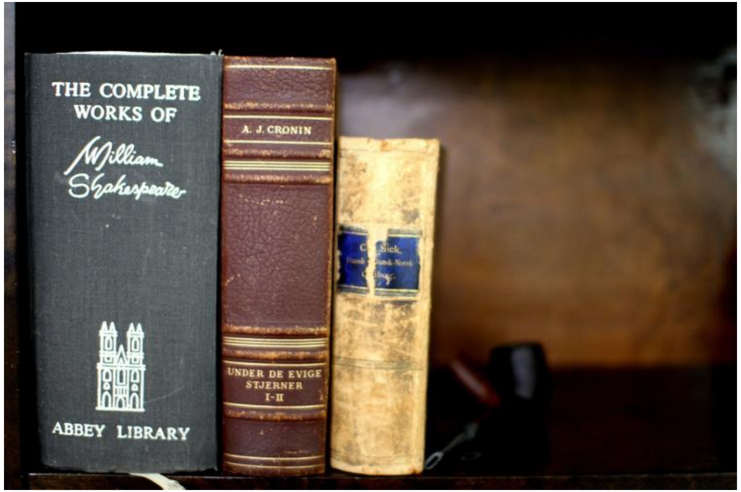
greatlearning
Learning for Life

Natural Language Processing

Agenda

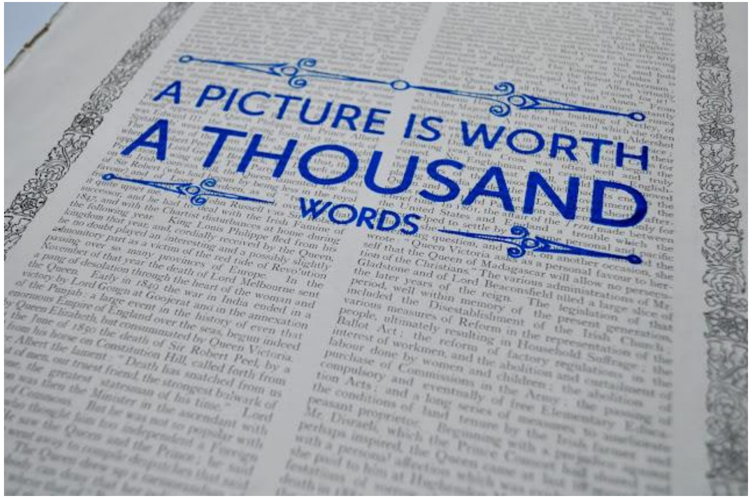
- ➢ Understanding NLP
- ➢ Bag of Words Model
  - ○ Count Vectorizer
  - ○ Tf-IDF Vectorizer
- ➢ Stemming , Lemmatization
- ➢ Sentiment Analysis
- ➢ Case studies
  - ○ SMS classification and sentiment analysis



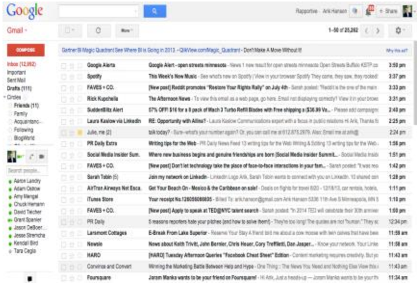
Natural Language Processing (NLP)

.. a sub-field of AI with focus on enabling machines to understand and process human languages

NLP vs Computer Vision



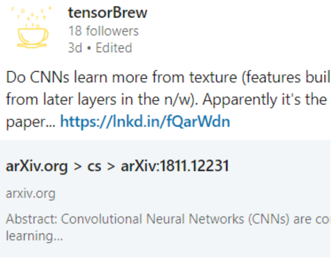
Extracting meaning out of Language data in general is more complex than Vision data



Gartner BI Magic Quadrant See Where BI is Going in 2013 - QlikView.com/Magic_Quadrant - Don't Make A Move Without It! Why this ad?

Google Alerts	Google Alert - open streets minnesota - News 1 new result for open streets minnesota Open Streets Buffalo KSTP.co	3:59 pm
Spotify	This Week's New Music - See what's new on Spotify | View in your browser Spotify They came, they saw, they rocked!	3:37 pm
FAVES + CO.	[New post] Reddit promotes "Restore Your Rights Rally" on July 4th - Sarah posted: "Reddit is the one of the main	3:33 pm
Rick Kupchella	The Afternoon News - To view this email as a web page, go here. Email not displaying correctly? View it in your brows	3:31 pm
SuddenBlitz Alert	57% OFF! $16 for a 8 pack of Mach 3 Turbo Refill Blades with Free shipping a ($36.99 Va... - Please add campaign	2:48 pm
Laura Kaslow via LinkedIn	RE: Opportunity with Allina? - Laura Kaslow Communications expert with a focus in public relations Hi Arik, Thanks fo	2:25 pm
Julie, me (2)	talk today? - Sure--what's your number again? Or, you can call me at 612.875.2879. Alex: Email me at arik@	2:24 pm
PR Daily Extra	Writing tips for the Web - PR Daily News Feed 13 writing tips for the Web Writing & Editing 13 writing tips for the Web	1:58 pm
Social Media Insider Sum.	Where new business begins and genuine friendships are born [Social Media Insider Summit... - Social Media Insid	1:51 pm
FAVES + CO.	[New post] Don't let technology take the place of face-to-face interactions in your fam... - Sarah posted: "It was rec	1:42 pm
Sarah Tobin (5)	Join my network on LinkedIn - LinkedIn Logo Arik, Sarah Tobin wants to connect with you on LinkedIn. 13 shared con	1:28 pm
AirTran Airways Net Esca.	Get Your Beach On - Mexico & the Caribbean on sale! - Deals on flights for travel 8/20 - 12/18/13, car rentals, hotels,	1:11 pm
iTunes Store	Your receipt No.128056086835 - Billed To: arik.hanson@gmail.com Arik Hanson 5336 11th Ave S Minneapolis, MN 5	1:10 pm
FAVES + CO.	[New post] Apply to speak at TED@NYC talent search - Sarah posted: "In 2014 TED will celebrate their 30th anniver	1:00 pm
PR Daily	5 reasons reporters hate your pitches (and how to solve them) - They're too long! The quotes are not "human." They sc	12:34 pm
Larsmont Cottages	E-Break From Lake Superior - Reserve Your Stay A friend told me about a cow moose with twin calves that have bee	11:59 am
Newsle	News about Keith Trivitt, John Bernier, Chris Heuer, Cory Treffiletti, Dan Jasper... - Know your network. Your Linke	11:58 am
HARO	[HARO] Tuesday Afternoon Queries "Facebook Cheat Sheet" Edition - Content marketing requires creativity. But yo	11:43 am
Convince and Convert	Winning the Marketing Battle Between Help and Hype - One Thing :: The News You Need and Nothing Else View this i	11:43 am
Foursquare	Joram Manka wants to be your friend on Foursquare! - Hi Arik, Just a heads-up — Joram Manka wants to be your fri	11:34 am

Emails

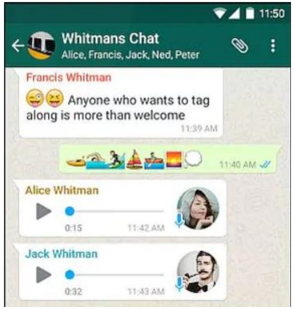


tensorBrew
18 followers
3d • Edited

Do CNNs learn more from texture (features buil from later layers in the n/w). Apparently it's the paper... https://lnkd.in/fQarWdn

arXiv.org > cs > arXiv:1811.12231
arxiv.org
Abstract: Convolutional Neural Networks (CNNs) are co learning...

Social Media Updates



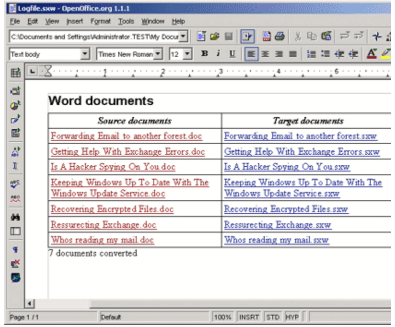
Francis Whitman
😜😝 Anyone who wants to tag along is more than welcome
11:39 AM

11:40 AM

Alice Whitman
0:15 11:42 AM

Jack Whitman
0:32 11:43 AM

Chat interactions


Logfile.sxw - OpenOffice.org 1.1.1
File Edit View Insert Format Tools Window Help
C:\Documents and Settings\Administrator.TEST\My Docu
Text body
Times New Roman
12
B i U
Page 1 / 1
Default
100%
INSRT
STD
HYP

Word documents

Source documents	*Target documents*
Forwarding Email to another forest.doc	Forwarding Email to another forest.sxw
Getting Help With Exchange Errors.doc	Getting Help With Exchange Errors.sxw
Is A Hacker Spying On You.doc	Is A Hacker Spying On You.sxw
Keeping Windows Up To Date With The Windows Update Service.doc	Keeping Windows Up To Date With The Windows Update Service.sxw
Recovering Encrypted Files.doc	Recovering Encrypted Files.sxw
Ressurecting Exchange.doc	Ressurecting Exchange.sxw
Whos reading my mail.doc	Whos reading my mail.sxw

7 documents converted

Office documents

...other audio data

Some examples of Language Data

NLP presents a huge opportunity...

Most organization have humongous amount of textual data but struggling to get value out of it.

Check Credit worthiness

Language Translation

Sentiment Analysis

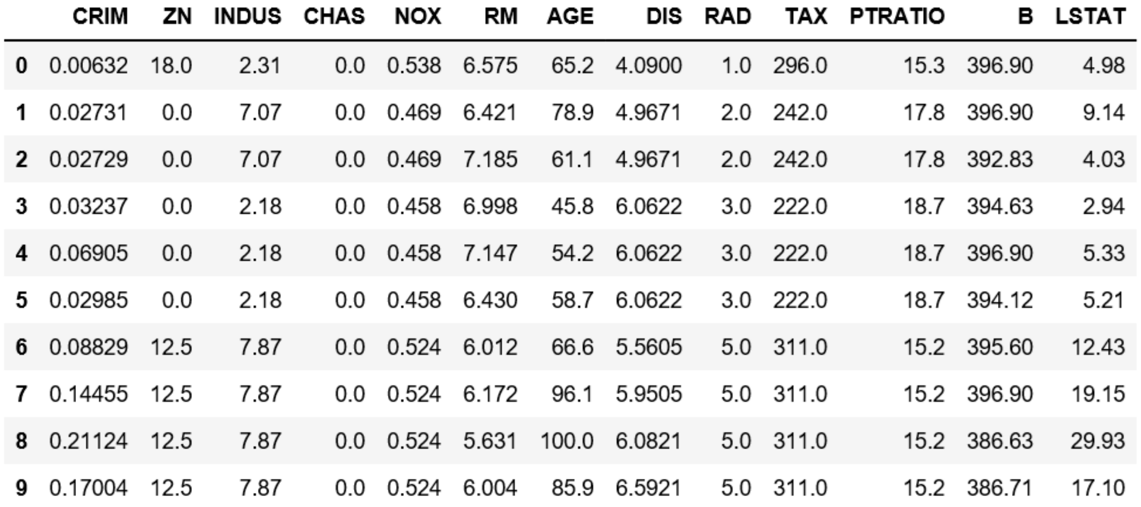
Customer Support

Work Routing

Identify Similar Legal cases
(Document similarity)

Some NLP based solutions

Working with Textual Data

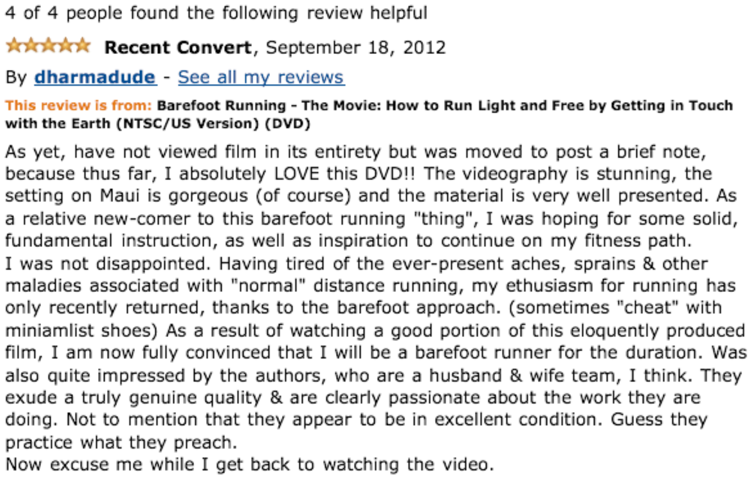


	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33
5	0.02985	0.0	2.18	0.0	0.458	6.430	58.7	6.0622	3.0	222.0	18.7	394.12	5.21
6	0.08829	12.5	7.87	0.0	0.524	6.012	66.6	5.5605	5.0	311.0	15.2	395.60	12.43
7	0.14455	12.5	7.87	0.0	0.524	6.172	96.1	5.9505	5.0	311.0	15.2	396.90	19.15
8	0.21124	12.5	7.87	0.0	0.524	5.631	100.0	6.0821	5.0	311.0	15.2	386.63	29.93
9	0.17004	12.5	7.87	0.0	0.524	6.004	85.9	6.5921	5.0	311.0	15.2	386.71	17.10

Boston Housing Price dataset

Structured Data

1. has features (columns)
2. All records have same features
3. Features maintain order across examples



4 of 4 people found the following review helpful

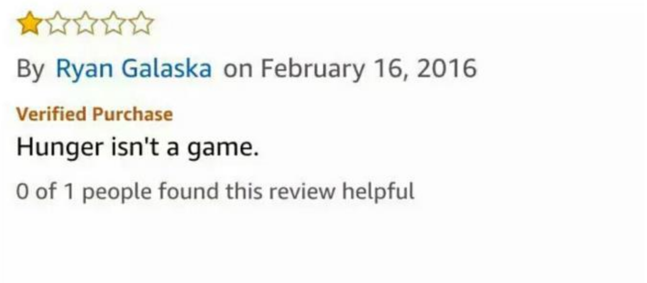
★★★★★ **Recent Convert**, September 18, 2012

By **dharmadude** - See all my reviews

This review is from: **Barefoot Running - The Movie: How to Run Light and Free by Getting in Touch with the Earth (NTSC/US Version) (DVD)**

As yet, have not viewed film in its entirety but was moved to post a brief note, because thus far, I absolutely LOVE this DVD!! The videography is stunning, the setting on Maui is gorgeous (of course) and the material is very well presented. As a relative new-comer to this barefoot running "thing", I was hoping for some solid, fundamental instruction, as well as inspiration to continue on my fitness path.
I was not disappointed. Having tired of the ever-present aches, sprains & other maladies associated with "normal" distance running, my ethusiasm for running has only recently returned, thanks to the barefoot approach. (sometimes "cheat" with miniamlist shoes) As a result of watching a good portion of this eloquently produced film, I am now fully convinced that I will be a barefoot runner for the duration. Was also quite impressed by the authors, who are a husband & wife team, I think. They exude a truly genuine quality & are clearly passionate about the work they are doing. Not to mention that they appear to be in excellent condition. Guess they practice what they preach.
Now excuse me while I get back to watching the video.

Hunger Games (2012)



★☆☆☆☆

By Ryan Galaska on February 16, 2016

Verified Purchase

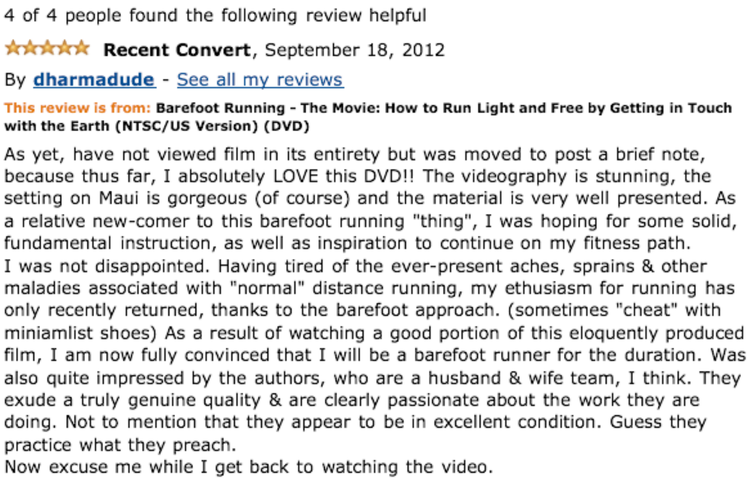
Hunger isn't a game.

0 of 1 people found this review helpful

Textual Data

1. Do not have features like in tabular data.
2. Examples usually have different size.

What could be the features of textual Data?



4 of 4 people found the following review helpful

★★★★★ **Recent Convert**, September 18, 2012

By **dharmadude** - See all my reviews

This review is from: **Barefoot Running - The Movie: How to Run Light and Free by Getting in Touch with the Earth (NTSC/US Version) (DVD)**

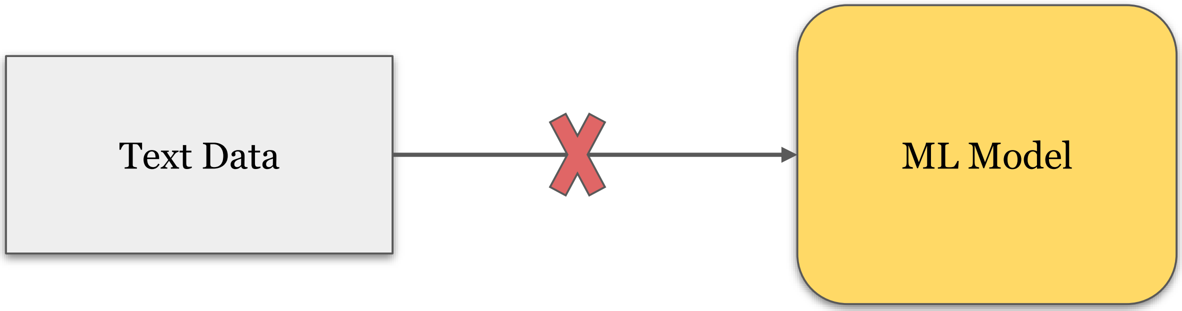
As yet, have not viewed film in its entirety but was moved to post a brief note, because thus far, I absolutely LOVE this DVD!! The videography is stunning, the setting on Maui is gorgeous (of course) and the material is very well presented. As a relative new-comer to this barefoot running "thing", I was hoping for some solid, fundamental instruction, as well as inspiration to continue on my fitness path.

I was not disappointed. Having tired of the ever-present aches, sprains & other maladies associated with "normal" distance running, my ethusiasm for running has only recently returned, thanks to the barefoot approach. (sometimes "cheat" with miniamlist shoes) As a result of watching a good portion of this eloquently produced film, I am now fully convinced that I will be a barefoot runner for the duration. Was also quite impressed by the authors, who are a husband & wife team, I think. They exude a truly genuine quality & are clearly passionate about the work they are doing. Not to mention that they appear to be in excellent condition. Guess they practice what they preach.

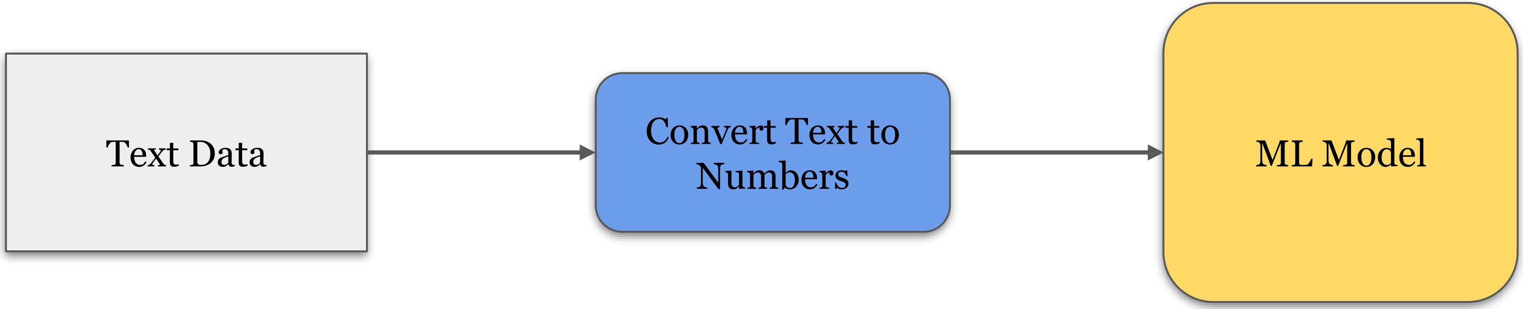
Now excuse me while I get back to watching the video.

Features in textual Data

1. Words?
2. Characters?
3. Combination of words (n-grams)?
4. Sentences?
5. What else?



Math works with numbers



Text Data

Convert Text to Numbers

ML Model

How to convert text into Numbers

Bag of Words

1. A simple feature extraction approach in NLP
2. Ignores grammar / structure
3. Represents each document by measuring presence of vocabulary words

Document #1

He is a good boy. She is also good.

Document #2

Radhika is a good person.

Vocabulary

a, also, boy, good, He, is, person, She, Radhika

Document #1

He is a good boy. She is also good.

Document #2

Radhika is a good person.

Vocabulary

a, also, boy, good, He, is, person, She, Radhika

	a	**also**	**boy**	**good**	**He**	**Is**	**person**	**She**	**Radhika**
Index	0	1	2	3	4	5	6	7	8

Assign index for each word in Vocabulary

Document #1

He is a good boy. She is also good.

Document #2

Radhika is a good person.

Vocabulary

a, also, boy, good, He, is, person, She, Radhika

	a	also	boy	good	He	Is	person	She	Radhika
Index	0	1	2	3	4	5	6	7	8
Document# 1									

Count how many times each word in Vocabulary appears in Document #1

Document #1

He is a good boy. She is also good.

Document #2

Radhika is a good person.

Vocabulary

a, also, boy, good, He, is, person, She, Radhika

	a	also	boy	good	He	Is	person	She	Radhika
Index	0	1	2	3	4	5	6	7	8
Document #1	1	1	1	2	1	2	0	1	0

Document #1 = [1, 1, 1, 2, 1, 2, 0, 1, 0]

Document #1

He is a good boy. She is also good.

Document #2

Radhika is a good person.

Vocabulary

a, also, boy, good, He, is, person, She, Radhika

	a	also	boy	good	He	Is	person	She	Radhika
Index	0	1	2	3	4	5	6	7	8
Document #1	1	1	1	2	1	2	0	1	0
Document #2	1	0	0	1	0	1	1	0	1

Document #2 = [1, 0, 0, 1, 0, 1, 1, 0, 1]

Document #1

He is a good boy. She is also good.

Document #2

Radhika is a good person.

Vocabulary

a, also, boy, good, He, is, person, She, Radhika

	a	also	boy	good	He	Is	person	She	Radhika
Index	0	1	2	3	4	5	6	7	8
Document #1	1	1	1	2	1	2	0	1	0
Document #2	1	0	0	1	0	1	1	0	1

Count Vector

SMS Classification: Ham or Spam

Hands-On

TF-IDF Vector

Not just simple counting

Document #1

He is a good boy. She is also good.

He	1
is	2
a	1
good	2
boy	1
she	1
also	1
Total	**9**

$$TF = \frac{Frequency\ of\ the\ word\ in\ a\ Doc}{Total\ number\ of\ words\ in\ the\ Doc}$$

TF(He, doc#1) = 1/9 = 0.11

TF(good, doc#1) = 2/9 = 0.22

TF captures how important a word is to the document (without looking at other documents in the dataset)

Document #2

Radhika is a good person.

Radhika	1
is	1
a	1
good	1
person	1
Total	**5**

$$TF = \frac{Frequency\ of\ the\ word\ in\ a\ Doc}{Total\ number\ of\ words\ in\ the\ Doc}$$

TF(He, doc#2) = 0/5 = 0

TF(good, doc#2) = 1/5 = 0.2

Document #1

He is a good boy. She is also good.

Document #2

Radhika is a good person.

He	1
is	2
a	1
good	2
boy	1
she	1
also	1
Total	**9**

Radhika	1
is	1
a	1
good	1
person	1
Total	**5**

$$IDF = log(\frac{Num\ of\ Docs}{Word\ in\ Num\ of\ Docs})$$

IDF(He) = log(2/1) = 0.301

IDF(good) = log(2/2) = 0

IDF tells us if a word (feature) can be used to distinguish documents. If a word appears in majority of the documents then IDF will be close to '0' *i.e.* give low weightage to that feature.

Document #1

He is a good boy. She is also good.

Document #2

Radhika is a good person.

He	1
is	2
a	1
good	2
boy	1
she	1
also	1
Total	**9**

Radhika	1
is	1
a	1
good	1
person	1
Total	**5**

$$IDF = log(\frac{Num\ of\ Docs}{Word\ in\ Num\ of\ Docs})$$

IDF(He) = log(2/1) = 0.301

IDF(good) = log(2/2) = 0

TF-IDF(He, doc#1) = 0.11 * 0.301 = 0.03311

TF-IDF(good, doc#1) = 0.22 * 0 = 0

TF-IDF(He, doc#2) = 0 * 0.301 = 0

TF-IDF(good, doc#2) = 0.2 * 0 = 0

Document #1

He is a good boy. She is also good.

Document #2

Radhika is a good person.

Vocabulary

a, also, boy, good, He, is, person, She, Radhika

	a	also	boy	good	He	Is	person	She	Radhika
Index	0	1	2	3	4	5	6	7	8
Document #1				0	0.03311				
Document #2				0	0				

TF-IDF Vector

TF-IDF in Scikit-Learn

Hands-On

Hands-On: Sentiment Analysis

Text Preprocessing

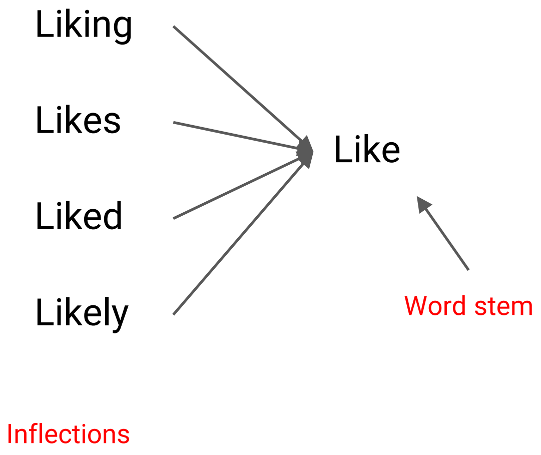
Stopwords Removal

You The

a and

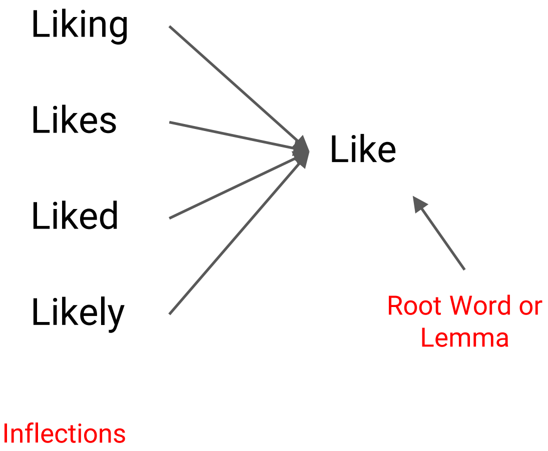
1. High frequency words *i.e* present in most documents
2. Can not be used to distinguish between documents
3. Hence can be removed as features

Stemming



1. Converts inflections to root or word stem
2. Used for dimensionality reduction
3. Word stem may **not be present in dictionary**
4. Popular algorithms include Potter Stemmer, Lovins Stemmer etc

Lemmatization



Liking

Likes

Liked

Likely

Like

Root Word or Lemma

Inflections

1. Very similar to Stemming
2. Converts inflections to root word or **Lemma**
3. Word stem may **not be present in dictionary**

Using NLTK

Information Retrieval

Hyderabad is the capital of the Indian state of Telangana occupying 650 square kilometres (250 sq mi) along the banks of the Musi River. Hyderabad City has a population of about 6.9 million, making it the fourth-most populous city in India.

Established in 1591 by Muhammad Quli Qutb Shah, Hyderabad remained under the rule of the Qutb Shahi dynasty for nearly a century before the Mughals captured the region.

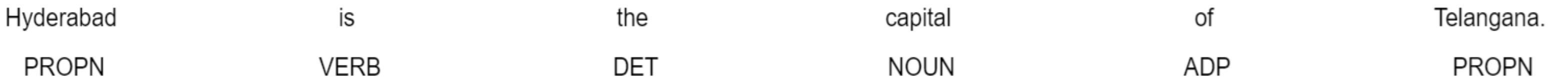
Hyderabad

- Capital of ?
- How populated is Hyderabad?
- Who established Hyderabad?

Understanding Language Structure & Syntax

Bag of words does not keep order of words and hence can not be used to understand the meaning of the text.

Part-of-Speech (POS) Tagging

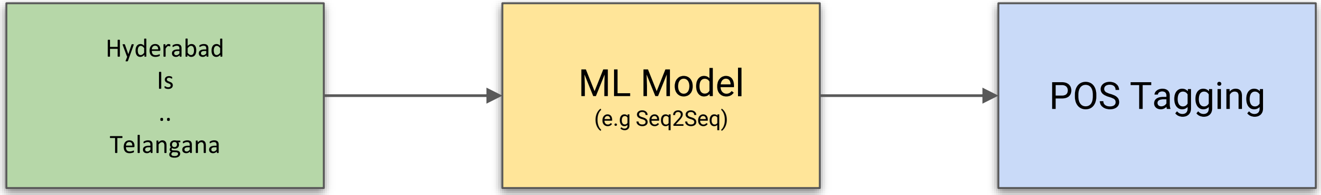


Hyderabad	is	the	capital	of	Telangana.
PROPN	VERB	DET	NOUN	ADP	PROPN

1. Assign grammatical properties (e.g. noun, verb, adverb, adjective etc.) to words.
2. Allows understanding of language structure and syntax.
3. These properties can used to extract information by using language rules.
4. Multiple NLP libraries support POS tagging e.g. NLTK, spaCy

Part-of-Speech (POS) Tagging

Hyderabad	is	the	capital	of	Telangana.
PROPN	VERB	DET	NOUN	ADP	PROPN



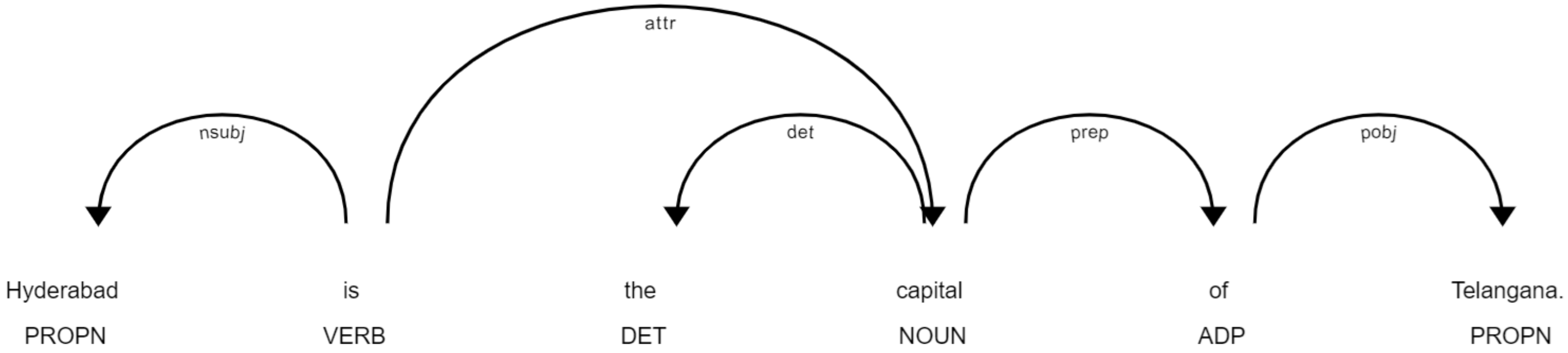
Hyderabad
Is
..
Telangana

ML Model
(e.g Seq2Seq)

POS Tagging

POS tagging is done using a trained ML Model

Dependency Parsing



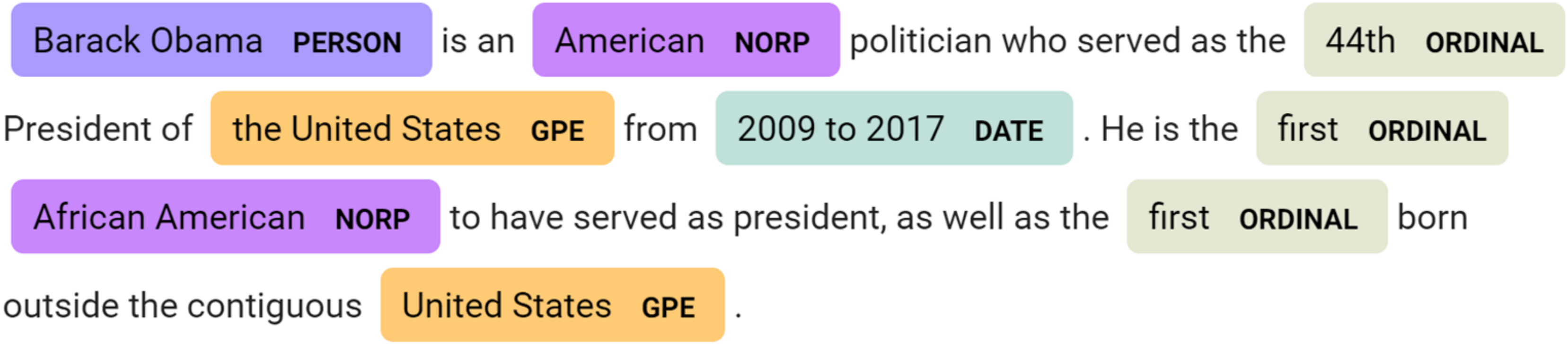
1. Shows how words in a sentence relate to each other.
2. Allows further understanding of language structure and syntax.

Named Entity Recognition (NER)

Barack Obama is an American politician who served as the 44th President of the United States from 2009 to 2017. He is the first African American to have served as president, as well as the first born outside the contiguous United States.

1. Classifies text into predefined categories or real world entities.
2. Used for information extraction, improve search algorithms, content recommendations.

Named Entity Recognition (NER)



Barack Obama **PERSON** is an American **NORP** politician who served as the 44th **ORDINAL** President of the United States **GPE** from 2009 to 2017 **DATE** . He is the first **ORDINAL** African American **NORP** to have served as president, as well as the first **ORDINAL** born outside the contiguous United States **GPE** .

1. Classifies text into predefined categories or real world entities.
2. Used for information extraction, improve search algorithms, content recommendations.