

Data Science Foundations

Master in Big Data Solutions 2017-2018



Liana Napalkova
liana.napalkova@bts.tech

Ludovico Boratto
ludovico.boratto@bts.tech

Francisco Gutierrez
francisco.gutierrez@bts.tech

Session 5

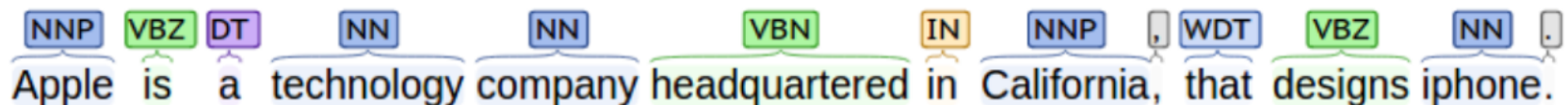
Introduction to Text Mining (Part 2)

Tokenization

- Definition: the process of chopping a sequence of characters into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation.
- Usage: the first step in NLP applications, generating the smallest language units for statistical modeling
- Example
 - Original sentence: Apple is a technology company headquartered in California, that designs iphone.
 - Tokenized sentence: ["Apple", "is", "a", "technology", "company", "headquartered", "in", "California", "that", "designs", "iphone"]

Part-of-speech tagging

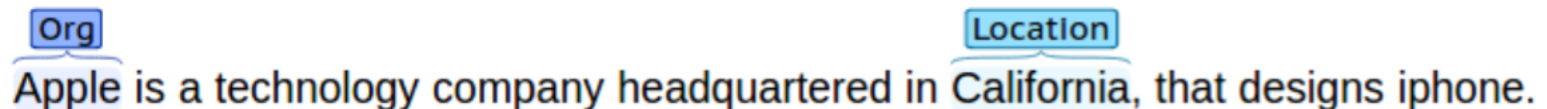
- Definition: the process of marking up a word as a particular part of speech such as nouns, verbs, adjectives, etc., based on its grammatical relationship with other words in the text.
- Usage: distinguishing words of different meaning but have the same form, generating phrases.
- Example



- Full list of part-of-speech tags: [Penn Treebank](http://www.nlp.cs.brown.edu/penn-treebank/)

Named entity recognition

- Definition: the process of identifying and classifying elements in text into pre-defined categories such as the names of persons, organizations, locations and so on.
- Usage: semantic analysis.
- Example

Apple is a technology company headquartered in California, that designs iphone.

Lemmatization / stemming

- Definition: the process of reducing the inflectional forms or derivationally related forms of a word to a common base form.
- Usage: reduce redundant information and computational cost, but also increases the risk of losing information
- For example:
 - am, is, are → be
 - car, cars, car's, cars → car

More NLP techniques

- Sentence splitting
 - Definition: the process of dividing a continuous document into separate sentences by identifying the boundaries of the sentences.
- Stop words removal
 - Definition: remove a set of most common words that are believed to carry little information
 - Stop words: there is no unique stopwords list, but some software such NLTK provides a set of most common functional words in English, and researchers can build their own list of stop words based on their domain knowledge.

DOCUMENT VECTOR REPRESENTATIONS

Finding salient tokens (words)

- Most frequent tokens?

Rank	Word	Rank	Word	Rank	Word	Rank	Word	Rank	Word
1	the	21	this	41	so	61	people	81	back
2	be	22	but	42	up	62	into	82	after
3	to	23	his	43	out	63	year	83	use
4	of	24	by	44	if	64	your	84	two
5	and	25	from	45	about	65	good	85	how
6	a	26	they	46	who	66	some	86	our
7	in	27	we	47	get	67	could	87	work
8	that	28	say	48	which	68	them	88	first
9	have	29	her	49	go	69	see	89	well
10	I	30	she	50	me	70	other	90	way
11	it	31	or	51	when	71	than	91	even
12	for	32	an	52	make	72	then	92	new
13	not	33	will	53	can	73	now	93	want
14	on	34	my	54	like	74	look	94	because
15	with	35	one	55	time	75	only	95	any
16	he	36	all	56	no	76	come	96	these
17	as	37	would	57	just	77	its	97	give
18	you	38	there	58	him	78	over	98	day
19	do	39	their	59	know	79	think	99	most
20	at	40	what	60	take	80	also	100	us

Document Frequency

- Intuition: Uninformative words appear in many documents (not just the one we are concerned about)
- Salient word
 - High count within the document
 - Low count across documents

TF•IDF score

- Term Frequency (TF)
 - $TF(x) = \log_{10}(1+c(x))$ or $c(x)$
 - $c(x)$ is the number of times x appears in the document
- Inverse Document Frequency (IDF):

$$IDF(x) = \log_{10} \left(\frac{N_{docs}}{DF(x)} \right)$$

- $DF(x)$ is the number of documents x appears in.