Text preprocessing level -1 :

Here the data may include text (categorical columns) but those with simple text data with repetitive words and with simple and single word can be handled with encoding techniques which can replace given word with some unique numerical value (rank wise) or with an vector of numericals.

So what if a column is a review or feedback or question or compalaint about any application. As a human we can understand it. And for machines it need to be a numerical. (It includes different approach to convert)

Eg: siri in iphone, Alexa of amazon, (spam classifier in Gmail)

NLTK is a very good library.

“import nltk” to import nltk library

And “nltk.download()” to download all its dependencies or whole repo from github.

Or ‘nltk.download(“punkt”)’ to download only what ever is required.

**Tokenization:**

convert Paragraph in to list of small sentences using below code; nltk.sent\_tokenize(paragraph)

convert paragraph or sentence in to list of words using following code:

nltk.word\_tokenize(paragraph)

.

Eg:

paragraph = """I have three visions for India. In 3000 years of our history, people from all over

the world have come and invaded us, captured our lands, conquered our minds.

From Alexander onwards, the Greeks, the Turks, the Moguls, the Portuguese, the British,

the French, the Dutch, all of them came and looted us, took over what was ours.

Yet we have not done this to any other nation. We have not conquered anyone.

We have not grabbed their land, their culture,

their history and tried to enforce our way of life on them.

Why? Because we respect the freedom of others.That is why my

first vision is that of freedom. I believe that India got its first vision of

this in 1857, when we started the War of Independence. It is this freedom that

we must protect and nurture and build on. If we are not free, no one will respect us.

My second vision for India’s development. For fifty years we have been a developing nation.

It is time we see ourselves as a developed nation. We are among the top 5 nations of the world

in terms of GDP. We have a 10 percent growth rate in most areas. Our poverty levels are falling.

Our achievements are being globally recognised today. Yet we lack the self-confidence to

see ourselves as a developed nation, self-reliant and self-assured. Isn’t this incorrect?

I have a third vision. India must stand up to the world. Because I believe that unless India

stands up to the world, no one will respect us. Only strength respects strength. We must be

strong not only as a military power but also as an economic power. Both must go hand-in-hand.

My good fortune was to have worked with three great minds. Dr. Vikram Sarabhai of the Dept. of

space, Professor Satish Dhawan, who succeeded him and Dr. Brahm Prakash, father of nuclear material.

I was lucky to have worked with all three of them closely and consider this the great opportunity of my life.

I see four milestones in my career"""

**Stemming and lemmatization:**

Stemming:

” Process of reducing the infected words to their word stem.”

Eg: the below list of words are converted to the one with same meaning or stem

[history, historical] =🡺 histori

[finally, final, finalized] =🡺 fina

[going, goes, gone] =🡺 go

**Why: stemming.**

Get the stem word of all the words with same origin of meaning. (either positive or negative). Which makes easier to give the less sized vector of numerical.

Lemmatization:

“process of getting the stem word but with exact and full meaning”

Eg: [history, historical] 🡺 history

[finally, final, finalized] 🡺 final

* Lemmatization takes more time to get the stem word than stemming, bcos it has to analyse first to get the meaningful word.
* Stemming: sentiment, gmail spam, (base word is enough)
* Lemmatization: chatbots, q&a , (get the meaningful word that’s understandable to human)

Stemming can be done using below code:

“from nltk.stem import PorterStemmer”

“stemmer = PorterStemmer()”

“stemmer.stem(word)” gives the word with the stem word as shown above eg;

For example ‘histori’ for history, ‘peopl’ fo people

**Problem with stemming:**

Produced immediate representation of the word may not have any meaning.

Eg: intelligen, fina, etc.

**Stopwords:**

These are the common words that are in any language which don’t have any meaning but helps in creating a meaning full sentence after all adding them in btw the words.

Eg: stopwords of English are :

['i',

'me',

'my',

'myself',

'we',

'our',

'ours',

'ourselves',

'you',

"you're",

"you've",

"you'll",

"you'd",

'your',

'yours',

'yourself',

'yourselves',

'he',

'him',

'his',

'himself',

'she',

"she's",

'her',

'hers',

'herself',

'it',

"it's",

'its',

'itself',

'they',

'them',

'their',

'theirs',

'themselves',

'what',

'which',

'who',

'whom',

'this',

'that',

"that'll",

'these',

'those',

'am',

'is',

'are',

'was',

'were',

'be',

'been',

'being',

'have',

'has',

'had',

'having',

'do',

'does',

'did',

'doing',

'a',

'an',

'the',

'and',

'but',

'if',

'or',

'because',

'as',

'until',

'while',

'of',

'at',

'by',

'for',

'with',

'about',

'against',

'between',

'into',

'through',

'during',

'before',

'after',

'above',

'below',

'to',

'from',

'up',

'down',

'in',

'out',

'on',

'off',

'over',

'under',

'again',

'further',

'then',

'once',

'here',

'there',

'when',

'where',

'why',

'how',

'all',

'any',

'both',

'each',

'few',

'more',

'most',

'other',

'some',

'such',

'no',

'nor',

'not',

'only',

'own',

'same',

'so',

'than',

'too',

'very',

's',

't',

'can',

'will',

'just',

'don',

"don't",

'should',

"should've",

'now',

'd',

'll',

'm',

'o',

're',

've',

'y',

'ain',

'aren',

"aren't",

'couldn',

"couldn't",

'didn',

"didn't",

'doesn',

"doesn't",

'hadn',

"hadn't",

'hasn',

"hasn't",

'haven',

"haven't",

'isn',

"isn't",

'ma',

'mightn',

"mightn't",

'mustn',

"mustn't",

'needn',

"needn't",

'shan',

"shan't",

'shouldn',

"shouldn't",

'wasn',

"wasn't",

'weren',

"weren't",

'won',

"won't",

'wouldn',

"wouldn't"]

And stopwords are available in any language but the library provides it for the few languages, the languages list is as follows.

'arabic',

'azerbaijani',

'bengali',

'danish',

'dutch',

'english',

'finnish',

'french',

'german',

'greek',

'hungarian',

'indonesian',

'italian',

'kazakh',

'nepali',

'norwegian',

'portuguese',

'romanian',

'russian',

'slovene',

'spanish',

'swedish',

'tajik',

'turkish'

**Lemmatization:**

Before this pls try to download wordnet from nltk as below

‘nltk.downlaod(‘wordnet’)’

It can be used in the code with following library “NLTK”

‘from nltk.stem import WordNetLemmatizer”

‘lematizer = WordNetLemmatizer()’

‘lematizer.lemmatize(word)’ to get the meaning full word and it takes a lot of time to analyze and gives the output.