NLP Tasks

About the Module

- Corpus, Tokens and N-Grams
- Tokenization
- Stemming
- Lemmatization
- Part of Speech Tagging
- Dependency Grammar

Corpus, Tokens and Ngrams

- Corpus : Collection of text documents
- Corpus > Documents > Paragraphs > Sentences > Tokens
- Tokens : Smaller units of a text (words, phrases, ngrams)
- Ngrams : combinations of N words / Characters together

Sentence : I love my phone

Unigrams (n =1) : I, Love, my, phone
Bigrams (n=2) : I Love, Love my, my phone
Trigrams (n=3) : I love my, love my phone

Tokenization

- Process of splitting a text object into smaller units (tokens)
- Smaller Units : words, numbers, symbols, ngrams, characters
- White space tokenizer / Unigram tokenizer

  Sentence : “I went to New-York to play football”
  Tokens : “I”, “went”, “to”, “New-York”, “to”, “play”, “football”

- Regular expression tokenizer

  Sentence : “Football,Cricket;Golf Tennis”

  re.split(r'[;,\s]', line)

  Tokens : “Football”, “Cricket”, “Golf”, “Tennis”

Normalization

- Morpheme : base form of a word
- Structure of token : <prefix> <morpheme> <suffix>

  Example : Antinationalist : Anti + national + ist
- Normalization : Process of converting a token into its base form (morpheme)
- Helpful in reducing data dimensionality, text cleaning
- Types : Stemming and Lemmatization

Normalization: Stemming

- Elementary rule based process of removal of inflectional forms from a token
- Outputs the stem of a word
- “laughing”, “laughed”, “laughs”, “laugh” >>> “laugh”
- May generate non-meaningful terms
- his teams are not winning
  >> hi team are not winn

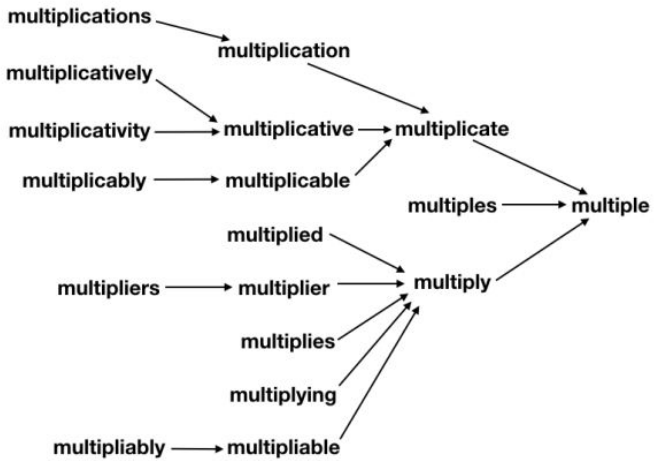
Form	Suffix	Stem
studies	-es	studi
studying	-ing	study
niñas	-as	niñ
niñez	-ez	niñ

Normalization : Lemmatization

- Systematic process for reducing a token to its lemma
- Makes use of vocabulary, word structure, part of speech tags and grammar relations
- Example :

  am, are, is >> be
  running , an , run , rans >> run

- Running, ‘verb’ >> run
  Running, ‘noun’>> running



Part of Speech Tags

- Defines the syntactic context and role of words in the sentence
- Common POS Tags : Nouns, Verbs, Adjectives, Adverbs

Sentence : David has purchased a new Laptop from Apple Store

- Defined by their relationship with the adjacent words
- Machine learning or Rule based process

Part of Speech Tags

1.	CC	Coordinating conjunction	25.	TO	*to*
2.	CD	Cardinal number	26.	UH	Interjection
3.	DT	Determiner	27.	VB	Verb, base form
4.	EX	Existential *there*	28.	VBD	Verb, past tense
5.	FW	Foreign word	29.	VBG	Verb, gerund/present participle
6.	IN	Preposition/subordinating conjunction	30.	VBN	Verb, past participle
7.	JJ	Adjective	31.	VBP	Verb, non-3rd ps. sing. present
8.	JJR	Adjective, comparative	32.	VBZ	Verb, 3rd ps. sing. present
9.	JJS	Adjective, superlative	33.	WDT	*wh*-determiner
10.	LS	List item marker	34.	WP	*wh*-pronoun
11.	MD	Modal	35.	WP$	Possessive *wh*-pronoun
12.	NN	Noun, singular or mass	36.	WRB	*wh*-adverb
13.	NNS	Noun, plural	37.	#	Pound sign
14.	NNP	Proper noun, singular	38.	$	Dollar sign
15.	NNPS	Proper noun, plural	39.	.	Sentence-final punctuation
16.	PDT	Predeterminer	40.	,	Comma
17.	POS	Possessive ending	41.	:	Colon, semi-colon
18.	PRP	Personal pronoun	42.	(	Left bracket character
19.	PP$	Possessive pronoun	43.	)	Right bracket character
20.	RB	Adverb	44.	"	Straight double quote
21.	RBR	Adverb, comparative	45.	‘	Left open single quote
22.	RBS	Adverb, superlative	46.	“	Left open double quote
23.	RP	Particle	47.	’	Right close single quote
24.	SYM	Symbol (mathematical or scientific)	48.	”	Right close double quote

Part of Speech Tags

- Uses:
  - Text cleaning
  - Feature engineering tasks
  - Word sense disambiguation

Sentence1 : Please book my flight for NewYork;
Sentence 2: I like to read a book on NewYork

Sentence1: book / Verb
Sentence2: book / Noun

Constituency Grammar

- Constituents : Words / phrases / group of words
- Constituency Grammar : Organize any sentence into constituents using their properties
- Properties: Part of Speech Tags / Noun Phrases / Verb Phrases

Sentence : <subject> <context> <object>

<subject> The cats / The dogs / They
<context> are running / are barking / are eating
<object> in the park / happily / since the morning

Another view (using part of speech)

< DT NN > < JJ VB > < PRP DT NN > ------> The dogs are barking in the park

Dependency Grammar

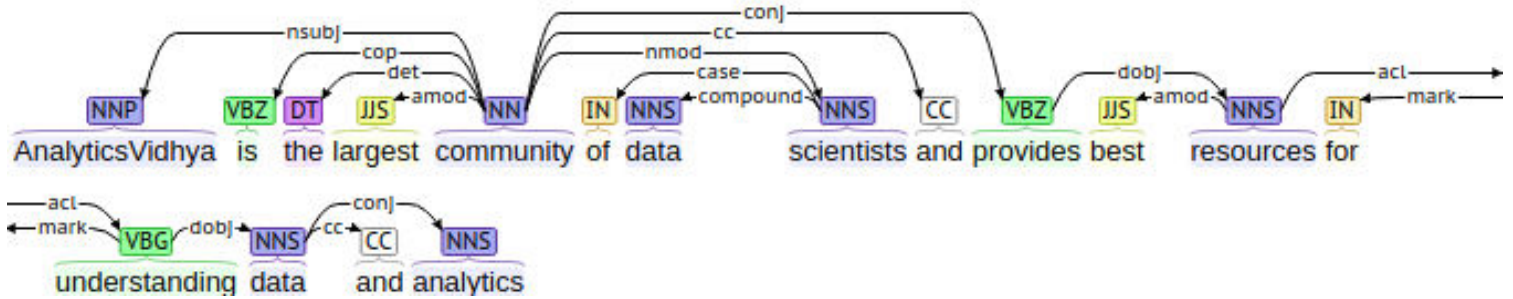
- Words of a sentence depends on which other words (dependencies)

  Example : Modifiers (barking dog)

- Organize words of a sentence according to their dependencies
- All the words are directly or indirectly linked to the root using links
- These dependencies represents relationship among the words in a sentence

Dependency Grammar

- Sentence : *AnalyticsVidhya is the largest community of data scientists and provides best resources for understanding data and analytics*



- *Relation : (Governer, Relation, Dependent)*

  <Analyticsvidhya> <is> <the largest community of data scientists>

Dependency Grammar – Use Cases

- Named Entity Recognization
- Question Answering Systems
- Coreference Resolution
- Text Summarization
- Text Classifications