio-pubs-static.s3.amazonaws.com/460514_4377b4f645a944d788ae7300782123f3.html

Statistical Models: TF-IDF

Documents are converted to vector models (or vectorized form) using the number of the times a token appears in **one** document and in **all** the documents.

Order is ignored.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

Use log to dampen the effect of large corpus

(df+1) is used instead for terms that do not occur in the vocabulary to avoid Zero Division.

Image ref: http://filotechnologia.blogspot.com/2014/01/a-simple-java-class-for-tfidf-scoring.html

Model Evaluation for Text Classification

		Predicted: Negative	Predicted: Positive
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{All Samples}}$$

Model Evaluation for Text Classification

		Predicted: Negative	Predicted: Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Recall

Precision

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision + Recall}}$$

Thank You

Q/A