Handling Text Data - IPBA 10

23 April 2022 13:17

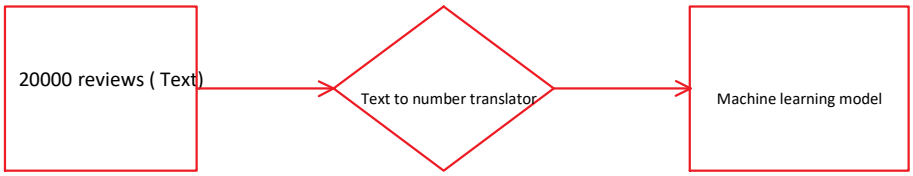
How can we define text ? What is text data
- Text is nothing but data present in the form on natural language. (NLP)
  - English
  - Hindi
  - Japanese
  - Local language.
- Natural Language Processing (NLP)
- Audio and Video Analytics is also NLP problem.
  - Translation of Audio/Video in Text (Regular NLP Modules)
- Image is also NLP ?
  - Use cases where you need to extract the text from the image and then some processing on that text -> All those use cases are NLP. (OCR = Optical Character Recognition) Google Vision API can do this OCR. Amazon Textract.

Popular sources of text data :-
- Social Media - LinkedIn, Twitter and Facebook
- Emails
- Customer Reviews/Feedback/Comments
- News
- Blogging Websites
- System (IT) Application logs
- Transcript
- Medical Records in hospital
- Research Papers
- Resumes
- Corporate LAN
- Customer Support function
  - User Tickets
  - Customer calls
- Digital Books

How to collect text data (Social media data)
- Web scrapping or Web crawling (Not ethical)
- APIs (Think of API as a package provided by organization)
  - To get twitter data in Python (Tweepy)
  - To get facebook data in R (Rfacebook)
    - With free version only a limited set of data
- Social Media Listening tools - Radian6, Buzz Matrix. (Social Media aggregators)

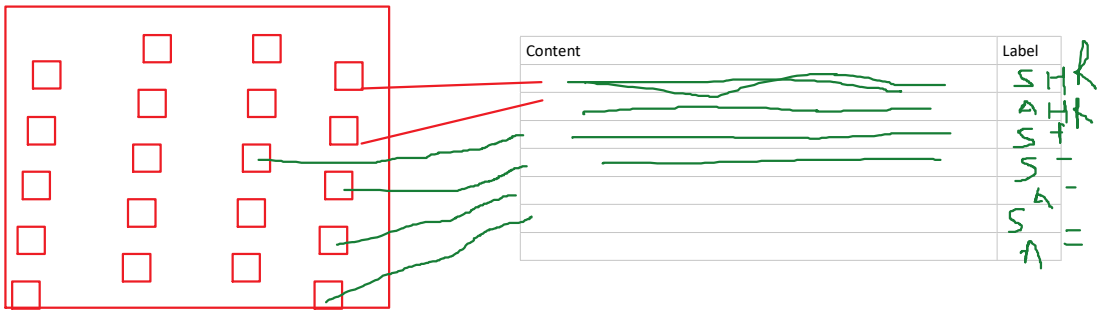
Applications that you can build on text data ?
- Sentiment Analysis of reviews
- SPAM Detection
- Fake News detection (Research Topic)
- Generate leads from the social media data
- Automatic classification of support tickets
- Over the call identify if the customer is happy with the conversation or not.
  - Or also flag the need of support.
- Automatic screening of resume.
- Group similar documents/reviews together.
- Can we build an AI agent to predict the story point of a story at the time of estimation (JIRA)
- Trolling detector

Why we can't use text data as it is in machine learning models ?
- Machine learning models understands ONLY NUMBERS.



Text to number translators :-
- Count Vectorization, Tokenization, Bag of Words (BoW)
- TF-IDF (Term Frequency & Inverse Document Frequency)

- Cosine Similarity



Corporate LAN/Folder with multiple files

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
	Text Reviews	**Sentiment**	Tokens	India	economy	is	growing	sale	of	fast	food	decreasing	Average	age	INDIA	population	has	decrease	Label
	India's economy is growing India	Good	{ India, economy, is, growing}	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	Good
	Sale of fast food is decreasing	Good	{ Sale, of, fast, food, is decreasing	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	Good
	Average age of INDIA's population has decrease	Bad	{ Average, age, of, INDIA, population, has decrease}	0	0	0	0	0	1	0	0	0	1	1	1	1	1	1	Bad

Each review is called as **document**
Collection of documents is called as **Corpus**.
Each individual word is called as **Token** or **term**.

Text data pre-processing steps:-

- Create tokens from the text
  - Convert all words in proper case (lower case conversion)
  - Remove special characters
  - Remove stop words (commonly used word in the language)
  - Remove digits
  - Remove white space
  - Language translator (to have all review in one language)
  - Remove duplicate tokens
  - Spelling mistake correction (Thanks, Thnaks)
  - Emoticons
    - Can we get the textual meaning behind the emoji.
  - Slang (Thanks, Thnks, Thx…)
  - Stemming (Blind - To reach to the root word)
    - Walks - Walk
    - Dances - Dance
    - Plays - Play
    - Running - Run
    - Lying - Ly
    - Studies - Studi
    - Decreasing - Decreas
  - Leammatization
    - Decreasing - Decrease
    - Lying - Lie
    - Studies - Study
    - Recommendation - Recommend
    - Best - Good

-
```
['This is sentence one', 'This is sentence two', 'This is sentence three']
```

Six features in the dataset.

Review	Class
This is sentence one	1
This is sentence two	2
This is sentence three	3

-

	is	one	sentence	this	three	two
0	1	1	1	1	0	0
1	1	0	1	1	0	1
2	1	0	1	1	1	0

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	Review	Label		This	is	Sentence	One	Two	Three	Label		One	Two	Three	Label
	This is sentence one	Class-1		1	1	1	1	0	0	Class-1		1	0	0	Class-1

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Review	Label		This	is	Sentence	One	Two	Three	Label		One	Two	Three	Label
This is sentence one	Class-1		1	1	1	1	0	0	Class-1		1	0	0	Class-1
This is sentence two	Class-2		1	1	1	0	1	0	Class-2		0	1	0	Class-2
This is sentence three	Class-3		1	1	1	0	0	1	Class-3		0	0	1	Class-3

Limitation of Count Vectorizer approach is that it gives equal importance to all the features.

- If a word is appearing many times in a document. This word is important to classify this document.
  Term Frequency
- If the same word is appearing rarely in other documents.
  Inverse document frequency

TF-IDF approach - this is improvisation of Count vectoriser approach.

- Uniqueness ?